IMPROVING THE SENTIMENT CLASSIFICATION OF STOCK TWEETS

by

SHENG LI

A thesis submitted to
The University of Birmingham

for the degree of

DOCTOR OF PHILOSOPHY (PHD)

Department of English
School of English, Drama and American & Canadian Studies
College of Arts and Law
University of Birmingham

February, 2014

ABSTRACT

This research focuses on improving stock tweet sentiment classification accuracy with the addition of the linguistic features of stock tweets. Stock prediction based on social media data has been popular in recent years, but none of the previous studies have provided a comprehensive understanding of the linguistic features of stock tweets. Hence, applying a simple statistical model to classifying the sentiment of stock tweets has reached a bottleneck. Thus, after analysing the linguistic features of stock tweets, this research used these features to train four machine learning classifiers. Each of them showed an improvement, and the best one achieved a 9.7% improvement compared to the baseline model. The main contributions of this research are fivefold: (a) it provides an in-depth linguistic analysis of stock tweets; (b) it gives a clear and comprehensive definition of stock tweets; (c) it provides a simple but effective way to automatically identify stock tweets; (d) it provides a simple but effective method of generating a localised sentiment keyword list; and (e) it demonstrates a significant improvement of stock tweet sentiment classification accuracy.

DEDICATION

謹以此文獻給我摯愛的父母。

I dedicate this dissertation to my dearest parents.

ACKNOWLEDGEMENTS

Doing a PhD has been the most challenging and arduous time in my life, and over these four years, I received tremendous help. To start, I would like to thank my parents. Without their generous financial and spiritual support, I would never have accomplished this dream. Then, I am thankful to my supervisors, Dr. Nicholas Groom and Dr. Neil Miller for all of the discussions about concepts and techniques in sentiment analysis and corpus linguistics as well as for all of the drafts they have reviewed. Also, I am grateful to my initial supervisor Dr. Oliver Mason for his invaluable support in the first three years of my PhD and for leading me to the programming and natural language processing world. Next, I appreciate all the staff in the Department of English, especially those who gave me lectures during my MPhil and those who pushed me hard during the panels. I would also thank my bachelor referees, Prof. Heyu Wang and Prof. Chuncan Feng who helped me to come out from the darkest time in my life and recommend me to Birmingham. The last thank you goes to my friends in Britain and back in China who have helped me to make it to the end of this process, particularly to Yingying Weng, Taylor Xu, and Xi'er Zhang (who helped me by commenting on the very first draft of my thesis), Amy Aisha Brown and Dr. Emma Tonkin (who proofread the final draft of my thesis), Dr. Alastair Sanderson (who patiently helped me to sort out R programming problems), Dr. Andy Pryke and Phil Smith (who discussed the details of my experiments with me), Dr. Manal Makoshi (who helped me with proofreading the two posters on my PhD project), and Ruodan Zhang (who criticised but also motivated me greatly in the first two years of my PhD).

Contents

Chapter 1 Introduction **21**

1.1 Rationale of the Research 23

1.1.1 Economics 24

1.1.2 Sentiment analysis 26

1.1.3 Linguistic features of tweets 27

1.2 Statement of the Problem 29

1.3 Structure 31

Chapter 2 Background **36**

2.1 Economics Theories Related to Stock Prediction 36

2.1.1 The predictable market 37

2.1.2 Data sources for stock prediction 40

2.1.3 Using social media as the data source 42

2.1.4 The relationship model between tweets and the stock market 43

2.1.5 Some criticisms of stock prediction based on tweets 45

2.2 Introduction of Twitter 47

2.2.1 Features of Twitter 48

2.2.2 Data releasing of Twitter 51

2.2.3 The development of Twitter 54

2.2.4 Features of Tweets 56

2.2.5 Prediction power of Twitter 60

2.2.6 Influence of Twitter on the market 62

2.3 Summary 65

Chapter 3 Previous Research **67**

3.1 Introduction of Sentiment Analysis 67

3.1.1 Main characteristics of sentiment analysis 69

3.1.2 Development of sentiment analysis 70

3.1.3 Main classification approaches to sentiment analysis 71

3.1.4 Summary 73

3.2 Previous Studies on Stock Prediction Based on Social Media 73

3.2.1 Overall market prediction based on public mood 74

3.2.2 Individual ticker prediction based on ticker-specific mood 77

3.2.3 Other approaches to using tweets to predict the market 79

3.2.4 Summary 80

3.3 Definition of Stock Tweets 81

3.3.1 Collecting general tweets 81

3.3.2 Collecting ticker-specific tweets 83

3.3.3 Other approaches to collecting tweet data 84

3.3.4 Defining stock tweets 85

3.4 Linguistic Features of Stock Tweets 86

3.4.1 Vocabulary features 86

3.4.2 Conversation features 87

3.4.3 Hashtag and cashtag features 88

3.4.4 Topic features 89

3.4.5 Summary 90

3.5 Identification of Stock Tweets 90

3.6 Sentiment Classification in Stock Prediction 91

3.6.1 Bag-of-words Model 93

3.6.2 Semantic orientation 97

3.6.3 Emoticon model 98

3.6.4 Stock tweets classification strategy 98

3.7 Summary 100

Chapter 4 Tweet Data Collection and Annotation **103**

4.1 Difficulties of Data Preparation 103

4.1.1 The paradoxical sparseness of massive tweet data sets 104

4.1.2 Frequent changes of Twitter's data releasing policy 104

4.1.3 Non-standard expressions in tweets 107

4.1.4 Disagreement of the trinary sentiment relationship 108

4.1.5 Different time scale of tweets 108

4.2 Stock Tweet Collection 108

4.2.1 Collection criteria 109

4.2.2 Raw data structure 111

4.2.3 Twitter Search API 114

4.2.4 Keyword selection 115

4.2.5 Server configuration 117

4.2.6 Collection rate 118

4.2.7 Summary 122

4.3 Raw Tweet Manipulation 123

4.3.1 Original folder structure 123

4.3.2 New record format 124

4.3.3 Escape characters 125

4.3.4 Retweets 126

4.3.5 Repeated tweets 127

4.4 Manual Annotation of Stock Tweets 131

4.4.1 Introduction of the tweets from the ticker tweet crawler 131

4.4.2 Manual annotation 135

4.4.3 The multi-level classification criteria 136

4.4.4 Non-stock-related tweets (NSR) 139

4.4.5 Non-ticker-related tweets (NTR) 140

4.4.6 Ticker-related tweets 143

4.4.7 Negative tweets (NEG) 145

4.4.8 Neutral tweets (NEU) 146

4.4.9 Positive tweets (PST) 148

4.4.10 Further discussion on classification criteria 149

4.4.11 Summary 151

4.5 Relationship of Annotated Tweets 152

4.6 Summary 153

Chapter 5 Temporal Relationship of Stock Price and Tweet Sentiment 154

5.1 Stock Price Data Collection 154

5.2 Introduction of Time Series Analysis 156

5.2.1 Time series data 156

5.2.2 Test of correlation between different time series data 158

5.2.3 Introduction of the ACF and the CCF functions in R 159

5.2.4 Correlation direction of the ACF and the CCF function 161

5.2.5 Plotting the ACF test results in R 162

5.2.6 Pearson product-moment correlation coefficient 164

5.2.7 Imputation 164

5.3 Time Series Analysis of Raw Data 166

5.3.1 Time correlation of the raw GE ticker's prices 166

5.3.2 Time correlation of the raw GE tweet sentiments 169

5.3.3 Time correlation between the raw GE price and tweet sentiments 171

5.3.4 Discussion 172

5.4 Time Series Analysis of Transformed Data 173

5.4.1 Percentage normalisation 174

5.4.2 Window function Transformation 180

5.4.3 Moving average smoothing 183

5.4.4 Discussion 191

5.5 Discussion of Neutral Sentiment 192

5.5.1 Three neutral sentiments 192

5.5.2 Overall trends of the different neutral sentiments 194

5.6 Summary 196

Chapter 6 Methodology for Analysis and Classification 198

6.1 A Linguistic Approach to Tweet Analysis 198

6.1.1 Introduction to corpus linguistics 198

6.1.2 Frequency analysis 199

6.1.3 Concordance analysis 200

6.1.4 Keyness analysis 201

6.1.5 Choice of using keyness analysis 201

6.1.6 Tweet normalisation 202

6.1.7 Stemming 203

6.1.8 Part-of-speech tagging 205

6.1.9 Summary 205

6.2 Statistical Tests 206

6.2.1 Normality tests 206

6.2.2 Significance tests 207

6.3 Machine Learning 209

6.3.1 Introduction to machine learning 209

6.3.2 Motivation of applying machine learning 210

6.3.3 Supervised and unsupervised learning 211

6.3.4 Decision tree classification 212

6.3.5 Random forests classification 213

6.3.6 Naïve Bayes classification 214

6.3.7 Support vector machine classification 215

6.3.8 Cross validation 216

6.3.9 Test set of CAT tweets 216

6.4 Summary 217

Chapter 7 Sentiment Classification Based on the Internal Linguistic Features 218

7.1 Test of Word Lists in Previous Studies 219

7.2 Frequency Analysis of Stock Tweets 221

7.2.1 The most frequent unigrams in stock tweets 221

7.2.2 The least frequent unigrams in stock tweets 225

7.2.3 Discussion 228

7.3 Design of A Localised Bag-of-words Model 229

7.3.1 Stemming ticker-related tweets 229

7.3.2 Keyness analysis of the unigram lists 232

7.3.3 Generating a word matrix based on a localised word list 236

7.3.4 Test of the bag-of-words model based on a localised word list 240

7.3.5 Discussion 241

7.4 Sentiment Classification Based on the Bag-of-words Model 242

7.5 Part-of-speech Analysis 245

7.5.1 Part-of-speech annotation 245

7.5.2 Analysis of POS tagged unigrams 248

7.6 Sentiment Classification Based on Part-of-speech Features 253

7.7 Discussion 256

7.8 Summary 259

Chapter 8 Sentiment Classification Based on the External Linguistic Features 261

8.1 External Linguistic Features of Stock Tweets 262

8.1.1 Word count including URLs 263

8.1.2 Character count including URLs 265

8.1.3 Word count excluding URL 267

8.1.4 Character count excluding URL 269

8.1.5 Hashtag count 270

8.1.6 Cashtag count 272

8.1.7 Retweet count 273

8.1.8 Discussion 275

8.2 Automatic Classification of the Ticker-related Tweets 276

8.2.1 Data preparation for the classification 277

8.2.2 Stock level training 279

8.2.3 Ticker-level training 279

8.2.4 Polarity-level training 281

8.3 Discussion 282

8.4 Summary 283

Chapter 9 Sentiment Classification Based on the Hybrid Features 284

9.1 Review of Sentiment Classifications in the Previous Chapters 284

9.2 Classification Based on the Combined Internal Features 286

9.3 Classification Based on the POS and External Features 288

9.4 Classification Based on the Bag-of-words and External Features 289

9.5 Classification Based on the Combined Internal and External Features 291

9.6 Discussion 292

9.7 Summary 293

Chapter 10 Conclusion 295

10.1 Research Context 295

10.2 Discussion 296

10.2.1 Linguistic features of stock tweets 296

10.2.2 Definition of stock tweets 298

10.2.3 Identification of stock tweets 300

10.2.4 Definition of sentiment polarities in stock tweets 301

10.2.5 Processing of neutral tweets 302

10.2.6 Generating a localised word list for sentiment classification 303

10.2.7 Sentiment classification of stock tweets 304

10.3 Conclusion 306

10.4 Limitation and Future Work 309

Appendix A Scripts 310

Appendix A.1 Scripts for Data Collection 310

Appendix A.1.1 Scripts for crawling ticker tweets 310

Appendix A.1.2 Scripts for crawling ticker tweets 310

Appendix A.1.3 Scripts for crawling retweets 310

Appendix A.2 Scripts for Data Analysis 311

Appendix B Address Book for Data Collection 312

Appendix B.1 Address Book for Ticker Tweets 312

Appendix B.2 Address Book for Media Tweets 313

Appendix B.3 Address Book for Reweets 316

Appendix C Normality test results 320

Appendix C.1 Normality Test Results for Chapter 7 320

Appendix C.1.1 Analysis of bag-of-words scores 320

Appendix C.1.2 Analysis of POS tags 321

Appendix C.2 Normality Test Results for Chapter 8 322

Reference 327

List of Figures

Figure 1.1.1 The structure of background knowledge 24

Figure 2.1.1 The relationship model of tweets and the stock market 44

Figure 2.2.1 Daily tweet counts during 2007-2012 55

Figure 3.7.1 The screenshot of Derwent Capital Markets' final online auction . . . 102

Figure 4.2.1 The statistics of ticker tweets in April, 2012 120

Figure 4.2.2 The statistics of media tweets in April, 2012 121

Figure 4.4.1 Statistics of working days of each tweet crawler, April-August, 2012 . . 132

Figure 4.4.2 Daily counts of the tweet collection in logarithm scaling, April-August, 2012 133

Figure 4.4.3 Daily counts of the 153-DAY tickers in logarithm scaling, April-August, 2012 134

Figure 4.4.4 Box plot of the 10-MOST tickers tweets in logarithm scaling, April-August, 2012 135

Figure 4.4.5 The relationship of different types of tweets 138

Figure 4.5.1 Distribution of different types of GE tweets after the manual annotation153

Figure 5.2.1 Cross-correlation result of GE price as shown in Section 5.3 160

Figure 5.2.2 Example of a heat map of a set of multi-matrix data 163

Figure 5.2.3 Plot of GE prices with replaced NA values, April-August, 2012 165

Figure 5.3.1 The normal plot of price trends of the GE ticker, April-August, 2012 . 167

Figure 5.3.2 The CCF test of four prices of the GE ticker, April-August, 2012 168

Figure 5.3.3 Plot and the ACF test result of different types of GE tweet raw counts, April-August, 2012 170

Figure 5.3.4 The CCF test result of between the price and sentiment of GE ticker, April-August, 2012 171

Figure 5.4.1 Plot and the CCF test result normalised GE ticker-related tweets at, April-August, 2012 175

Figure 5.4.2 Transformed and retransformed percentage normalisation of GE tweets, April-August, 2012 177

Figure 5.4.3 The CCF test result of the percentage normalised price and the sentiment of the GE ticker, April-August, 2012 179

Figure 5.4.4 Plot of GE price normalised by a 5-step window function, April-August, 2012 181

Figure 5.4.5 Cross-correlation result of GE sentiment and price by a 5-step window function, April-August, 2012 182

Figure 5.4.6 Daily GE price trend with a 7-day moving average 184

Figure 5.4.7 Cross-correlation result of GE sentiment and price with a 7-day moving average 185

Figure 5.4.8 Cross-correlation result of GE sentiment and price with a 15-day moving average 186

Figure 5.4.9 Cross-correlation result of GE sentiment and price with a 30-day moving average 187

Figure 5.4.10 Cross-correlation result of GE sentiment and price with a 14-day moving average 189

Figure 5.5.1 Comparison of different neutral sentiment trends of GE 193

Figure 5.5.2 Comparison of different neutral sentiment trends of GE in percentage scale 195

Figure 7.2.1 Statistics of unigrams in different polarities of tweets 225

Figure 7.3.1 Score distribution comparison of the localised sentiment word list and the list by Davies and Ghahramani (2011) 239

Figure 7.5.1 Distribution of POS tags of ticker-related tweets 247

Figure 7.5.2 Five groups of the most frequent part-of-speech features in ticker-related tweets 252

Figure 8.1.1 The relationship of annotated tweets as introduced in Chapter 4 263

Figure 8.1.2 Boxplot of word count including URLs 264

Figure 8.1.3 Boxplot of character count including URLs 266

Figure 8.1.4 Boxplot of word count excluding URLs 268

Figure 8.1.5 Boxplot of character count excluding URLs 269

Figure 8.1.6 Boxplot of hashtag counts in the annotated tweets 271

Figure 8.1.7 Boxplot of cashtag counts in the annotated tweets 272

Figure 8.1.8 Boxplot of retweet counts in the annotated tweets 274

List of Tables

Table 2.2.1 Comparison of different tweet data services 52

Table 3.2.1 Comparison of Stanford final projects, 2011 (I) 75

Table 3.2.2 Comparison of Stanford final projects, 2011 (II) 76

Table 3.2.3 Comparison of Stanford final projects, 2011 (III) 77

Table 3.3.1 Statistics of the sample size based on the first approach 82

Table 3.3.2 Statistics of the sample size based on the second approach 84

Table 3.3.3 Statistics of the sample size based on other approaches 85

Table 3.6.1 Statistics of previous studies of stock prediction based on tweet data . . 93

Table 3.6.2 Sentiment word lists used in previous studies 96

Table 4.2.1 Escober's (2011) suggestions for monitoring social network 110

Table 4.2.2 List of 30 companies and their tickers in the Dow Jones Industrial Average 116

Table 4.2.3 Statistics of some breaking news Twitter accounts 117

Table 4.2.4 Final collection rate of each crawler 122

Table 4.3.2 Statistics of processed stock tweets 130

Table 4.4.1 Statistics of the collected tweets from three crawlers, April - August, 2012 131

Table 4.4.2 The relationship of the positive and negative relationship between the market trend and the investment action 143

Table 5.1.1 Sample of converted GE price data 155

Table 5.2.1 Sample data of GE sentiment 157

Table 5.2.2 Autocorrelations of series ‘X’, by lag 160

Table 5.2.3 The effect size of Pearson’s *r* 164

Table 5.4.1 Comparison of the day counts of the transformed and retransformed percentage normalisation of GE tweets at 4-8, 2012 178

Table 7.1.1 Comparison of four popular word lists in previous studies 220

Table 7.2.1 Meaningful unigrams in the top 50 frequent unigrams of ticker-related tweets 222

Table 7.2.2 The relationship of unique unigrams and total tokens 226

Table 7.2.3 The coverage of unique unigrams 227

Table 7.2.4 Relationship of one-occurrence unigrams and the whole unigrams . . . 228

Table 7.3.1 Comparison of the word types in different polarities before and after stemming by the Porter stemmer 230

Table 7.3.2 Sample of the top 20 frequent unigrams after stemming 231

Table 7.3.3 Comparison of words with top 20 keyness scores in negative and positive category 233

Table 7.3.4 Size comparison of the localised sentiment word list and Davies and Ghahramani (2011)’s list 235

Table 7.3.5 An extract of the word matrix based on the annotated GE tweets 236

Table 7.3.6 Statistics of average score of each polarity group in the GE tweets . . . 237

Table 7.3.7 Coverage comparison of the localised sentiment word list and the list by Davies and Ghahramani (2011) 238

Table 7.3.8 Summary of significance tests of three groups of sentiment scores . . . 241

Table 7.4.1 Statistics of average score of GE training and test set 243

Table 7.4.2 Summary of the best accuracy of four classifiers based on the bag-of-words model 244

Table 7.5.1 Statistics of the POS tagging results of five groups of tweets by the ARK tweet POS tagger (The original tags are in brackets.) 246

Table 7.5.2 Top 20 frequent adjectives in the ticker-related tweets 249

Table 7.5.3 Top 20 frequent verbs in the ticker-related tweets 250

Table 7.5.4 Summary of significance tests of five most frequent POS tag categories 253

Table 7.6.1 A sample set of POS features generated from annotated GE ticker-related tweets 254

Table 7.6.2 Summary of the accuracy of four classifiers based on the most frequent part-of-speech features 255

Table 7.6.3 Summary of the accuracy of four classifiers based on the use of all frequent part-of-speech features 255

Table 8.1.1 Summary of using the external features to identify the type of tweets . . 275

Table 8.1.2 Average counts of the external linguistic features 276

Table 8.2.1 An extract of the external features of the annotated GE tweets 278

Table 8.2.2 The best results of four classifiers at the stock level 279

Table 8.2.3 The best results of four classifiers at the ticker level 280

Table 8.2.4 The best results of four classifiers at the polarity level 281

Table 8.3.1 Statistics of the accuracy of each level of four classifiers 282

Table 9.1.1 Sentiment classification accuracy at the training stage based on different sets of features in previous chapters 285

Table 9.1.2 Summary of significance tests of three sets of features 286

Table 9.2.1 Sample of the training model based on the combined internal features . 287

Table 9.2.2 Classification results based on the combined internal features 288

Table 9.3.1 Sample of the training model based on POS and external features . . . 288

Table 9.3.2 Classification results based on POS and external features 289

Table 9.3.3 Comparison of the classification results based only on part-of-speech or external features 289

Table 9.4.1 Sample of the training model based on combination of the bag-of-words model and external linguistic features 290

Table 9.4.2 Classification results based on the combination of bag-of-words and external features 290

Table 9.5.1 Sample of the training model based on the combined internal and external features 291

Table 9.5.2 Classification results based on combined internal and external features 292

Table 9.6.1 The best accuracy (in bold) of each classifier based on different features 293

Table 10.2.1 Comparison of the best training accuracy at the stock level and the ticker level 300

Table 10.2.2 The accuracy comparison at the training stage 305

Table 10.2.3 Statistics of sentiment classification accuracy in previous studies (same as Table 3.6.1) 306

Table C.1.1 The Shapiro test of the overall BoW scores in NEG, NEU, and PST tweets320

Table C.1.2 The Shapiro test of the positive BoW scores in NEG, NEU, and PST tweets320

Table C.1.3 The Shapiro test of the negative BoW scores in NEG, NEU, and PST tweets 320

Table C.1.4 The Shapiro test of adjectives in NEG, NEU, and PST tweets 321

Table C.1.5 The Shapiro test of common nouns in NEG, NEU, and PST tweets . . . 321

Table C.1.6 The Shapiro test of prepositions in NEG, NEU, and PST tweets 321

Table C.1.7 The Shapiro test of proper nouns in NEG, NEU, and PST tweets 321

Table C.1.8 The Shapiro test of verbs in NEG, NEU, and PST tweets 322

Table C.2.1 The Shapiro test of word count including URLs in NSR, NTR, NEG, NEU, and PST tweets 322

Table C.2.2 The Kolmogorov-Smirnov test of word count including URLs in ticker-related and stock-related tweets 322

Table C.2.3 The Shapiro test of character count including URLs in NTR, NSR, NEG, NEU, and PST tweets including URLs 323

Table C.2.4 The Kolmogorov-Smirnov test of character count including URLs in ticker-related and stock-related tweets 323

Table C.2.5 The Shapiro test of word count excluding URLs in NSR, NTR, NEG, NEU and PST tweets 323

Table C.2.6 The Kolmogorov-Smirnov test of word count excluding URLs in ticker-related and stock-related tweets 324

Table C.2.7 The Shapiro test of character count excluding URLs in NSR, NTR, NEG, NEU and PST tweets . . . 324

Table C.2.8 The Kolmogorov-Smirnov test of character count excluding URLs in ticker-related and stock-related tweets . . . 324

Table C.2.9 The Shapiro test of hashtag count in NSR, NTR, NEG, NEU and PST tweets . . . 324

Table C.2.10 The Kolmogorov-Smirnov test of hashtag count in ticker-related and stock-related tweets . . . 325

Table C.2.11 The Shapiro test of cashtag count in NSR, NTR, NEG, NEU and PST tweets . . . 325

Table C.2.12 The Kolmogorov-Smirnov test of cashtag count in ticker-related and stock-related tweets . . . 325

Table C.2.13 The Shapiro test of word count including URLs in NSR, NTR, NEG, NEU and PST tweets . . . 326

Table C.2.14 The Kolmogorov-Smirnov test of retweet count in ticker-related and stock-related tweets . . . 326

Chapter 1 Introduction

Growing from 100 million active users in 2011 to 200 million in 2012, from 5,000 daily posts in 2007 to 400 million in 2012, Twitter is a massive social network service with a great number of users' statuses posted daily (Hepworth, 2012; Twitter, 2011d, 2012d; Weil, 2010). The language on Twitter presents a number of unique features, and in particular, "the extremely large volume of naturally occurring language is of great interest, as data, to linguists" (Zappavigna, 2012, p. 4). Twitter discourse covers various topics, some people thus regard Twitter as a useless platform as it can only provide limited useful information (Brown, 2012), but others take a contrary position. For example, many investors participated in the website Stocktwits (`http://stocktwits.com`), and used it as an important information resource for predicting stock trading. Indeed, Twitter has become increasingly important as a data source for stock prediction. As an illustration, Bloomberg Terminal[1], one of the most important and popular trading tools being used by stock traders, has recently integrated tweets as one of its data resources (Indvik, 2013). This presents a promising blueprint for using tweet data in an innovative way.

Moreover, advances in computer science and linguistics encourage human beings to extract increasing amounts of information from increasingly large sets of data. Therefore, sentiment analysis, with a focus on investigating the hidden emotion information in large-scale textual data, has become vital to both academia and industry. With a short history of only ten years, there is great potential for it to develop. This sub-branch of natural language processing (NLP) in computational linguistics has already benefited wide areas, with more and more scholars devoting time to such research. One concrete example is that some studies have applied sentiment analysis to predict stock movements based on tweet data, including Ruiz, Hristidis, Castillo, Gionis, and Jaimes (2012) and Bollen, Mao, and Zeng (2011). Indeed, stock prediction has attracted long-lasting interest, and assisted by sentiment analysis, recent developments in it have become increasingly

[1] Bloomberg Terminal is a computer system developed by Bloomberg L.P to access its finance service and data.

elaborate.

A good illustration of stock prediction based on tweets is that during January, 2013, a number of reports suggested that some tweets from hoax accounts might be the main cause of dynamic changes in the NASDAQ market (National Association of Securities Dealers Automated Quotations), and the authorities began to investigate the claims (McCrank & Gaffen, 2013; Vlastelica, Bases, & Flitter, 2013). In both cases, the hoax Twitter accounts used fraudulent identifications to publish misleading information, and soon after, the market underwent correspondingly aggressive movements. If both cases stand, they indicate that tweets, even fraudulent tweets, can have a huge impact on the stock market. Furthermore, it suggests that suspicious tweets like these can bring further difficulties to stock prediction based on tweets.

By stock tweets, this research adopts a narrow definition: a tweet that contains one or more cashtags—a convention on Twitter that uses the dollar sign and ticker symbol to introduce a ticker—and focuses on a topic relevant to the stock market. This is a widely applied concept that was first introduced by Sprenger and Welpe (2010). As an illustration, the tweet shown in Sample 1.0.1 reports the selling of the ticker General Electric (GE), so its author puts "$ge" at the end to indicate the topic. The research reported in this thesis used cashtags as keywords and classified them accordingly.

Sample 1.0.1 general electric short 16.98 $ge

This research centres on the three fundamental issues: the definition and features of stock tweets, the automatic identification of stock tweets and their sentiments, and whether including an analysis of linguistic features can improve the accuracy of sentiment identification of stock tweets. It presents the following stages of investigations: first, the investigation of the temporal correlation between the sentiment of stock tweets and the price performance of an individual quote, known as a "ticker" in the business and finance context; then, the analyses of the linguistic features of stock tweets, such as frequent words and part-of-speech features, and the use of them to train sentiment classifiers based on different algorithms, in order to set up a sentiment classification baseline; next, the analyses of the other low-level linguistic features, such as word count and cashtag count and the use of them to train different sentiment classifiers to see whether they can outperform the baseline accuracy; and finally, the combination of different linguistic features to further train the sentiment classifiers based on the previous analyses, and then the comparison of them with the baseline accuracy.

1.1 Rationale of the Research

The research mainly involves three areas: economics, sentiment analysis, and stock prediction based on tweets, but with a particular interest in the linguistic features of stock tweets (see Figure 1.1.1). The interdisciplinary nature of the research requires extra attention to the background knowledge of these relevant fields. To map out the rationale of the research, this section briefly overviews the current context of these areas. More in-depth discussions follow in Chapter 2.

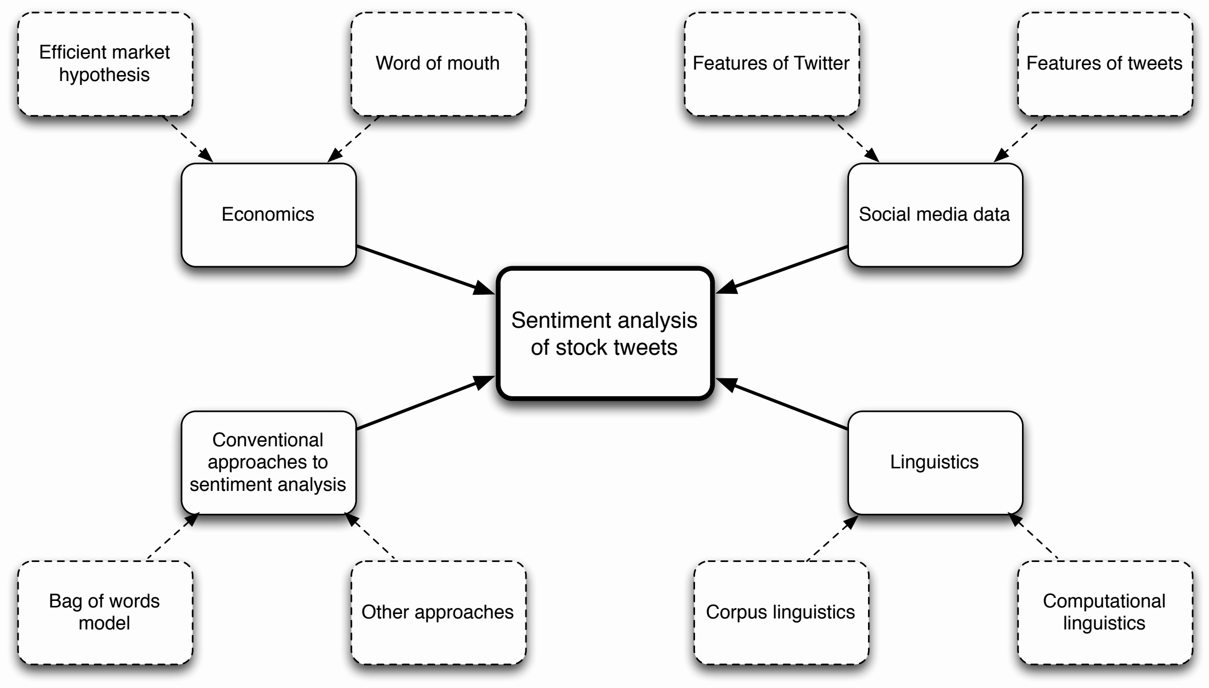


Figure 1.1.1 The structure of background knowledge

1.1.1 Economics

In economics, there are numerous discussions on the connection between available information and future developments. The efficient market hypothesis (EMH) by (Fama, 1970) is one of the most widely used theories. In addition, the other support comes from marketing, for example, the word of mouth (WoM) theory also suggests a relationship between information spread and buying behaviours. The thesis follows these two broadly adopted theories to explain the principles of this study.

The efficient market hypothesis mainly explains the influence of information distribution towards the market. According to the hypothesis (Fama, 1965a, 1965b), an efficient market has three forms: weak, semi-strong, and strong. The weak form only considers historical information. The semi-strong form also includes publicly available information. The strong form involves inside information, which is not legally available to public. In reality, the semi-strong form of the efficient market is the most common form. Based on this, researchers and business analysts make enormous effort to make predictions of the market with a hope to achieve the maximum profit. This research is no exception: It assumes that the market follows the semi-strong form of the efficient market, theorising the market as partially predictable. In other words, not all changes can be foreseen.

Developed in the 1960s (Dichter, 1966), word of mouth theory, also known as story-telling, mainly interprets the motivation of investment behaviours. At its early stage, it mainly explains the marketing mechanism, especially consumer's buying behaviours. Ditcher (1966) defines word of mouth as follows:

> When the consumer feels that the advertiser speaks to him as a friend or as an unbiased authority, creating the atmosphere of Word-of-Mouth, the consumer will relax and tend to accept the recommendation. (p. 148)

The traditional word of mouth mechanism can thus magnify the effect of information spread, because it is usually in a face-to-face form. More recently, this theory has also been used to interpret investing behaviours in the stock market (Argan, Sevil, & Yalama, 2011).

Additionally, story-telling on developing social networks may perform in a similar way to the word of mouth mechanism, and can be more robust because they involve "different kinds of multi-party, co-constructed narration" (Page, 2012, p. 13). Therefore, the discussion of social media often involves participants from different geographic positions. Moreover, the retweet mechanism on Twitter expands the duration of information spread, so participants can get involved at different time points. This further amplifies the word of mouth effect. Furthermore, the co-constructed discussion on social media is particularly beneficial to investors because the reliability of information can be cross-checked from different channels.

1.1.2 Sentiment analysis

Sentiment analysis has developed rapidly since 2001 (Pang & Lee, 2008, p. 4). Basically, it uses computational methods to assist the identification of sentiment information in texts, such as human subjective feelings and opinions. These methods are known as machine learning algorithms in computer science. More details of machine learning methods are described in Chapter 6, but for the moment, it is enough to note that these methods are a significant advance on traditional methods of textual analysis, which can only deal with limited quantities of data. Sentiment analysis emphasises enabling computers to automatically evaluate the emotional information in quantities of data that would be impossible to analyse manually.

In general, applying sentiment analysis to stock prediction is a specific and developing area, and it has attracted a number of researchers from different backgrounds. Some researchers have computer science backgrounds, particularly computational linguists, so they mainly rely on machine learning techniques to analyse data. For example, Hassan and Nath (2005), and Rao and Hong (2012) discussed using different machine learning methods, such as hidden Markov models and support vector machine, to forecast the market. They contributed to algorithm design and related areas, but their focus is restricted to numeric data calculation and not fully on the textual data. Also, scholars with a background of economics or finance emphasise the theoretical analysis, but may be limited in dealing with textual data (Sprenger & Welpe, 2010). For obvious reasons, their studies prefer to use stock models to interpret stock changes than to reveal the textual information hidden behind the model. In short, while both perspectives are useful, both are limited; in particular, both approaches lack a sophisticated analysis of the language of stock tweets. The key question that this thesis aims to address, therefore, is whether bringing a linguistic dimension into the analysis will substantially enhance this interdisciplinary area.

An extensive discussion of sentiment analysis is given later in Chapter 3.

1.1.3 Linguistic features of tweets

Currently, sentiment analysis has an obvious bottleneck in terms of accuracy, particularly when it deals with context-free data. The tweet data are particularly short and noisy, so the accuracy of automatic sentiment classification on this type of data is more subject to the consistency of distinct features than other types of textual data. For instance, Mukherjee, Bhattacharyya and Balamurali (2012) noticed that the short contents could negatively affect sentiment analysis based on supervised machine learning methods. In other words, subtle differences in the data may bring different automatic classification results. From a linguistic perspective, the main reason might be that the current sentiment analysis on tweets lacks an integration of the linguistic features of the tweets.

As discussed in previous literature, the language being used in tweets is considerably different from other domains. For example, Zappavigna (2012) indicated that the most frequent words in the HERMES tweet corpus are completely different from Davies' (2008) corpus, COCA (Corpus of Contemporary American English). However, there are only a limited number of analyses of the linguistic features of tweets, and analyses of the linguistic features of stock tweets are limited further still. Therefore, previous sentiment analyses projects on stock tweets took little consideration of the tweets' linguistic features.

However, even basic linguistic features can have a significant contribution to the analysis of tweets. For instance, Kiciman (2010) pointed out that "there is a strong correlation between language and metadata features" (p. 51) in tweet data. By metadata, he analysed the geographic location of posted tweets and the follower number of Twitter users. Following this, this research hypothesises that using a similar approach based on the low-level linguistic features could assist the recognition of the category of tweets, and as well as their included sentiment. It therefore might improve the accuracy of sentiment analyses of stock tweets because these features are more consistent than other higher level linguistic features.

In line with this, Zappavigna (2012) applied basic corpus analysis approaches, such as frequency analysis to explore her data, and considered that they could "guide the eye of the analyst to regions of meanings that are likely to be fruitful sites for close discourse analysis" (p. 192). Such an approach is critical for analysing tweets because tweet data sets are often large, and sampling such vast quantities of data is particularly challenging. Thus, analysing the basic features might be the most practical way to explore them at a macro level .

To illustrate how these low-level linguistic features might help in identifying the category of tweets or their sentiment, it is better to take a sample tweet from the data. Sample 1.1.1 is a typical example of a stock tweet. Posted on August 1, 2012, it contains 15 cashtags, which indicates 15 different tickers. The main idea of this tweet is simple: It aims to promote the watch list in the link at the end of this tweet. Therefore, this stock tweet does not have a particular focus on any specific tickers. Moreover, it does not show any explicit sentiment. Accordingly, this research considers it as a stock-related tweet, but not a ticker-related tweet.

> Sample 1.1.1 $AA, $BAC, $BK, $CSX, $EMC, $F, $GE, $GLW, $HPQ, $KR, $LNC, $MS, $OI, $TER, $VLO: Watchlist Aug 2, 2012 _THIS_IS_A_URL_LINK_

Simply counting the number of the cashtags might be limited in other cases, but if it combines other low-level linguistic feature, this approach can be more robust. Particularly, using machine learning methods to train the classifiers based on the combination of these features, it can have much more potential. Mukherjee et al. (2012) added several "light-weight" linguistic features, such as n-grams and stop words, with the bag-of-words feature to the classification model, and their results outperform the baseline accuracy.

However, there is a concern in combining linguistic features for sentiment analysis: the computing efficiency. In other words, the more elaborate the linguistic features are, the more computing power is needed, and the more potential errors there are that could be generated.

In short, the absence of linguistic features as part of current sentiment analysis method has led to a bottleneck in terms of accuracy, and the addition of linguistic features to the analysis could potentially change the situation. Therefore, bringing a comprehensive analysis on the low-level linguistic features of stock tweets is particularly necessary, and it might be a helpful way to improve the accuracy of sentiment analyses of stock tweets.

1.2 Statement of the Problem

As mentioned above, there have been a number of attempts to model different sources of data in order to predict future movements, and stock prediction have been among the most popular topics. Different data sources were used for the stock prediction: printed newspapers (Awan, 2010; Schumaker & Chen, 2009), and various electronic media, such as online bulletin boards or online forums (Mizrach & Weerts, 2009), blogs (Gilbert & Karahalios, 2009), online news (Ahmad, Cheng, & Almas, 2006; Goonatilake & Hearth, 2007; Kevin, Yang, & Hsin-Hsi, 2008; Koppel & Shtrimberg, 2006; Mizumoto, Yanagimoto, & Yoshioka, 2012), and even web search queries (Bordino et al., 2012). More recently, there has been an increase in the literature on the possibility of using social network data. In particular, Twitter provides a great possibility to study these questions.

Prediction based on social media data considers social media as a proxy for investors' discussions, as well as a source of public mood, because social media are, as Page (2012) suggested, "Internet-based applications that promote social interaction between participants" (p. 5). The rapid development of social networks provokes discussions of various topics, particularly many of them cover the market related topics, for example, products, services or even marketing (Ruiz et al., 2012). Therefore, mining social networks has been an important area of market prediction in recent years. Furthermore, Sprenger and Welpe (2010) suggested that Twitter has "the widest acceptance in the financial community" than other social media domains, so it may have richer market information than other social websites.

Specifically, Bollen et al. (2011) pointed out that a relatively positive relationship between tweet data and stock trends exists, and using tweet data can achieve an accuracy as high as 87.6%. Also, they suggested that stock prices are largely driven by new information, for instance, relevant news, rather than present and past prices. Furthermore, Sprenger and Welpe (2010) indicated that financial markets are "informationally efficient" (p. 5). In other words, market prices are able to reflect known information. Both of them demonstrated that there is a certain connection between market and information.

One thing should be borne in mind is that stock prediction based on tweets actually has two separated questions: (a). what is the sentiment in tweets, and (b). what is the relationship between the extracted sentiment and market changes. Both questions attempt to find the best accuracy. This research focuses on the first question, so the aim is to answer the following question.

> **Does linguistic analysis improve the sentiment identification accuracy of stock tweet sentiment analysis?**

This question is considered as the overarching question of the entire research, and can be extended to the following specific questions:

> Question 1. Does a clearer definition of stock tweets improve the quality of an analysis of such tweets?
>
> Question 2. Are stock tweets a linguistically distinct type of tweet? What specific linguistic features do they have?

Question 3. If stock tweets have explicit linguistic features, is it possible to automatically identify the stock tweets based on their linguistic features?

Question 4. How can a robust sentiment word list for tweet sentiment classification be designed?

Question 5. Does a more precise definition of a positive, neutral, or negative stock tweet in accordance with market values help to improve the quality of stock tweet sentiment analysis?

Question 6. How can the neutral sentiment category of stock tweets be defined and processed?

These six specific research questions construct the entire research, so after a further introduction of background information is provided in Chapter 2, Chapter 3 discusses the previous studies on these topics, and the remaining chapters try to answer above questions accordingly.

1.3 Structure

The following chapters are organised as below.

Chapter 2 discusses the empirical context of this research from two perspectives: economics and Twitter. These two areas may not be familiar enough to a linguistic audience, so this chapter introduces the relevant aspects of the two areas in order to provide sufficient background knowledge. LyricSheffield: Heading off to the launch of this at Blackwells shortly. Some seriously talented academics in here. #HbkStyle http://t.co/0AOV3b7fzA Chapter 3 first surveys sentiment analysis, including its characteristics and developments, and then reviews relevant studies of stock prediction based on tweets. Prediction based on such a data source is a new and specific area, so there have been only a limited number of studies; thus, the research reviews them thoughtfully. Finally, the chapter elaborates on the six research questions raised in Chapter 1 in detail with relevant studies.

Data are central to this thesis because it mainly relies on quantitative analysis. Chapter 4 introduces the preparatory stage of the data analysis, and Chapter 5 presents an analysis of the temporal correlation between tweet sentiment and stock price in order to discuss the feasibility of the main analysis. Chapter 6 explains the quantitative methods being used in the main analysis. The key part of the thesis is the main analysis presented in Chapters 7, 8, and 9, which used statistical analysis, machine learning techniques, and data visualisation to explore sentiment classification based on different linguistic features.

All data analyses in this dissertation use R language (R Team, 2013), which is an open source programming language with a focus on statistical computing. For the purposes of this research, the packages in R were mainly used to carry out statistical analysis, conduct time series analysis, and develop machine learning based classifiers to automatically classify data. The details of the usage of R are given in the separate methodology parts of Chapters 5 and 6.

Chapter 4 is an introduction to the data preparation. It first introduces the difficulties that might have been encountered during the collection phase, and then discusses the procedure of collecting stock tweets through the Twitter Search API. Five thresholds are focused upon in detail: collection criteria, raw tweet structure, collection design, server configuration, and collection rate. It then provides a detailed explanation of tweet data manipulation. The fifth section reports on the annotation of stock tweets about General Electric (GE), discusses the annotation criteria, and introduces the annotation procedure. The annotation applied a hierarchy classification model. In brief, it classified the raw tweets as non-stock-related tweets (NSR) and stock-related tweets. The content of the NSR tweets does not have any relationship with the stock market. The stock-related tweets are then classified as non-ticker-related tweets (NTR) and ticker-related tweets. The NTR tweets discuss topics relevant to stock, but they do not focus on any specific ticker, so they provide little help in predicting the individual ticker's performance. Finally, the ticker-related tweets are classified as negative (NEG), neutral (NEU), and positive (PST) polarities. This three-level hierarchy classification model is different from other sentiment studies in stock prediction conducted before because most of them conventionally classify NSR and NTR tweets as NEU tweets. The last section of this chapter discusses the relationship of annotated tweets.

Chapter 5 reports on the investigation of the temporal relationship between tweet sentiment and stock price changes. This is a critically important part to the study because if there is no temporal correlation or a very weak correlation, then the stock tweets have little value in predicting market changes. Different temporal data manipulation methods are used to extract the relevant information from the noisy tweet sentiment data, and then time series analysis is used to test if there is any correlation between stock tweet sentiment and price change. The result suggests that the moving average, a commonly-used smoothing method in stock analysis, can provide the best correlation between tweet sentiment and the price change. The last part compares the temporal correlation (based on the hierarchy classification annotation undertaken as outlined above) and the conventional sentiment classification result. It suggests that the hierarchy sentiment classification method used in this study can slightly improve the correlation, but it takes more effort to classify. Thus, in practice, it may not be a better choice.

Chapter 6 introduces three groups of methods applied in the main analysis: linguistic analysis, statistical analysis, and machine learning. With a focus on the linguistic features of tweets, the main analysis requires a comprehensive linguistic analysis. The first section of this chapter introduces different corpus linguistics methods, including frequency analysis, concordance analysis, and keyness analysis. To cover more linguistic features of the annotated tweets, it also includes three computational linguistic methods in this section: tweet normalisation, stemming, and part-of-speech tagging. The second part of this chapter focuses on statistical analysis, so it explains normality and significance tests, the covering Shapiro-Wilk test, the Kolmogorov-Smirnov test, the Wilcoxon rank sum test, the Kruskal-Wallis test, and the pairwise Wilcox test. The last section of this chapter introduces machine learning. It first explains the motivation of applying machine learning for the purposes of the main analysis, and then discusses different supervised learning methods, including decision tree, random forests, Naïve Bayes, and support vector machine methods. Finally, it introduces cross validation to evaluate the performance of machine learning.

As a number of previous studies used the bag-of-words model to classify sentiment of tweets, Chapter 7 first tests this approach. It is common in the natural language processing field to use a predesigned sentiment word list for training a bag of words model, so this research first investigates some of the widely used lists, and as shown later, none of them satisfy the needs of this research. Therefore, the internal linguistic features of the annotated stock tweets are analysed in order to design a more robust sentiment word list. The first internal linguistic features to be analysed are the most and least frequent unigrams; however, neither of them are found to be distinct or stable enough to indicate sentiment. Thus, a stemmer is then employed to conflate the tweets, and keyness analysis is used to classify their polarity. According to this, a localised sentiment word list is then generated. Compared with other sentiment word lists, this localised list provides a better coverage and more balanced positive and negative words. Next, four supervised machine methods—decision tree, random forests, Naïve Bayes, and support vector machine—are used to train sentiment classifiers based on the above features. The classification based on the bag of words model achieved a reasonably good result, so it is thus regarded as the baseline accuracy. However, these four classifiers are dependent on the topic. Therefore, the part-of-speech tags of the stock tweets are then annotated and their features analysed. The part of speech of these annotated stock tweets mainly occur in the following categories: *common noun*, *adjective*, *verb including copula*, *auxiliaries* and_ preposition or postposition, conjunction categories_. Four out of five of them have a significant effect on the difference in terms of the polarity. Then the classifiers are trained based on the part-of-speech features; nevertheless, they had a poorer performance.

To reduce the topic dependence of sentiment classification, the research therefore focuses on the external linguistic features as reported in Chapter 8. The first analysis applied different statistical methods to examine the external linguistic features, such as word length, character length, cashtag, hashtag, and retweet, to see if these features help in identifying stock tweets and included tweet sentiment. These analyses indicate that stock tweets have a significant effect of difference at different levels. According to this, it is possible to develop an automatic classifier to identify the stock tweets, in order to reduce the noise for the later sentiment classification. Each of the classifiers present a perfect performance of identifying stock-related tweets and a moderate performance of identifying ticker-related tweets. However, these external linguistic features are not sufficient to differentiate different polarities of stock tweets.

Chapter 9 combines the sentiment classifications described in the previous two chapters, in order to understand if combining linguistic features can improve the sentiment classification accuracy. It reports on four experiments. The first experiment combines the bag-of-words model with part-of-speech features, the second one combines the part-of-speech features with external linguistic features, the third one combines the bag-of-words model with the external linguistic features, and the last one combines all features together. Although each supervised learning method achieves the best accuracy based on different combined features respectively, all best accuracies outperform the baseline accuracy. The best accuracy of 0.685 is yielded by the random forests classifier based on all combined features in the last experiment. This shows that combining linguistic features can improve the accuracy of sentiment classification.

Finally, Chapter 10 discusses the research results of the main study, draws conclusions, and provides an outline for future development.

Chapter 2 Background

Stock prediction has a long history, and has involved a number of theories and techniques. How to use different information sources to predict the market has been a long lasting research question, and in particular, how Twitter, as one of the dominant social media platforms, can be used, has recently attracted more attention from both researchers and investors. Some relevant research was carried out, such as Bollen et al. (2011) and Sprenger and Welpe (2010). However, this is a highly interdisciplinary area, so it requires extensive background knowledge from these areas. Therefore, this chapter aims to provide background knowledge for the entire research by covering the two main areas: economics and Twitter.

In terms of economics, it discusses five questions: (a) Is the market predictable? (b) What were the available data sources for stock prediction? (c) Is social media a good data source for stock prediction? (d) What is the relationship between tweet sentiment and the stock market? (e) What doubts are raised by the main criticisms of stock prediction based on tweets? The next part of this chapter then introduces Twitter's characteristics, including features of the Twitter platform, the data releasing policy of Twitter, the historical development of Twitter, and technical features of tweets. Finally, the chapter discusses whether Twitter is a good source for stock prediction, mainly investigating the coverage of Twitter and influence of Twitter on stock market.

2.1 Economics Theories Related to Stock Prediction

> We're not sure what joke to make here. Finally a business model for Twitter? A sure sign of a bubble? But where? On Twitter? Or on the stock market? Or both? (Gobry, 2010)

A number of theories support the developing prediction techniques, particularly from economics and marketing perspectives. These theories, to some extent, demonstrate the feasibility of stock prediction, and the most common ones include the efficient market hypothesis and word of mouth theory. As briefly introduced in Chapter 1, these two theories are completely different, and therefore provide support for different aspects to stock prediction. Efficient market hypothesis focuses on the relationship between the market and information, mainly the market response to information (Fama, 1996); therefore it has often been used to explain the motivations of investments. On the other hand, word of mouth theory concentrates more on the relationship between advertisers and receivers. In the stock market, both advertisers and receivers may participate in the investment, so the second theory has been used to understand the interaction among different participants. Thus, these two theories cover different aspects of stock prediction.

While these theories have received much support during the development of stock prediction, a number of doubts still remain. Briefly, in the present research, for the purpose of the primary questions are as follows:

1. Is the stock market predictable? And to what extent?
2. Is it feasible to use social media to predict stock trends?

Thus, the following sections discuss these two questions by introducing relevant aspects of the above two theories and explore the relationship between information and the market. Finally, there is a discussion of some doubts upon the possibility of stock prediction, with a particular focus on the results of the recent tweet-based prediction studies.

2.1.1 The predictable market

Whether the development of stock is predictable is a frequently discussed question in related areas. One of the mainstream theories is Fama's efficient market hypothesis (EMH) (Fama, 1965a, 1965b, 1970, 1996). As briefly introduced in Chapter 1, EMH defines the efficient market in three forms: weak, semi-strong, and strong. In particular, the strong form of an efficient market, namely, an ideal market, should meet three conditions as below :

1. There are no transactions costs in trading securities;
2. All available information is costlessly [*sic*] available to all market participants;
3. All agree on the implications of current information for the current price and distributions of future prices of each security (Fama, 1970, p. 387).

If these three conditions are achieved, the market movements will reflect the available information effectively. In other words, the preceding information can indicate the future changes in the market. Thus, the strong form of an efficient market is a perfectly predictable market.

According to Fama (1965a, 1965b), the key factor of an efficient market is the availability of information: the market is only fully efficient when information is completely free to all market participants, but this 'completely free' situation is rare in reality. Thus, the strong form of EMH is hard to achieve in the real market, unless the investor takes risks by doing illegal inside trade. It suggests predicting future movements is notoriously difficult because it can only be on the basis of limited information. Otherwise, if a market is fully predictable, then it loses the possibility of being predicted, because any changes will be foreseen or known.

Moreover, due to the inefficiency of information distribution, disparities between the actual price and expected price often exist. Overestimation and underestimation of the price remain balanced on a macro level: In Fama's opinion, these disagreements are 'systematic', neither random nor coincidental (Fama, 1965a, 1965b). The systematic disagreements in the market trigger investors to make predictions, so the estimated price moves towards the actual price. Furthermore, he suggested that investors' behaviours would 'neutralise' the difference between the actual price and the estimated price (Fama, 1965a, p. 65), so the efficiency of the market may not be noticeable. Fama

(1965b) also noted that any overestimation or underestimation is independent, which suggests that changes in the estimated price are influenced only by the preceding occurrence of relevant events, but not by the preceding prices. With the assumption of independence, the market is thus defined as a random walk market (Fama, 1965a).

Yet some more recent studies have posed an opposite opinion against the EMH theory as that ‘the qualitative information is not reflected fully and instantly in market prices’ (Sprenger & Welpe, 2010, p. 6). For instance, a more detailed opposed voices can be found in Malkiel (2003)‘s discussion, who summarised several typical contrary examples of the efficient market, and therefore denied the efficiency in the market. Nevertheless, he still held the opinion that ’pricing irregularities and predictable patterns in stock returns can appear over time and even persist for short periods’ (p. 33).

Responding to these questions, Sprenger and Welpe (2010) further suggested that such failures may be caused by the unavailability or inefficiency of computational linguistic methods during the time when those studies were carried out. In other words, EMH theory could still be valid if information can be processed properly. In addition, it could also be valid if other sources of information can be used to predict the market changes. Based on this inference, this thesis is to use data from other source–Twitter to explore if the accuracy of a more advanced computational linguistic method–sentiment analysis being used in stock prediction–can be improved.

Most importantly, the focus of this thesis is not to discuss the validity of using EMH theory to predict the market changes. Simply put, the thesis accepts the assumption that available information can reflect the market changes, so the market changes can be predicted by the available information. This follows the previous studies in using available information to predict the market, such as Bollen et al. (2011)’s and Sprenger and Welpe (2010)’s.

2.1.2 Data sources for stock prediction

Before the emergence of social media, academics and industries have attempted using various data sources to predict market changes, including corporate annual reports, printed newspapers, online bulletin boards, online forums, blogs online news, and search queries.

Historically, it is not unusual to see that a number of studies have focused on more closely relevant data sources, including corporate annual reports and preliminary earning announcements, in an attempt to understand the future movements of a particular company. A typical example can be seen in Patell & Wolfson (1984)'s research, which discovered that news of earnings and dividend announcements exerted an influence on stock prices over different spans of time. Another recent example can be found in F. Li (2006)'s research, which investigated words related to risk or uncertainty in corporate annual reports. One of the major findings in this study was that there seems to be a connection between risk sentiment expressions and negative future earnings.

Printed newspapers have been the all-time favorite for stock prediction, and the recent availability of the electric version of the traditional printed newspapers has made this collection much easier to be accessed. For instance, the typical applications can be found in (Awan, 2010; Schumaker & Chen, 2009). The first one included both news outlets and news wire services as the components of their data sources, with an attempt to examine the instant effect of the available information towards the market. The second one used Google search engine to collect data from news wire services, in order to design a more robust sentiment classifier for stock prediction.

Lately, Internet has been more popular in the stock prediction literature with its latest popularity in public. Online bulletin boards or online forums were among the first choices for scholars and investors to carry out stock prediction. Taking Mizrach and Weerts (2009)'s work as an illustration, they analysed data from a public online chat room–Activetrader, which was regard as a 'cooperative venture' by the researchers (p.268), and their finding suggested that the participants using available information in this chat room were profitable than the others.

Later on, other online data sources have gained more popularity among the academia and industry. For example, the emergence of blogs has soon attracted researchers' interests, and Gilbert and Karahalios (2009) were among the first to explore the influence of this data source towards the market. Their result suggested that the anxiety expressed in the blog data from LiveJournal can predict the downward pressure on the S&P 500 index.

Apart from the electric version of printed newspapers, online-only news has also become much more popular. For instance, Ahmad et al. (2006) used the online news data from two Reuters corpora in English and Arabic and a Chinese corpus to identify the frequent patterns in financial reports, in order to understand the changes in the market. Another example is that Mizumoto et al. (2012) used online news as the data source to train a semi-supervised classifier, with an attempt to make a more robust dictionary for stock prediction, and they built a polarity dictionary automatically and were able to determine the polarity of stock market news by using the their dictionary.

Interestingly, even web search queries were used to predict the market changes. Mao, Counts, & Bollen (2011) conducted a comparison of the stock prediction power among different data sources, and their result suggested that 'weekly Google Insight Search volumes on financial search queries do have predictive value' (p. 1). This was confirmed by Bordino et al. (2012), who used the search queries from Yahoo to predict the NASDAQ market changes, and claimed that trading volumes correlated to the amount the daily search queries.

Even so, as suggested in the opposite literature towards the EMH theory, the theory may not be as efficient as described. Although most of the above studies suggested positive results in favor to this long-lasting theory, it is worth using a set of more real-time data to examine. Therefore, this thesis first uses Twitter data to verify the correlations between the information available on

Twitter and the market, providing support for further analysis.

2.1.3 Using social media as the data source

The recent uptake of social media makes it an important data source. This section, in a more detailed way, discusses reasons for using social networks as a data source for market prediction.

The main reason for using social media in market prediction is based on the word of mouth theory. As briefly introduced in Section 1.1.1, word of mouth theory (WoM) emphasises interpersonal communication in marketing, while the emergence of social media extends interpersonal communication geographically and temporally. In particular, discussions of the market on social media fit the word of mouth theory better than the traditional word of mouth marketing, because investors aim to pursue the maximum outcome from **exclusive information** available on social media immediately.

The most important part of this theory is that, if consumers consider that a recommendation is from a friend or an authority, they would believe it without any suspicions. This type of communication is frequently applied in practice intentionally or unintentionally, which results in the improvement of the efficiency of marketing. Particularly in the stock market, people tend to obtain 'inside information', even sometimes only rumours, because they believe that this type of information is exclusive and helpful to capture future price changes. With the benefit of this type of exclusive information, they believe that they are able to maximise the outcome of their investments.

However, it cannot be ignored that the word of mouth theory has two directions: One is positive, as discussed above, and the other is negative. Some research has focused on the negative aspect of WoM. For example, Richins (1984) raised the concept of 'Systematic NWOM' where *N* stands for negative. According to this, external factors, such as negative feedback, also play an important role in word of mouth. Moreover, the Technical Assistance Research Program (1981) suggested that negative WoM is more frequent than positive and may be able to outweigh the positive direction of word of mouth. This may be more common in stock market discussions than in other contexts, because investors become alert when they receive negative or uncertain information (Xue Zhang, Fuehres, & Gloor, 2010). This was supported by Bollen et al. (2011), who pointed out that calmness has the closest relationship to DJIA's movements.

The other form of support comes from behavioural economics. For example, Barber and Odean (2007) have argued 'that individual investors are net buyers of attention-grabbing stocks, e.g., stocks in the news, stocks experiencing high abnormal trading volume, and stocks with extreme one day returns' (p. 785). In addition, Chan (2003) discovered a major difference in terms of return patterns between two sets of stocks with and without news: the stock returns in the set with news show decreases following bad news. This was confirmed by Leinweber and Sisk (2010)'s work, which further suggested that these kind of effects are 'larger for smaller capitalization firms' (p. 3). These studies again provided support to the possibility of using information to predict changes in the market.

Based on these, this research therefore uses Twitter as a data source to evaluate how sentiment analysis for stock prediction can be improved with the addition of linguistic power.

2.1.4 The relationship model between tweets and the stock market

To understand the relationship between tweets and market, it is better to use a simplified model to illustrate (see Figure 2.1.1). This illustration consists of two parts: speculation on the left and share price on the right. They have a bilateral relationship: According to the EMH model, the market is reflected by information, and it is affected by information.

Briefly, in the left part of the illustration, the model considers the mass media and social media

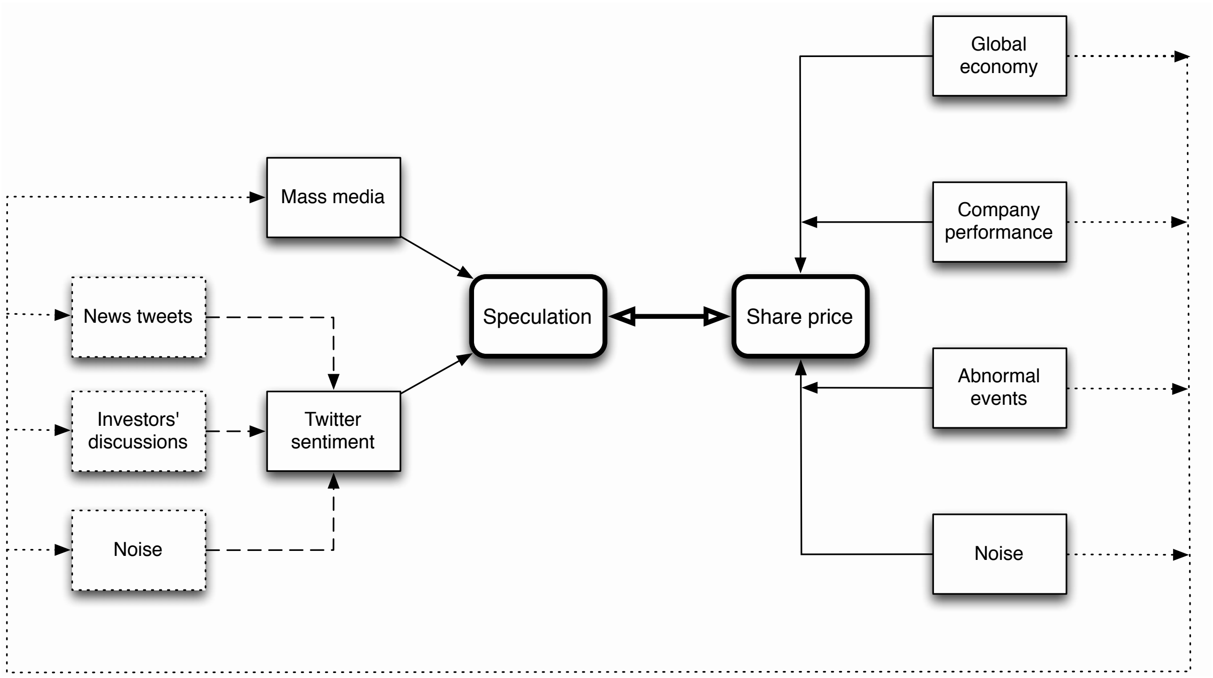


Figure 2.1.1 The relationship model of tweets and the stock market

as the main information sources. The research focuses on social media. Social media covers a number of platforms, but the research only uses Twitter, specifically, the sentiment extracted from relevant stock tweets, as representative of social media. The mass media is not the focus of the research, so the model ignores its influence. On Twitter, there are two major groups of tweets relating to the market: news tweets and tweets of investors' discussions, which are greatly influenced by the mass media. Certainly there is a large amount of noise in tweets, such as spam tweets.

The right part of the illustration uses share price to represent the market. Share price has four main sources of influence: global economy, company performance, abnormal events and noise. It is mainly formed by company performance. The global economy is an external factor: Most listed companies are multinational corporations, so they receive more influence from the global economy than the local economy. Following this, the local economy is neglected. Abnormal events, such as natural disasters or law cases, can have a particularly huge impact on share price. Similarly, there are various sources of noise, such as rumours, that may have influences on performance as well. These four elements can have a direct influence on price; or they can first be reflected by the mass media or social media (as shown by the dotted lines), and then impact upon share price accordingly.

Based on this simplified model and previous studies such as Bollen et al. (2011) and Sprenger and Welpe (2010), this thesis assumes that a correlation between tweet sentiment and stock movements exists (an extensive discussion of the temporal correlation between the Twitter sentiment and stock price is presented in Chapter 5). Therefore, it focuses on improving the accuracy of the identification of the tweet sentiment through the addition of linguistic analysis.

2.1.5 Some criticisms of stock prediction based on tweets

Some researchers have criticised and cast doubts upon the feasibility of stock prediction, especially recent developments in prediction based on tweets. Not only have academics expressed considerable doubts, but industry experts also have their suspicions. In particular, plenty of reports in the mass media questioned stock prediction. Some typical examples are discussed below.

Ingram (2011) argued that there might be many spammers that can affect the prediction accuracy. Theoretically, this could seriously undermine the development of stock prediction. However, Finger and Dutta (2013) pointed out that there was a significant drop in spam tweets since 2012: from 11% in August, 2012 to 1% in February, 2013. Thus, the possibility of 'polluted' content affecting prediction accuracy on Twitter is lower than before. In this sense, Ingram's assumption is less reasonable. However, considering the possibly high frequency of polluted content, the current research will design a hierarchical system to filter out irrelevant tweets, aiming to increase the data quality in the later sentiment analysis. This is reported in full in Chapter 4.

Similarly, Zweig (2011) also has concerns about this prediction model:

1. the future is the realm of surprises; no one, no matter how expert, can reliably foresee what will happen and how people will react to it.
2. Experts are inconsistent in how they analyse complex data.

The first of these doubts is hard to deny. However, it does not directly relate to stock prediction itself, and only concerns the decision based on the prediction. EMH also discusses this question. For example, Timmermann and Granger (2004) suggested that the early EMH model has an obvious shortcoming that does not cover the influence of investors' decisions. Therefore, they defined the *efficient market* as:

> An efficient market is thus a market in which predictability of asset returns, after adjusting for time-varying risk-premier and transaction costs, can still exist but only 'locally in time' in the sense that once predictable patterns are discovered by a wide group of investors, they will rapidly disappear through these investors' transactions (p. 21).

According to this, the stock market is predictable, but only within a limited time span and specific environment. Such a prediction chance exists in an open markets, but depends on investor's evaluations. Additionally, the market may be insufficient at developing new technology. Hence, Zweig's (2011) argument can be explained: With recent development in stock prediction based on tweet data, the market might not be simultaneously efficient because investors who use the new technology may not be familiar with the tool. However, gradually, they will handle such a tool more effectively, so they may capture more predictable opportunities. Thus, the market can be regarded as more efficient. Adopting new tools may result in the loss of some prediction chances for some investors at the beginning, but this does not mean that the market itself becomes insufficient or unpredictable.

Regarding Zweig's (2011) second doubt, the current approaches to sentiment analysis largely rely on computer applications, with the help of technologies such as machine learning. Thus, the results can be more consistent than would be by any analysis conducted by human beings. Hence, this doubt is less relevant than once was.

Another main doubt concerns the effectiveness of stock prediction based on tweet data. A representative voice here is that of Paul Nolte (Morgan, 2010):

> These kinds of studies are little more than data mining cherry picking information that creates a pattern, then rationalising the results after the fact.

This is a fundamental question, but it ignores the predictability of an efficient market as discussed earlier in this chapter. According to Fama (1965a, p. 56), 'the successive price in individual securities will be independent', so there is no such a pattern in the market trend. However, no pattern does not mean not predictable. On the contrary, under the random walk theory, an efficient market is full of 'rational, profit-maximizers actively competing, with each trying to predict future market values of individual securities' (Fama, 1965a, p. 56). In this sense, he indicated that the efficient market is predictable.

These doubts seem lack of a comprehensive understanding of current mainstream economics theories, so they become skeptical when a full reasoning is presented.

2.2 Introduction of Twitter

Twitter is the data source used in this research. Each post, or tweet, has a 140-character limit, which is the most well-known feature of this type of data. To further discuss whether Twitter is a good data resource for stock prediction, this section introduces the features of Twitter and tweets in detail. As a relatively new type of social media, it is useful to provide a comprehensive and accessible introduction to Twitter before proceeding any further; therefore, this section starts by discussing the following aspects of Twitter: its conciseness, its real-time characteristics, its free-to-use policy, and the wider media ecosystem in which Twitter participates. In addition, as Twitter's data releasing policy is unique, this section discusses its details of it from two perspectives: free tweet data and charged tweet data. The rapid development of Twitter also raises major issues on the question of how to sample tweet data; therefore, this section briefly introduces the history of Twitter's development as well. Furthermore, tweets contain a number of unique features, so the next section introduces such features, including @ mentions, RTs and retweeting, hashtags, cashtags, URL links, meta information, protected accounts and time. Finally, regarding the relationship of tweets and the stock market as introduced in Section 2.1.3, the last section discusses the coverage and influence of Twitter on the market.

2.2.1 Features of Twitter

Twitter, as one of the largest and most popular social network sites, has the following features: real-time updates, a free-to-use policy, and a well-built ecosystem.

Real-time updates Bollen, Pepe and Mao (2009) pointed out that tweets are 'necessarily associated with a specific moment' (p. 8), so they regarded each tweet as 'a microscopic, temporally-authentic instantiation of sentiment'. This is supported by Zappavigna (2012), who defined *real-time* as 'web content [that] is streamed to users via syndication' (p. 4), providing 'a semiotic world in which users have almost immediate access to what is being said in their social networks at any given moment' (p. 4). Benefiting from its own innovation, Twitter can present a considerable number of real-time tweets simultaneously, which has never been achieved before. For example, both Sprenger and Welpe (2010) and Oh and Sheng (2011) found that the post time of stock tweets corresponds to the market opening time well, so this real-time information is increasingly frequently used by financial professionals to predict market trends.

Depending on the user scale, Twitter receives thousands of tweets per second at peak. For instance, during 2011, the highest peak occurred at August 28, when the MTV (Music Television) Video Music Awards were held. The TPS count (tweets per second) was 8868 (Twitter, 2011c). The enormous amount of instant data bring extreme difficulties to Twitter's operations as the huge amount of highly unbalanced incoming data burdens servers severely. Suffering from many outages at the beginning of its development, Twitter can well handle such peaks now. Accordingly, presenting real-time contents is much improved as well, and many studies or services based on this feature have become available; for example, some researchers used Twitter to predict earthquakes or tsunamis, such as Doan, Vo and Collier (2012) and Earle, Bowden and Guy (2012).

Free-to-use policy Free usage is a common characteristic of social network services, and Twitter is no exception. Using Twitter is free, which is a good selling point to attract users. In addition, Twitter began to release tweet archives to its users in December, 2012 (Vandor, 2012), so individual users could retrieve their own tweets without any cost, which had been not possible earlier because access was restricted to the most recent updates only. However, Twitter began to introduce promoted tweets in April, 2010 (B. Stone, 2006). This service, which is similar to charged advertising, mainly focuses on corporate users.

Ecosystem The concept of 'ecosystem' originates from biology, where it means 'a biological system composed of all the organisms found in a particular physical environment, interacting with it and with each other' (Dictionary, n.d.). And for Twitter, this concept mainly has three interpretations: open source, open API (application programming interface) and the open business model.

From its foundation, Twitter was 'built on open source software' (Pass, 2009). Open source is a popular concept in the software industry, which generally means that the product or service is open and available, and can be used and shared for free. Twitter itself not only uses but also develops many open source products; for example, they released Bootstrap 2.1 (`http://blog.Twitter.com/2012/08/bootstrap-21-and-counting.html`) in August, 2012 for improving the user interface (UI) design of web pages (Otto, 2012). The open source strategy significantly reduces costs for Twitter, which results in a better free service and more popularity among its users.

APIs are a popular way to release data in the current Internet. Generally, a website uses an API protocol to share its data, and the developers can follow the protocol to obtain the needed data, instead of crawling the entire dataset from that website. Thus, it can improve the efficiency of data release and retrieval. Considering the sheer quantity of data, Twitter has applied the API policy since its birth. The details of Twitter's API policy will be introduced in the next section.

APIs also have brought an open business model to Twitter. Integration with other platforms or products via APIs makes sharing much easier. For example, if a user reads a news story and shares it on Twitter by the built-in sharing function on the news site, it may attract more Twitter users to read this story. This is a win-win situation: the easier sharing makes the interactions on Twitter more frequent and efficient, while also bringing more traffic to the news site. Such a sharing function is not rare, and according to Twitter's Chief Executive Official Biz Stone, the API received more than 10 times the amount of traffic than the main site itself received in 2007 (Ammirati, 2007). This suggests that the APIs have brought enormous business opportunities to Twitter.

Furthermore, according to Twitter, from 2010 to 2011, more than 500 million dollars have been spent on the ecosystem (Twitter, 2011b). This suggests that Twitter is confident on this business model.

2.2.2 Data releasing of Twitter

Twitter provides both free and charged tweet data releasing services. The free tweet data are mainly released by the REST APIs, while the charged tweet data, named Firehose data, are mainly released by three data retailers: Topsy (`http://topsy.com`), GNIP (`http://gnip.com`) and Datasift (`http://datasift.com`) at the time of writing (Williams, n.d.).

Free Tweet data REST (Representational State Transfer) is an 'architectural style for distributed hypermedia systems' introduced by Fielding (2000, p. 76), and it is one of the core protocols of HTTP 1.1. REST API, short for RESTful API, is a web API based on the REST protocol, which has been widely used for transferring data under the HTTP protocol (Hypertext Transfer Protocol). The REST API has four main methods: GET, PUT, POST and DELETE. Twitter mainly uses the REST GET method to release part of its data via two APIs without charges: the REST GET Search API and the Streaming API.

The REST GET Search API is a straightforward way to obtain data as it 'returns relevant tweets that match a specified query' (Twitter, 2012c). The default data format of this API is JSON (JavaSript Object Notation), which did not need authentication in Twitter API version 1.0. However, it only releases 'an index of recent tweets' with a time range of '6~9 days' (Twitter, 2012c), which suggests that the data are not completely real-time. In addition, it has an unclear rate limit, which depends on 'the complexity and frequency' (Twitter, 2012c). The recent change in API version 1.1 requires authentication, and the rate limit is set to 180 calls per 15 minutes (Twitter, 2013b, 2013c).

The Steaming API gives 'low latency access to Twitter's global stream of tweet data' (Twitter, 2012f). The major difference between the Streaming API and other REST APIs is that the Streaming API provides continuous real-time data, which requires a persistent HTTP connection. It also provides data in the JSON format and needs authentication. There is a clearly stated rate limit of the Streaming API of 150 times per hour. The FAQ (Frequently Answered Questions) page suggests only ‘a small fraction of the total volume of tweets at any given moment’ will be released through the Streaming API (Twitter, 2013a). Nevertheless, ‘a small fraction’ of data from the Streaming API is far larger than the data from the REST Search API.

Charged Tweet data From March, 2010, Twitter started to provide charged data, named as Firehose data, which derived from the Streaming API (Sarver, 2010).

GNIP was the first company to provide the Firehose data. Currently it provides various types of tweet data: Historical PowerTrack for Twitter provides all data since 2006 (Gnip, 2012), Twitter PowerTrack provides full access to real-time coverage data (Gnip, 2013), and Detahose provides at least 10% of the real-time volume data (Gnip, 2013). DataSift and Topsy are the other two major tweet data resellers, providing similar data products.

Indeed, these charged data provide a more comprehensive view of tweet data, but they are expensive. For example, Datasift charges $0.1 per 1000 tweets (Datasift, 2012). These real-time data include a large proportion of irrelevant data, so the cost of obtaining relevant data would be much higher.

Using Twitter as a data resource Considering the complexity of the data and the budget, the research chooses the Search API to collect data (see Table 2.2.1).

Table 2.2.1 Comparison of different tweet data services

Type of data	Price	Amount	Real time	Rate limit	Authentication
Rest Search API	free	very limited	No	Yes, but unclear	No in API 1.0, Yes in API 1.1
Streaming API	free	top to 1%	Yes	Yes	Yes
Firehose data	charged	can be full access	Both	No	No

Using Twitter as a data resource to predict stock changes has at least four advantages.

First of all, the data are free to use, which is the greatest advantage of Twitter. On Twitter, information is open to any users, so theoretically, if users follow the ‘right person’ on Twitter, they can obtain ‘the right information’ immediately without any cost. Moreover, the followership, specifically referring to the relationship of following on Twitter, has no restriction, so users can follow or be followed freely. Thus, this environment can, to a large extent, satisfy both the second and third condition of the strong form of the efficient market hypothesis that all available information is available to all market participants, and all agree on the implications of current information for the current price and distributions of future prices of each security. Although Twitter has made its data policy much stricter, which makes it much harder to obtain data than before, it is still possible and legal to retrieve a small portion of data from the platform without any cost.

Then, tweet data are also well-structured. As will be shown in Section 4.2.2, tweets are presented in the JSON format (JavaScript Object Notation), which makes the further manipulation and analysis easier. Tweet data contain rich meta information. Unlike the information presented to general users, tweets in the JSON format contain more comprehensive information for each tweet. For example, it contains the post time, which is critical to this research.

Next, the machine-readable data makes the search more convenient, so it is easier to involve discourses in different geographic regions. boyd (2011 *sic*)[2] defined this as the ‘searchability’ of Twitter (p. 46). For example, an American Twitter user can participate in a discussion of stock movements of the London Index even though the event itself happened while he was asleep. Although Zappavigna (2012) indicated that ‘the searchability is particularly useful for linguists collecting particular kinds of discourse’ (p. 6), this is only partly true, because it is neither easy to collect tweet data from Twitter, nor easy to carry out quantitative analysis on such a huge scale of data.

Moreover, the last two features, from a macro perspective, make the discourses on Twitter dis-

[2] boyd uses lowercase to spell her name in her publications.

tinctly different from traditional discourses. The traditional discourses involve a specific number of participants, but the discourses on Twitter are open to any Twitter users, so it involves potentially unlimited participants. Also, traditional discourses are usually restricted by the geographical location, while the discourses on Twitter can overcome this restriction.

2.2.3 The development of Twitter

Twitter has quickly evolved since it was first launched. The official blog of Twitter (`https://blog.twitter.com`) has used different measurements to evaluate its growth: user numbers, count of active users, counts of daily tweets, peak TPS (tweet per second), average TPS, and peak TPM (tweet per minute). Among these measurements, the count of daily tweets is the most frequently mentioned as shown in Figure 2.2.1 (Garrett, 2010; Hepworth, 2012; Thau, 2010; Twitter, 2012e; TwitterEng, 2011; Weil, 2010). In this figure, each point represents the daily tweet counts with the blog post title that mentioned it. According to this, it is obvious that, during this five-year period, Twitter has maintained a high rate of growth speed since the beginning.

The main concern brought by this immense and rapidly increasing amount of data is how to sample the tweet data. Twitter's restrictive data releasing policy seems to provide a controversial answer: first, they only release a small portion of data through APIs, either the Search API or the Streaming API; second, these data are randomly sampled by Twitter's algorithms, but there is no available documentation about these algorithms. This means that collecting data through Twitter's API does not leave many choices, but it completely depends on Twitter (more details can be found in Section 4.1.2).

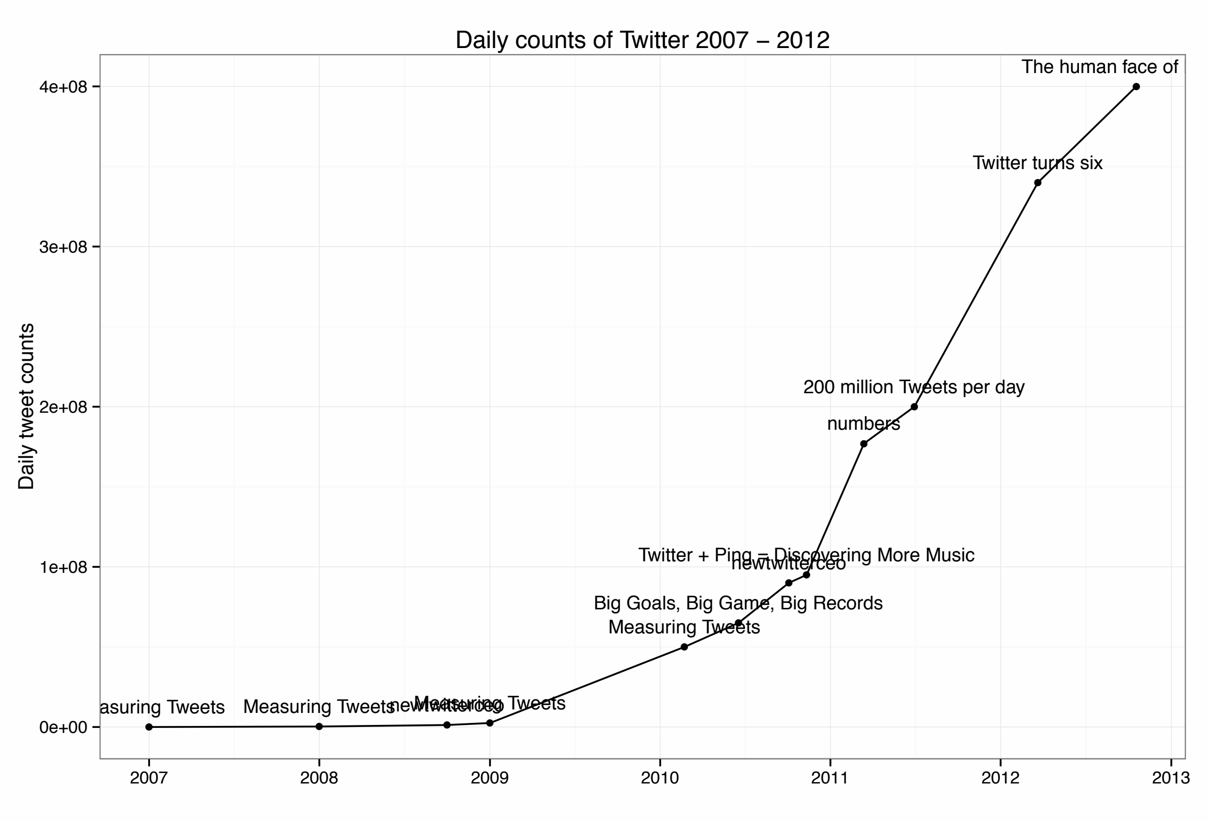


Figure 2.2.1 Daily tweet counts during 2007-2012

2.2.4 Features of Tweets

Page (2012) pointed out that 'the narrative dimensions of Twitter stories are shaped by the emergent, collaborative character of social media forms that are heightened in the specific environment of Twitter itself' (p. 116). Specifically, the specific environment greatly involves the length limit of tweets. Due to its restrictive nature, Twitter has developed a number of conventions. Compared to other written materials, tweets have a number of unique features. Here are two sample tweets:

> Sample 2.2.1 BBC News team detained and beaten up in #Libya by Colonel Gaddafi's security forces - full story in text and video http://bbc.in/erETP5

> Sample 2.2.2 Yes, use on business documents RT @elyssa_rae: @goldkorn Out of pure curiousity, do you really have a legal Chinese name?

These two samples contain three main devices used on Twitter: @ mention, RTs (retweets), hashtags, and as well as some other features that do not usually occur in other written domains. The following sections introduce these features in turn.

@ mentions In a tweet, a user name often follows the @ symbol to indicate a particular user; for instance, *@goldkorn* in Sample 2.2.2. This is called an @ mention on Twitter. In the two cases above, the @ mention increases the user's exposure on Twitter (Sprenger & Welpe, 2010). Depending on different position of this mechanism, it has different functions as a 'deictic marker' (Zappavigna, 2012, p. 34), and these functions can be summarised as below:

1. If it occurs at the beginning of a tweet, this is usually a conversational tweet or reply tweet, so *@USERNAME* functions as addressing that user. This tweet can only be seen by the user being replied to, and the followers who follow both users.

2. If it occurs in the middle of a tweet, it often has the function of mentioning some users.
3. If it occurs at the end of a tweet, it usually has more than one addresses, so it functions as forwarding the tweet to those users.
4. Alternatively, there is an increasing tendency of putting a full stop at the beginning. As discussed above, a reply tweet can only be seen by the user being replied to and their mutual followers. However, in some circumstances, the author not only wants to reply to that user, but also wants to broadcast the tweet to a broader audience, so this strategy is applied. Putting a full stop at the beginning interrupts Twitter's automatic user reply identification, so it is not regarded as a reply tweet by Twitter, and can be read by other followers. Adding a full stop at the beginning is not only hardly noticeable, but also space-saving.

Apart from using as an addressing mark, the @ symbol can also be used to indicate a place, or used in an emoticon. These types of uses can be found in other domains, so they are not discussed here.

Retweet and retweeting RT is the abbreviation of *retweet*. On Twitter, *retweet* can be used as either as a noun or as a verb, referring to the republishing or repetition of another person's tweet. Basically, retweets have two forms:

1. Users simply republish the original post without adding their own comments. When users repost a tweet, they simply copy the original tweet, so the content of the retweet is same as the original tweet.
2. Users cite the original tweet, and add comments at the beginning or end, so it is regarded as a **quoted tweet**. In some cases, the original content might be cut; this is considered as an **omitted tweet**.

Both forms were invented by users, and then have been adopted by Twitter. Applying this mechanism can 'significantly amplify the reach of a tweet' (Zappavigna, 2012, p. 36), so it is also an important convention on Twitter.

Hashtags On Twitter, a hashtag is an important mechanism to provide **searchable talk**; users therefore can 'bond around particular values' (Zappavigna, 2012, p. 1). For instance, *#Libya* in Sample 2.2.1, referring to a theme, works as a signal word. Twitter regards a hashtag as a special word: if a user clicks a hashtag, it will redirect to a search page where the hashtag is the keyword. Twitter indexes hashtags separately, which therefore classifies tweets more efficiently. Moreover, users can search and follow a hashtag easily. Zappavigna (2012) pointed out that using hashtags is 'an emergent convention for labelling the topic of a micropost and a form of metadata incorporated into posts' (p. 1).

Usually, there are two basic types of hashtags: mid-sentence hashtags and end hashtags. Usually, the mid-sentence tag functions as a normal word in the sentence, so it functions as an actual word; while the end hashtag more often bears no grammatical relationship to the sentence, only denoting a theme symbol at the end of a tweet. Hence, some previous studies took this mid-sentence hashtag as a practical approach to analyse tweets (Go & Bhayani, 2010). Zappavigna (2012) defined this mechanism as 'social tagging', which 'engages communities of general users' (p. 37), so this practice can attract more interactions from other users apart from the direct followers.

Cashtags A cashtag is similar to a hashtag: using the $(dollar mark) instead of a # (hash mark), it focuses on stock-related topics, so it is convenient to track the stock-related tweets by crawling tweets containing a specific cashtag. Before Twitter officially introduced this device on July 31, 2012, users had already been using this convention to discuss stock. Since July 31, 2012, cashtags are indexed by Twitter, so they perform in the same way as hashtags. Similarly, by clicking the cashtag, Twitter will redirect users to a search page where the cashtag is the keyword. The introduction of cashtags by Twitter has attracted more users to try this device. The only problem is that cashtag before the announcement date are not searchable, because Twitter considered the dollar sign as a normal word. As shown in Chapter 4, there is a huge increase of tweets containing a cashtag after Twitter's above announcement.

URL links As discussed in 2.2.1, to meet the length limit of tweets, normal URL (Uniform resource locator) links are converted to a shortened form, leaving room for the rest of content. Twitter not only converts a normal URL to a shortened t.co link, but also converts all shortened URLs according to this format. Thus, collecting data through Twitter API, the raw tweets only contain t.co links, while the complete URL or a shortened URL by a third-party service can be found in the meta information. Moreover, the shortened links often include advertisement links or spam links, so they become an important indicator of spam tweets (Benevenuto, Magno, Rodrigues, & Almeida, 2010; Kwak, Lee, Park, & Moon, 2010).

Meta information Usually, a general user can only see a tweet as shown in the above Sample 2.2.1 and 2.2.2, because 'most of the metadata collected by Twitter is not presented directly to the general user' (Zappavigna, 2012, p. 3). As mentioned above, however, a complete tweet is stored in a JSON file, which contains much more meta information. The meta information includes the tweet owner's information, post time and place, complete URL of the shortened URL and integrated picture, hashtag information and reply information. Section 4.2.2 provides a detailed account of a complete tweet with rich meta information.

Protected account Twitter provides an option to let users set their accounts as private, which means that others need to ask approval to follow those users, and also tweets posted by those users are not visible to the public. Thus, a protected user's tweets will not be collected by any means.

However, protected accounts are not common on Twitter, so their existence will not affect the data collection considerably.

Time Zappavigna (2012) points out that 'time is an important variable in microblogging', because it offers 'a view of how particular couplings of meanings shift and change, enabling us to consider relationship between linguistic features in time series' (p. 39). Generally, a tweet may have two time indications: one is the post time in the meta data, and another one is the time mentioned in the tweet, which is optional. However, retweets may have three time indications, as they may also include the post time of the retweet. In this research, the time indication of a tweet is particularly important, because the temporal correlation to stock changes is one of the main relationships between tweets and market. As shown in Chapter 4, each collected tweet contains at least two time indications: the collected time and the post time. Some of them also contain the time indication in the tweet itself. The choice of the proper time indication in this research will be discussed in Section 4.3.5.

2.2.5 Prediction power of Twitter

Apart from being used in stock prediction in the previous studies, Twitter has also been used to predict natural disasters, outbreak of diseases or political elections, and these predictions have attracted long-lasting and wide interests from both industry and academia. This section focuses only on the studies of using tweets to predict earthquakes, due to the increasing popularity of the application in this area.

In the area of predicting natural disasters, tweets have been shown to be effective as a means of predicting and thus offering early warnings about earthquakes. One of the first studies in this line was conducted by P. Earle et al. (2010), who used Twitter data to obtain firsthand accounts of earth tremors in order to detect and map an earthquake. By analysing tweet counts and the geographic information contained in tweets during the 2009 MW 4.3 Morgan Hill, CA earthquake, this study showed both the potentials and limitations of using Twitter to predict an earthquake: it was faster than the traditional methods, such as the Advanced National Seismic System (ANSS) developed by United States Geographic Service (USGS), to obtain firsthand information about the occurrence of an earthquake from tweet users around the epicentre, but the obtained information was not sufficient to provide a seismographically detailed picture of the event.

Following this, a number of studies have been conducted in Japan due to its high frequency of earthquakes. For instance, based on studies of discussions of earthquakes on Twitter, Sakaki, Okazaki, & Matsuo (2010) developed an algorithm for detecting new seismic events, which in turn led to the development of an earthquake reporting system which can distribute earthquake alarms faster than the conventional alarm system maintained by Japan Meteorological Agency (JMA). A similar result can be found in Doan et al. (2012)'s analysis of subsequent responses to the 2011 Tohoku earthquake on Twitter. In this study, Twitter was found to be useful for 'tracking the public mood of populations affected by natural disasters as well as an early warning system'.

A further study by P. S. Earle et al. (2012) emphasized two major advantages of using tweets to predict earthquake over the conventional approaches: 1. The detections based on tweets are faster than seismographic detections in poorly equipped regions. 2. Tweets can provide firsthand narratives immediately from people who experienced the disaster, and this could never be achieved by the conventional approaches.

Overall, Twitter has extended possibilities in some conventionally well-developed prediction areas, such as earthquake prediction, and similarly, it may also benefit the stock prediction area.

2.2.6 Influence of Twitter on the market

To carefully evaluate the influence of Twitter on the stock market is not practically realistic, because it is hard to isolate the influence from other sources. However, some negative indications can explain the mechanism from a different perspective. As mentioned earlier in Chapter 1, in January, 2013, the abrupt changes of two NASDAQ tickers might have connections with some fraudulent information from two hoax Twitter accounts. Therefore, these two cases are good illustrations of Twitter's influence.

According to Reuters (McCrank & Gaffen, 2013; Vlastelica et al., 2013), in the two cases, the victims were Audience Inc. (NASDAQ: ADNC.O) and Sarepta Therapeutics Inc. (NASDAQ: SRPT), and the main suspicious Twitter accounts were @Mudd1waters and @citreonresearc. The username of the first suspicions account is very similar to a famous consulting company Muddy Waters Research, and by February 2, 2013, it has posted 9 tweets in total. The name of the second account is similar to Citroen Research, but the account was suspended by Twitter soon after, so no data can be retrieved. Thus, this section only discusses the Audience Inc. and @Mudd1waters case.

The report further indicates that there were 800,000 shares being changed during that afternoon, which is nearly eight times the number of changes in the previous 25 days. These suspicious changes pulled the original price of $12 to $8.87 at 2:21pm (McCrank & Gaffen, 2013). Thus, the market, media and the regulators considered that these unusual trading activities might be caused by misleading information posted by the suspicious Twitter account @Mudd1waters.

The suspicious account @Mudd1waters was registered on January 25, 2013, but posted its first tweet four days later. By February 6, 2013, it only had 24 followers and 18 followees. This indicates that it had a very limited number of network relationships, so its influence on other users was limited. Also, checking the nine tweets posted by @Mudd1waters, only one was retweeted once by another user (see Sample 2.2.9), and that user only had approximately 300 followers. Therefore, those tweets could only reach a limited audience theoretically.

Sample 2.2.3 to 2.2.9 are tweets posted by @Mudd1waters: the first eight were posted on January 29, 2013, and only the last one was posted on February 2, 2013. Three of these tweets are reply tweets (see Sample 2.2.8, 2.2.10 and 2.2.11). The receivers are @CMEGroup, @MerrillLynch, and @NicoSEnea: The first two are stock-related corporation accounts, and the last is an account of a stock analyst according to their profiles. However, there is no interaction between this account and other Twitter users: None of these three tweets received a single reply from any other receivers; thus, these tweets just published the information as other non-reply tweets, but only with particular foci. Moreover, it is obvious that seven tweets (see Sample 2.2.3, 2.2.4 and 2.2.6 to 2.2.10) are highly similar. They simply repeated the same content that the target company is being investigated by the regulator. Ironically, five out of nine of these tweets had a clear misspelling, as they spelt *report* as ***reort***.

Sample 2.2.3 29-01-13 08:44 AM $ADNC AUDIENCE the noise suppression company being investigated by DOJ on rumoured fraud charges Full reort to follow

Sample 2.2.4 29-01-13 08:51 AM $ADNC AUDIENCE the noise suppression company being investigated by DOJ on rumoured fraud charges Full reort to follow later

Sample 2.2.5 29-01-13 08:56 AM $adnc annual report held back

Sample 2.2.6 29-01-13 09:28 AM $ADNC AUDIENCE noise suppression company being investigated by DOJ on rumoured fraud charges Full reort to follow

Sample 2.2.7 29-01-13 09:29 AM $ADNC AUDIENCE noise suppression company being investigated by DOJ on rumoured fraud charges

Sample 2.2.8 29-01-13 09:31 AM @CMEGroup $ADNC AUDIENCE noise suppression company being investigated by DOJ on rumoured fraud charges Full reort to follow later

Sample 2.2.9 29-01-13 09:35 AM $ADNC noise suppression company being investigated by DOJ on rumoured fraud charges Full reort to follow later

Sample 2.2.10 29-01-13 10:09 AM @MerrillLynch $ADNC AUDIENCE noise suppression company being investigated by DOJ on rumoured fraud charges (1 reply by @Bulltalk)

Sample 2.2.11 02-02-13 05:16 AM @NicoSEnea but no lack of money

Considering the numbers of @Mudd1waters' followers and the number of its tweets being retweeted, this account could only reach a very limited human audience in principle. Consequently, there was only a tiny opportunity for human investors to fetch this information. According to this, the only logical explanation is that some automatic crawlers have retrieved this information via Twitter's APIs and considered it as an authentic piece of information to be processed for automatic trading.

If the above inference stands, it is meaningful to this research. First, this suggests that Twitter has a certain influence on the market, and that this influence can be significant. Also, any automatic trading platform based on Twitter source needs a mechanism to filter out fraud information. Finally, evaluating the influence of tweets based on the count of the tweets needs more serious reconsideration, particularly if some highly similar tweets contain the same errors.

2.3 Summary

This chapter presented an overview of two fundamental aspects of background knowledge for this research: economics, and Twitter and tweets.

The first part summarised two economic theories, the efficient market hypothesis and word of mouth theory, to support the ongoing research. In economics, people have a long-lasting interest in stock prediction, and the efficient market hypothesis suggests that information can influence the market and that the stock market is a predictable market. In addition, the word of mouth theory points out that story-telling is an important approach to marketing. Social media satisfies both conditions, thus it can be regarded as a good information source for market prediction. It then discussed the simplified model of the relationship of information flow and market. Finally, it discussed several main criticisms of analysing social media data to predict stock movements, concluding that the main criticisms became less relevant when considered alongside current economic theories.

The emergence of Twitter brings a new data source that has unprecedented potential to satisfy this need. The next part introduced the general features of Twitter and tweets, as well as the information coverage and the additional challenges for academic and commercial analysts Twitter brings. The three main features of Twitter, namely real-time updates, a free-to-use policy and a well-built ecosystem, shape it into an information platform with full freedom, an open development platform for all kinds of related services, and a multi-win business model for the Internet. This action also discussed the main devices and other main features of tweets: They not only make tweets a unique domain to analyse, but also raise a great deal of challenges that were not found in other domains. Consequently, they bring more difficulties to the research as Twitter has forced numerous organisations to stop exporting tweet data, even if for academic purpose (Watters, 2011). Finally, it discussed the fraud tweet case in 2013, which is a good indication of Twitter's influence on market.

In addition, this chapter discussed the details of Twitter's data releasing policy. Twitter uses APIs to provide data: the free tweet data are limited, the charged data are enormous. Considering the complexity of the data and the budget, the research chooses the Search API to collect data.

To summarise, both opportunities and challenges are involved in developing stock prediction methods based on tweets.

Chapter 3 Previous Research

As stated in Chapter 1, the overarching question for this research is **Do linguistic analysis improve the accuracy of the sentiment identification of the stock tweet sentiment**. This chapter first introduces sentiment analysis and previous research on stock prediction based on tweet data, and then elaborates on each of the six specific research questions listed in Chapter 1 in detail, assessing the extent to which these questions have been addressed by previous studies. To reiterate, the six questions are as follows: 1. Does a clearer definition of stock tweets improve the quality of an analysis of such tweets (Section 3.3)? 2. Are stock tweets a linguistically distinct type of tweet? What specific linguistic features do they have (Section 3.4)? 3. If stock tweets have explicit linguistic features, is it possible to automatically identify the stock tweets based on their linguistic features (Section 3.5)? 4. How can a robust sentiment word list for tweet sentiment classification be designed (Section 3.6)? 5. Does a more precise definition of a positive, neutral, or negative stock tweet in accordance with market values help to improve the quality of stock tweet sentiment analysis (Section 3.6)? 6. How can the neutral sentiment category of stock tweets be defined and processed (Section 3.6)? These questions are addressed in turn in the following analyses.

3.1 Introduction of Sentiment Analysis

Pang and Lee (2008) stated that the demand of understanding customer feedback or political opinion has risen since the Internet appeared. In turn, it has prompted the initial motivation of the emergence of sentiment analysis because using humans to manually review a large quantity of data containing sentiment information is impractical. The history of sentiment analysis is not long, and the term first appeared in around 2001 (Pang & Lee, 2008). This starting point heavily depends upon a number of other developments, such as the rapid development of machine learning and database technology.

As a sub-branch of computational linguistics, sentiment analysis consists of two parts: the classification approach and the classification algorithm. At present, the classification approach makes little use of the linguistic features of the textural data, relying only on some simple statistical methods, such as the bag-of-words approach. The present research attempts to argue the addition of linguistic features can improve the accuracy of sentiment analysis on stock tweets. Therefore, a detailed discussion of current classification approaches is provided in Section 3.6. As for the classification algorithm, most studies have used popular machine learning algorithms, such as decision tree, random forests, Naïve Bayes or support vector machine. This is not the focus of the research, so Chapter 6 only gives a brief introduction of these algorithms.

Generally, there are two main classification strategies being used to classify data in sentiment analysis: polarity classification, and fine-grained classification (Täckström & McDonald, 2011). The first approach is a binary approach: The data are classified as positive or negative, and sometimes a neutral category is also included. The binary category is divided into different degrees, such as “very positive”, “positive”, “negative”, and “very negative”. The first approach is simple and straightforward. While many previous studies have focused on the first approach, the overall trend moves away from the polarity approach to the multi-dimensional approach, especially in industry. For example, Bollen et al. (2011) pointed out that using a binary approach to classify sentiment might “ignore the rich, multi-dimensional structure of human mood” (p. 3), and this was supported by Grimes (2013), who noticed that “the market has started to recognise the limitations of polarity-based sentiment analysis”. Section 3.6.4 discusses the classification strategy used in this research, Section 4.4 provides a detailed annotation scheme according to this classification strategy, and Section 5.5 discusses this annotation scheme by accessing the temporal correlations between stock price and tweet sentiments.

3.1.1 Main characteristics of sentiment analysis

Subjectivity and interaction with social media are the two main characteristics that differentiate sentiment analysis from other natural language processing (NLP) branches.

Opinions, evaluations, emotions and speculations are all difficult to observe or verify objectively as Pang and Lee (2008) indicated. This suggests that the accuracy of sentiment analysis cannot compete with other NLP tasks, for example, classifying the part-of-speech tags based on a statistical model can achieve an accuracy of more than 95 percent (Mason, 2004). The accuracy of sentiment analysis depends on many factors, such as the sample domain (Pang & Lee, 2008, p. 25) or the topic (Mei, Ling, Wondra, Su, & Zhai, 2007). Therefore, in general, accuracies between 60% and 70% are considered to be good, but this does vary (S. R. Das & Chen, 2007; Oh & Sheng, 2011; Pang, Lee, & Vaithyanathan, 2002). For some specific domains, the accuracy can be as high as 90% (Potts, 2011). Other projects achieved a much lower accuracy; for example, a cross-domain project by Bloom (2011) achieved an overall accuracy of 26.1%. This feature differs from other NLP tasks greatly.

Undoubtedly, sentiment analysis links with the demand to analyse Internet data for such things as product reviews (Turney, 2002) and box office predictions (Asur & Huberman, 2010). The dramatic growth in the update of social media has accelerated the evolution of sentiment analysis due to its “rich and diverse data” (Mitixa & Rana, 2013, p. 238). Thus, more and more studies in this area have turned to investigate social media data, such as Twitter or Facebook. Also, this new type of data brings more possibilities to expand sentiment analysis. For instance, combining social media data with geographical tags, sentiment analysis is able to provide prediction based on geographical information, such as earthquake prediction (Sakaki et al., 2010) or influenza prediction (Lampos & Cristianini, 2010).

3.1.2 Development of sentiment analysis

In a recent survey, Cambria, Schuller, Xia and Havasi (2013) pointed out that the development of sentiment analysis has three main directions: from heuristics to discourse structure, from coarse- to fine-grained, and from keyword- to concept-based analysis. In other words, sentiment analysis tends to use more elaborate approaches to investigate more complicated problems from some more comprehensive perspectives. To review the rapid development of sentiment analysis, two aspects should be taken into consideration: the data source and application.

Initially, sentiment analysis concentrated on **explicit** data, such as online product reviews (Turney, 2002) or movie reviews (Pang et al., 2002): These data directly and explicitly present users" evaluations. Moving on, the focus has gradually switched to more "implicit" data, such as social media data, which may not explicitly express users" evaluations or may integrate them with a great deal of irrelevant contents. This indicates that one of the challenges of sentiment analysis is the challenge of dealing with more and more domain-free data.

Pang and Lee (2008) summarised four major applications of sentiment analysis: review-related websites, sub-component technology, business and government intelligence, and across different domains. However, the evolution of sentiment analysis has brought more possibilities, and makes it hard to summarise. Generally, the trend is moving away from direct applications to more sophisticated applications. At the beginning, sentiment analysis aimed to answer the question of what a user's feedback of a product or service was. Thus, the applications of sentiment analysis (e.g, production review evaluations) were straightforward and had direct outputs. Gradually, the applications have become more complicated: They do not only simply answer the original question, but also seek to answer broader questions such as what can be done based on the feedback. Using sentiment analysis to predict the investment actions provides a good illustration of this: Sentiment analysis answers the question of what the market reaction was or is, as well as the question of what action will be taken next (Rambocas & Gama, 2013).

3.1.3 Main classification approaches to sentiment analysis

With the rapid development of sentiment analysis, these simple classification methods are slightly out of date. With various classification approaches to sentiment analysis, there are a number of criteria to group them, but some approaches combine several different methods. This section briefly discusses the main approaches of recent years, such as the bag-of-words model, semantic orientation, and emoticons. Most approaches make little integration of comprehensive linguistic features of data to identify sentiment. A more detailed discussion of the main sentiment classification approaches being used in stock prediction based on tweet data is given in Section 3.6.

In the early stage, sentiment analysis studies mainly used two approaches to perform analysis according to Whitelaw, Garg, & Argamon (2005b):

> The first (bag-of-words) attempts to learn a positive/negative document classifier based on occurrence frequencies of the various words in the document.
>
> Another main approach (semantic orientation) classifies words (usually automatically) into two classes, “good” and “bad”, and then computes an overall good/bad score for the text (p.1).

The first one is the bag-of-words model, which concentrates on the individual word frequency: It generates a frequency list from the data, calculates the weight of different polarity of words, and then determines the overall polarity by the polarity of individual words. Although the basic bag-of-words model uses unigram as the main feature, parsing texts into individual words, an increasing trend is to use bigrams as the main feature in this model. However, using two connected words as a pattern, the accuracy of sentiment analysis improves little (Pang et al., 2002).

The second approach is the semantic orientation model (Akshi Kumar & Sebastian, 2012), which first designed a seed word list for positive and negative sentiment respectively and then computed the distance between words in the text and in the seed lists. It considered the semantic direction of a word from its norm as the semantic orientation (Hatzivassiloglou & McKeown, 1997), so in order to generate the overall score of the text, it uses a designed seed word list, usually based on adjectives, to calculate the orientation of each adjective or phrase containing adjectives and adverbs (Turney, 2002).

Other approaches include appraisal framework and local grammar (Bloom, 2011). The first approach uses appraisal framework from linguistics to evaluate the intensity of adjectives in order to recognise the strength of sentiment. For example, Whitelaw, Garg and Argamon (2005a) and Whitelaw et al. (2005b) extracted appraisal groups, which they define as "those groups and phrases in a text giving what kind and intensity of appraisal is expressed". They adopt this simplified definition of appraisal to identify the sentiment in the texts. Local grammar, developed by Gross (1982) and his followers, such as Roche (1999), is used to extract verb patterns regardless of the domain of the text. In particular, Ahmad et al. (2006) used this approach to understand the sentiment of the in multi-lingual financial news by extracting specific verb patterns.

In addition, emoticons, combinations of characters and punctuation marks to represent specific emotions are considered as a useful indicator to identify sentiment information (Hogenboom et al., 2013; K. Liu, Li, & Guo, 2012; Read, 2005), and some studies integrated this approach with other approaches (Akshi Kumar & Sebastian, 2012). However, as will be shown in the later analyses, emoticons appear less frequently in stock tweets. Therefore, this approach cannot be applied to this thesis.

Also, hashtags are considered as an factor in tweet sentiment analysis (Davidov, Tsur, & Rappoport, 2010). Based on the hypothesis that hashtags can express, group and propagate people's sentiment with regard to some topics and events, Barbosa et al. (2012) aimed to use hashtags to develop an automatic sentiment detection of specific events and topics, and the qualitative result suggested that about one third of hashtags are "necessary for defining the message sentiment" (p. 2625).

3.1.4 Summary

These main characteristics make sentiment analysis not only a unique NLP sub-branch but also an increasingly important one. A detailed introduction of sentiment classification approaches being used in the stock prediction is given in Section 3.5.

3.2 Previous Studies on Stock Prediction Based on Social Media

Reading previous work on stock prediction, much effort has been taken to utilise sentiment analysis. It is clear that sentiment analysis is developing rapidly, and the classification model has become more and more elaborate to achieve better performance on complicated language structures. This approach has also been applied to stock prediction, especially based on social media data.

There are two main traditions in research on stock prediction based on tweet data, which are closely relevant to the scope of this thesis. The first tradition uses expressions of public mood in tweets to predict the overall market changes, for example, the movement of DJIA index (Bollen et al., 2011). Another one uses ticker-specific tweets to predict the individual ticker's performance, such as the changes in the Apple.Inc stock price (Y. Mao, Wang, Wei, & Liu, 2012). Each of these research strands is reviewed below in turn.

3.2.1 Overall market prediction based on public mood

In recent years, using public mood to track the overall market performance has become an increasingly popular approach. The first attempt to do this, carried out by Bollen et al. (2009), demonstrates that social events discussed on Twitter can have a significant effect on public mood. Following this result, Bollen et al. (2011) found that the indicators of public mood extract from tweet data can help to predict DJIA price changes. They then extracted six different public moods from tweets: *calm*, *alert*, *sure*, *vital*, *kind* and *happy*. The calm mood generated an accuracy of 87.6% in predicting daily changes in the closing values of the DJIA. This accuracy only indicates the temporal correlation between sentiment and price, but not the accuracy of the sentiment classification conducted in their research. Later in 2011, Mao, Counts and Bollen (2011) compared the data of tweets, online news headlines and Google search queries, and showed that using the tweet data (as well as Google search query data) is able to predict market changes. However, they rejected the notion that the traditional finance survey data can predict market changes.

Similarly, by analysing tweets only containing words such as *fear*, *worry*, and *hope*, Xue Zhang et al. (2010) found that the percentage of tweets in different polarities is negatively correlated with the Dow Jones, NASDAQ (National Association of Securities Dealers Automated Quotations) and S&P 500 (Standard & Poor's 500), but positively correlated to VIX (Chicago Board Options Exchange Market Volatility Index). Therefore, they suggested that using tweets might be of help in formulating the following day's trading strategy. However, this approach has a clear shortcoming that a single word can not represent the sentiment of a tweet as a whole.

Soon after, Stanford held a Machine Leaning course in autumn, 2011, and seven final projects replicated Bollen et al. (2011)'s experiment. The cross-comparison is shown in Table 3.2.1.

Table 3.2.1 Comparison of Stanford final projects, 2011 (I)

Project	**Correlation**	**Word list**	**Feature**	**Dimension**
Chyan, Lengerich, & Hsieh (2011)	80%	POMS	bag-of-words	NA
Chakoumakos, Trusheim, & Yendluri (2011)	78%	POMS	unigram	6
Kuleshov (2011)	60%	WordNet POMS	word vocabulary	NA
Debbini, Estin, & Goutagny (2011)	63%	Davies and Ghahramani (2011)	unigram	NA
Hsu, Shiu, & Torczynski (2011)	82%	Davies and Ghahramani (2011) A fiction word list	unigram collocations	6
Mittal & Arpit (2011)	76%	POMS	unigram	4
R. Chen & Lazer (2011)	70%	Davies and Ghahramani (2011) a pre-generated word list	NA	1/2/more

In brief, these seven projects share four main similarities. Firstly, none of them achieved Bollen et al. (2011)'s high temporal correlation. The best correlation amongst the seven projects varied from 60% to 82%. Secondly, all of the projects used pre-designed word lists, such as the Twitter Sentiment Analysis Word List by Davies and Ghahramani (2011), or a pre-designed list that had been developed from Profile of Mood States (POMS) by McNair, Lorr and Droppleman (2003). Only one used a hybrid approach by extracting the top 1000 frequent words (Debbini et al., 2011). The third similarity is that most of these studies focused on the unigram or bigram features of the data, which are similar to the bag-of-words model. Finally, four of them followed Bollen et al's (2011) sentiment classification category, using a multi-dimensional polarity (from 4 to 6 dimensions) to do sentiment analysis (Chakoumakos et al., 2011; R. Chen & Lazer, 2011; Hsu et al., 2011; Mittal & Arpit, 2011).

From these comparisons, three points are clear:

1. Despite the claim of Bollen et al. (2011) that calmness sentiment correlated most closely to stock trends, from these Stanford reports, Bollen et al.'s results do not appear to be replicable.

2. Focusing on the same data without regard of the size, the results of sentiment analysis can vary, and the main cause might be the selection of different features (see Table 3.2.2). Based on similar machine learning methods, the cross-validated results are similar. Thus, the present word-level approaches to sentiment analysis need to be implemented, and some more stable features need to be considered.

Table 3.2.2 Comparison of Stanford final projects, 2011 (II)

Project	**Correlation**	**Data**	**Machine learning method**
Chyan et al. (2011)	80%	SNAP	neural network
Chakoumakos et al. (2011)	78%	SNAP	NA
Kuleshov (2011)	60%	SNAP	neutral network, SVM
Debbini et al. (2011)	63%	SNAP	SVM
Hsu et al. (2011)	82%	SNAP	SVM, neutral network, RPT, SVD grouping
Mittal & Arpit (2011)	76%	SNAP	linear regression, logistic regression, SVM, SOFNN
R. Chen & Lazer (2011)	70%	NA	linear regression

3. The advantage of sentiment analysis based on the multi-dimensional approach in stock prediction appears not to be obvious. That the sentiment analysis accuracy based on different machine learning methods is similar illustrates that the performance of a multi-dimensional approach has limitations. Moreover, it is hard to be improved through machine learning methods (see Table 3.2.3).

Table 3.2.3 Comparison of Stanford final projects, 2011 (III)

Project	**Correlation**	**Machine learning method**	**Dimension**
Chyan et al. (2011)	80%	neutral network	NA
Chakoumakos et al. (2011)	78%	NA	6
Kuleshov (2011)	60%	neutral network, SVM	NA
Debbini et al. (2011)	63%	SVM	NA
Hsu et al. (2011)	82%	SVM, neutral network, RPT, SVD grouping	6
Mittal & Arpit (2011)	76%	linear regression, logistic regression, SVM, SOFNN	4
R. Chen & Lazer (2011)	70%	linear regression	1/2/more

On the basis of these studies, the prospects for using sentiment-based methods of predicting stock prices seem bleak. However, a more recent study by Rao & Srivastava (2012) showed a high correlation between DJIA prices and Twitter sentiments, and that in a short-term period, Twitter feeds had a strong influence on stock trends. This study suggests that it is worth continuing with this approach, although it is clear that a number of aspects of the approach still need considerable methodological development.

3.2.2 Individual ticker prediction based on ticker-specific mood

The other main trend to be reviewed here is one that uses ticker-specific tweets to predict the individual ticker's performance. The analyses by Sprenger and Welpe (2010) and Sprenger, Tumasjan, Sandner and Welpe (2013) on tweets and S&P 500 Index companies show that tweets contain helpful information, outperforming current market indicators. Their own research led them to propose the following five bold but testable claims about tweets as predictors of market movements, as follows:

1. Abnormal stock returns and the bullishness[3] of tweets can be correlated.
2. New information from tweets is reflected in market prices quickly.
3. Investors on Twitter follow a contrarian strategy.
4. Message volume can predict next-day trading volume.
5. Users who provide above-average-quality information often have more followers, and their tweets are more often retweeted.

Subsequent studies have generally agreed with these conclusions. Leeuwen (2011) used a set of fine-grained categories to classify tweets about stocks in the Amsterdam Exchange Index, and found that stock price and trading behaviour are predictable. However, it is not clear from this study whether the social media has any causality towards the market. Oh and Sheng (2011) focused on a smaller tweet dataset than that used by Sprenger and Welpe to analyse the relationship between tweets and ticker performance in NYSE, NASDAQ. They confirmed that tweets can predict simple returns and market-adjusted returns respectively, and the accuracy is consistent with the underreaction hypothesis in behavioural finance.

Another positive result was obtained by Smailović, Grčar and Znidarsic (2012). They collected tweets with a focus on Apple Inc. to identify important events and predict the price changes. The result again shows that tweet sentiment can predict the rise or fall in closing price, in this case, two full days before the actual changes.

Further support comes from Nann, Krauss and Schoder (2013)'s study. They compared tweet data between online and traditional news data in order to understand which data has the most predictable power of the S&P 500 index. Based on the findings, they then developed a trading strategy and obtained a positive return of up to 0.49% per trade in virtual trading.

[3] In economics context, *bearish* refers to a downward price movement of a stock, and *bullish* refers to a upward price movement of a stock.

The only note of caution thus far is found in Oliveira, Cortez and Areal (2013). Focusing on nine major technological companies, they found inadequate evidence that sentiment indicators can explain these stock returns. However, they nevertheless suggested that tweet volume correlates with the trading volume, and particularly with volatility. In short, then, this approach seems to show a great deal of promise.

3.2.3 Other approaches to using tweets to predict the market

In addition to the above, there are some projects taking a number of other approaches to predict movements of the market based on tweets.

Yi (2009) used supervised machine learning methods to identify stock-related tweets from raw tweets. He found that simple noun counting failed to correlate to the market, while some more complex models, such as the loose ngram model (co-occurring words within limited range) by Zhang and Zhu (2007), showed an increase of the correlation between tweets and the stock price.

By comparing the correlation between blogs, online news, tweets and NYSE stock price, Zhang and Skiena (2010) suggested that tweet data have the following characteristics:

1. Twitter is different in that its polarity also has some correlation with tomorrow's or the day after tomorrow's return.
2. Stock market incorporates Twitter sentiment slower than news and need two or three days.
3. The sentiment of Twitter is more persistent between neighbouring days.

Vincent and Armstrong (2010) captured new words appearing on Twitter feeds to predict breaking news in order to assist high-frequency trading. Their algorithm performed better than the benchmark algorithm in predicting the correlation between the currency market and breaking news.

Zhang, Fuehres and Gloor (2011) developed their previous work to correlate tweets with financial market movement such as gold price, crude oil price, currency exchange rates and stock market indicators. By collecting retweets posted in the US containing six market-related keywords, they showed that these tweets are correlated to market changes and that most of them correspond to next-day market performance.

Instead of directly searching for tweet sentiment about the market, Bar-Haim, Dinur, Feldman, Fresko, and Goldstein (2011) concentrated on identifying expert investors on Twitter in order to capture expert trading suggestions. Based on this model, the precision of predicting stock rise is relatively high.

Y. Mao et al. (2012) correlated tweets about the S&P 500 to corresponding market data on three levels: the market, the sector, and the company. They found that the daily volume of tweets is correlated to stock indicators.

Like Mao et al. (2011), Ruiz et al. (2012) also showed that the relevant tweet volume is correlated with the trading volume of stock and is stronger than the correlation with the stock price. However, they then developed a trading stimulator based on these results, which showed that the trading strategy based on relatively weak correlations between price and tweet data features still outperforms other baseline trading strategies.

3.2.4 Summary

Reviewing these relevant studies, it is clear that there are two main trends in using tweets to predict the market. Either extracting public mood from tweets in general to predict the overall market changes, or using ticker-specific tweets to predict the individual performance, has been studied in previous work. These studies agree that the correlation between tweets and the market exists, although they vary between different components, and the correlation between the tweet volume and the trading volume is the most proved correlation. However, none of the studies place an emphasis on the linguistic features of tweets to identify sentiment. Moreover, these studies leave one fundamental question unanswered: What is a stock tweet?

3.3 Definition of Stock Tweets

What a stock tweet is seems an obvious question, but the category of stock tweets does not exist on Twitter and is not a concept defined by Twitter. Moreover, though many studies indicate that they used stock tweets for analysis, there has not been a detailed discussion of what a stock tweet is so far. In the discussion that follows, it is shown that robust definitions of stock tweets can only emerge from the process of collecting tweet data (see Chapter 4) and studying the empirical features of these data in detail (see Chapter 7 and 8).

As discussed above, there are two main trends in collecting tweet data to predict the stock market: either using tweets to predict general market performance or using tweets to predict the performance of an individual company. Thus, accordingly, there are two main types of tweet data to be collected: general tweets or ticker-specific tweets.

3.3.1 Collecting general tweets

The first trend is to collect public tweets without setting any specific keywords, so many of the collected tweets may not be relevant to market topics. Bollen et al. (2011) were one of the first to apply this approach. They collected tweets with the patterns containing *I*, such as *I am feeling* or *I'm* to capture the public mood, but excluded tweets containing *http:* or *www.* This approach does not focus on any specific topics of tweets, but concentrates only on the language being used in the tweets, meaning the collected tweets might contain a higher percentage of irrelevant contents than focusing on specific topic words. Also, removing tweets with expressions such as *http:* and *www* in order to "avoid spam messages and other information-oriented tweets" (Bollen et al., 2011, p. 2) is questionable, because a number of tweets discussing stock-relevant topics contain URL links. A later research conducted by Mao et al. (2011) applied a similar approach, which included a "15%-30% random sample of all public tweets" (p. 3).

Following this approach, Xue Zhang et al. (2011) collected only retweets containing the words *hope*, *fear* or *worry* posted in America to conduct their research, and they classified retweets into corresponding categories. They considered these tweets to be able to present the sentiment of the predesigned keywords well, so they did not apply any sentiment analysis on the data. Such an approach could be seen to be problematic; however, because accuracy of using a single word in a tweet to evaluate the overall sentiment of that tweet is questionable, for example, if a tweet contains both *hope* and *fear*, then it can be put into either category according to Xue Zhang et al. (2011)'s definition, so both categories can gain a score, which means the overall accuracy of the sentiment will then be affected in turn.

Although the major advantage of these approaches is that it can easily collect enormous quantities of data (see Table 3.3.1), it can only be used to detect the relationship between the public mood and the market as a whole instead of individual ticker's performance. Moreover, this approach may bring more noise as it contains much content irrelevant to the market.

Table 3.3.1 Statistics of the sample size based on the first approach

Project	**Number of tweets**
Bollen et al. (2011)	9,853,498
Mao et al. (2011)	15%-30% of public tweets
Xue Zhang et al. (2011)	3,809,437

3.3.2 Collecting ticker-specific tweets

Another trend is to collect tweets containing keywords that are relevant to the market, in order to predict the performance of individual tickers. For instance, Sprenger and Welpe (2010) are one of the first to collect tweets containing the ticker names of S&P 500 companies with a dollar sign in front. They aimed to "investigate the most relevant subset of stock microblogs and avoid 'noise'" (p. 22). Following this approach, Brown (2012), Y. Mao et al. (2012), Smailović et al. (2012), and Oliveira et al. (2013) all applied the same strategy to collect data. Oh and Sheng (2011) took a similar strategy: They fetched tweets from Stocktwits.com and removed tweets "without any ticker, more than one ticker, or not in NASDAQ and NYSE exchanges" (p. 7). This procedure removed about two thirds of the downloaded tweets, which left only 72,221 tweets for the following analysis.

In addition, some of the previous studies used a broad definition of stock tweets. For example, Chakoumakos et al. (2011) pointed out that if only the tweets containing the ticker symbol are considered, then the data are sparse, so they included tweets containing more related information, such as company names, company leader names, or even informal company names. This is a controversial approach. On the one side, only using ticker names as query keywords to collect data yields focused but sparse data. On the other side, using broader concepts such as keywords in the collection can bring more noise. In addition, Ruiz et al. (2012) used a similar approach to collect stock tweets, and they included the company ticker names and hashtags associated with the company. According to their understanding, they considered cashtags as one type of hashtag, so they put expressions such as *#Yahoo*, *$YHOO*, or *#Yahoo* together to set their query keywords. This approach has a drawback in that it may cover more ambiguous tweets. They must have noticed, for example, that abbreviations of Yahoo are frequently used in news tweets by Yahoo News.

Although the second trend can reduce the proportion of irrelevant contents, it provides relatively limited data in the same time span compared to the above approach (see Table 3.3.2).

Table 3.3.2 Statistics of the sample size based on the second approach

Project	**Number of tweets**
Sprenger & Welpe (2010)	249,533
Sprenger et al. (2013)	249,533
Y. Mao et al. (2012)	9434 daily tweets
Brown (2012)	13,000
Smailović et al. (2012)	33,733
Oliveira et al. (2013)	NA
Oh & Sheng (2011)	72,221
Ruiz et al. (2012)	NA

3.3.3 Other approaches to collecting tweet data

Other studies use different strategies to collect data to make prediction. Yi (2009) used a completely different approach to find ticker-specific tweets. He first crawled around 60 million tweets without a selection of keywords, and then applied different NLP methods to find tweets with a specific focus on individual companies. Ignoring specific tweets, Vincent and Armstrong (2010) monitored the top new keywords on Twitter every minute and used these new words as the measurement of the volatility on Twitter. Both approaches may well suit their respective research goals, but they make it difficult to carry out sentiment analysis on specific issues of the stock market.

Finally, some studies have failed to give an explicit explanation of how they defined or collected stock tweets. Zhang and Skiena (2010), Xue Zhang et al. (2010) and Nann et al. (2013) briefly stated that they only collected tweet data in a specific period, without any explanation of what types of tweets they collected. Rao and Srivastava (2012) focused on DJIA, NASDAQ-100, and 11 major companies simply because they "are some of the highly traded and discussed technology stocks having very high tweet volumes" (p. 6), but they did not give a detailed explanation of their selection criteria of stock tweets.

In general, these studies collected many more tweets to extract relevant information (see Table 3.3.3).

Table 3.3.3 Statistics of the sample size based on other approaches

Project	**Number of tweets**
Yi (2009)	61,756,056
Nann et al. (2013)	1,801,345
Tushar Rao & Srivastava (2012)	1,964,044
T Rao & Srivastava (2012)	4,025,595
W. B. Zhang & Skiena (2010)	NA
Xue Zhang et al. (2010)	NA

3.3.4 Defining stock tweets

Although using tweet data to predict stock trends has become more and more popular, none of these studies discussed the definition of stock tweets. Therefore, the first research question of this study is apparent:

> Question 1. Does a clearer definition of stock tweets improve the quality of an analysis of such tweets?

Clearly defining stock tweets is critically important in this research because it is a data-driven study and all the following analyses rely on good quality data. In this sense, as stated in Chapter 1, this research adopts a narrow definition of “stock tweet”: a tweet that contains one or more than one cashtag and focuses on topics relevant to the stock market. This definition follows Sprenger and Welpe (2010)’s and Sprenger et al. (2013)’s approach, and as well as Brown (2012), Y. Mao et al. (2012), Smailović et al. (2012) and Oliveira et al. (2013). Such a definition can help collect tweets discussing about the performance of individual companies while also reducing the amount of irrelevant data because it uses cashtags as keywords. However, this definition only gives a general idea of stock tweets. A more in-depth definition is explored in this research, and an in-depth discussion can be found in Chapter 10.

3.4 Linguistic Features of Stock Tweets

Without a comprehensive definition of stock tweets, it is also unclear whether stock tweets have specific linguistic features, and more importantly, if they can be linguistically distinguished from other types of tweets. Furthermore, in the past few years, many scholars have conducted a number of studies on tweets, covering some aspects of the linguistic features of tweets in general, from vocabulary to grammar, from style to conversation, or some unique properties of tweets, such as hashtags, but the analysis of the linguistic features of stock tweets is still in its infancy. This section introduces the four main linguistic features of tweets being studied so far, namely vocabulary features, conversation features, hashtag or cashtag features, and topic features.

3.4.1 Vocabulary features

Some of previous studies focus on the vocabulary of stock tweets. Pak and Paroubek (2010), and Lake (2010) noticed that the distribution of word frequency in tweets follows Zipf’s law. This is similar to other linguistic domains. However, Yi (2009) mentioned that most of the vocabulary in the web domain made little contribution to his following analyses using machine learning methods to classify tweets. This is similar in the stock tweet domain, where most vocabulary are irrelevant to the stock context as shown in the later case studies in Chapter 7.

Another important vocabulary feature of tweets is their informality. There are a number of types of irregular language, such as misspelling and abbreviation. Go and Bhayani (2010) suggested that the frequency of misspelling and slang in tweets is much higher than in other domains, but Kaufmann (2010) argued that “misspelling are [*sic*] much more common in SMS than in tweets” (p. 5), and the shortening of words and repetition of spelling are typical forms of informal language on Twitter. Basically, they both agree that the frequency of misspelling in tweets is significantly high. These two phenomena have brought huge challenges to analysis. As Zappavigna (2012) noticed, it is “particularly true to microblogging” (p. 20). Stock tweets share this feature as well. As Bar-Haim et al. (2011) indicated, although stock tweets are “typically abbreviated and ungrammatical utterances” (p. 1311), they have some slang expressions and various forms of sentiment expressions that are “unique to the stock tweet community”. For example, *bought* can be spelt as *bot* or *bght* in stock tweets. This may bring further difficulties for processing stock tweets because a general text normaliser, which replaces the irregular language usage to its standard from, may not work well with these abbreviations.

3.4.2 Conversation features

Many Twitter users consider Twitter to be a communication platform and often use it to establish conversations with others. Honeycutt and Herring (2009) suggested that Twitter is a collaborative platform in which the conversation is the “essential component” (p. 9). They found that a typical conversation on Twitter “is continuous and proceeds in a mostly gradual, step-wise fashion” (p. 7).

However, some distinct differences exist between Twitter users when they have conversations.

Perreault and Ruths (2011) analysed tweets posted from mobile and non-mobile platforms. They found that the mobile contents have several characteristics: They are more conversational and personal, but contain fewer links. On the other hand, Naaman, Boase, and Lai (2010)'s findings suggested that tweets appear to be more conversational by users in the "information sharing" group, who share information as their main activity on Twitter. Observing the reply intervals between Twitter conversations, Chalmers, Fleming, Wakeman and Watson (2011) found that 10% of replies come within less than 1 minute and the majority are posted in a 30-minute interval. However, many users do not post reply tweets.

This feature is particularly important to tweets on stock topics because many investors use Twitter as a platform to discuss the market.

3.4.3 Hashtag and cashtag features

Hashtags are another important characteristic of tweets. Zappavigna (2012) suggests that the functions of hashtags are two-fold: making the tweet more searchable, and making it more easy to group tweets, in turn to form social groups on Twitter. As introduced in Chapter 2, Twitter makes hashtags clickable, so clicking them will automatically redirect to a search page. By this mechanism, Twitter users can easily find topics or users that interest them. This finding is supported by Doan, Ohno-Machado, and Collier (2012), who used tweets to "emphasise or group important information of topics" (p. 4) in order to indicate the ILI (Influenza-Like-Illnesses) rate.

In the stock context, cashtags, as introduced in Chapter 1 and 2, share similar features to those of hashtags in other domains. Sprenger and Welpe (2010) regarded cashtags as an important syntax element to structure the information flow in tweets, connecting tweets to a relevant topic and making them easy to find. Furthermore, using cashtags to collect tweets can reduce irrelevant content as suggested by Sprenger and Welpe (2010), Sprenger et al. (2013) and Oliveira et al.

(2013).

3.4.4 Topic features

Each tweet is short, but clustering relevant tweets to a topic is often a key concern in tweet processing. Yi (2009) observed that topic changes are more frequent on Twitter than in other domains. When the features are less related to the target company, they bring more noise so that models generally fail to capture any true relationship with respect to stock price.

Xue Zhang et al. (2010) indicated that the topic of a tweet can be captured by one or two keywords in the tweet because the length limits the topic. However, this opinion can be questioned because some tweets are too short to convey a clear meaning and some may contain more than one topic. For example, as shown before, the tweet “general electric short 16.98 $ge” only contains five words. Extracting any individual word from this tweet cannot provide the kind of comprehensive knowledge needed for computers to understand the overall meaning. In another case, Sample 3.5.1, as it is a comment on another tweet, it contains two topics: one is news posted by USER1 about the GE’s market strategy and the other is the author expressing his optimism about the news. There are two topics in this tweet, so using just one or two words would make it hard to capture its concept in their entirety:

> Sample 3.5.1 Something to look forward to RT @USER1: GE Wants Rooftop Solar @ $4/Watt by 2012 $GE _THIS_IS_A_URL_LINK_ via @solarfeeds

Identifying the topic of stock tweets may be of little benefit in increasing the accuracy of stock tweets classification, but it can provide more features for automatic identification, as Yi (2009) suggested. He also noted that topics change more frequently in shorter tweets, so the difficulty of topic identification will increase.

Thus, considering the difficulty of identifying stock topics in tweets and the fact that little benefit comes from topic identification, this thesis will not utilise a sophisticated topic identification method to recognise specific topics, instead it tries to identify tweets with a focus on specific tickers to reduce the noise as presented later in Chapter 8.

3.4.5 Summary

Although some of the above studies have noticed that noise exists in tweets and can affect the accuracy of the identification of stock tweets (Yi, 2009), none has discussed whether and how stock tweets differ from noisy data. Furthermore, as suggested by Xue Zhang et al. (2011), linguistic analysis "might offer additional valuable information" (p. 11) to stock prediction based on tweets. However, no discussion has yet been carried out on the detailed linguistic features of tweets. Thus, this research provides an in-depth discussion on this second question:

> Question 2. Are stock tweets a specific type of tweet? And what specific linguistic features do they have?

This question will be answered in Chapter 7 and 8.

3.5 Identification of Stock Tweets

Different approaches to obtain stock tweet data have been introduced above, but a question still remains after the raw data are collected: Are the collected tweets relevant to the stock topic? Most of the previous studies ignored this problem, and only very few realised the importance of refining the data.

Lacking a comprehensive linguistic analysis of stock tweets, previous studies apply some simple strategies to identify stock tweets from collected data. For example, Nann et al. (2013) used “scurrile and nasty language” (p. 4) to identify irrelevant contents in order to clean up their data. Similarly, Bar-Haim et al. (2011) used a word list based on their preliminary analysis to identify stock tweets from the raw data. The list contains words such as *in* and *out*.

Due to the sheer quantity of the data, it is not feasible to identify stock tweets from raw data manually, so developing an automatic classifier based on machine learning methods is necessary. According to the linguistic features, the research applies machine learning methods to develop an automatic classifier to identify stock tweets, in order to answer the following research question:

> Question 3. Is it possible to identify the stock tweets according to their linguistic features automatically?

This question is addressed in Chapter 8.

3.6 Sentiment Classification in Stock Prediction

As discussed earlier, in the literature regarding stock prediction based on tweet data, there is a discussion of how to apply sentiment analysis to this area. Whether tweets bring explicit sentiment is a key question in sentiment analysis. An early analysis by Bollen et al. (2009) suggested that tweets with explicit mood or sentiment are in the minority, which reduced their sentiment dataset to 10% of the raw dataset. Moreover, Bar-Haim et al. (2011) pointed out that the sentiment on Twitter may not be easily extracted as the result of the informal language used.

Sprenger and Welpe (2010) generated a frequency list for words occurring in “sell”, “buy” and “hold” tweets, and they demonstrated that “the most common words reasonably reflect the linguistic bullishness of the three classes” (p. 26). For example, positive words and bullish words occur more frequently in the “buy” tweets.

Xue Zhang et al. (2011) indicated that people prefer optimistic to pessimistic words because they find that the frequency of positive words is much higher than the negative ones. Considering the time span of their data collection from March to September, 2009, three out of four of their correlated market prices have increased (Dow, NASDAQ and S&P 500), so the corresponding tweet mood should remain positive. Therefore, the main cause of the high frequency of positive words is that the market has an influence on Twitter users' opinions, but not simply because of users' preferences. This is supported by Oh and Sheng (2011), who suggested that online investors tend to be over-confident or over-optimistic as they notice that there is more bullish sentiment than bearish sentiment even in the declining period.

As stated in Chapter 1, this thesis focuses on the sentiment classification of stock tweets, so it is worth looking at what kinds of approaches have been used in previous studies. Table 3.6.1 shows that most of the previous studies are based on the bag-of-words approach, and the rest use the variations of the bag-of-words approach, such as lexical scorers and POS features. Thus, the following sections will review these main approaches in turn. In addition, 7 out of 10 papers in this table did not mention the accuracy of the sentiment classification results at all. Sprenger and Welpe (2010) and Oh and Sheng (2011) achieved best scores of 0.643 and 0.663 respectively by using the bag-of-words model. Smailović et al. (2012) reported a best accuracy of 0.81 in their studies, but the dataset they used to achieve this accuracy is the dataset collected by Go, Bhayani and Huang (2009). Go et al. (2009)'s dataset is a collection of general tweets containing emoticons, which is completely different from the stock tweet data used in Smailovic's later analysis. Therefore, this result is not comparable to the other's results.

Table 3.6.1 Statistics of previous studies of stock prediction based on tweet data

Project	Sentiment classification approach	Best accuracy
Sprenger & Welpe (2010)	bag-of-words	0.643
Xue Zhang et al. (2010)	based on keywords	NA
W. B. Zhang & Skiena (2010)	bag-of-words	NA
Bollen et al. (2011)	bag-of-words	NA
Mao et al. (2011)	based on keywords	NA
Oh & Sheng (2011)	bag-of-words, lexical scorer & both	0.663
Xue Zhang et al. (2011)	based on keywords	NA
Smailović et al. (2012)	bag-of-words (unigram, bigram)	NA
Oliveira et al. (2013)	5 word lists, emoticons	NA
Nann et al. (2013)	bag-of-words, part-of-speech	NA

It is completely shocking and disappointing to see that these papers based on sentiment analysis never reported or misreported the sentiment classification accuracy in their experiments. Without the accuracy, how can other researchers in this area evaluate and refer to these papers?

3.6.1 Bag-of-words Model

The **bag-of-words model**, or the bag-of-features model, is a widely used model in NLP, which performs relatively well in text classification and image recognition. The concept of “bag-of-words” first appeared in Harris (1954)’ work, where it was suggested that “language is not merely a bag-of-words but a tool with particular properties which have been fashioned in the course of its use” (p. 156). In the model of distributional structure, he hypothesised that language consists of different “structures”, and each of them can function individually. Based on this, the distributional structure has an essential feature:

> Some elements are similar to others in terms of certain tests; or are similar in the sense that if we group these similar elements into sets ("similarity groupings"), the distribution of all members of a set (in respect to other sets) will be the same as far as we can discover (p. 158).

Thus, the "bag-of-words, model shares a similar concept: it treats text as 'an unordered bag of independent evidence points' " (Yarowsky, 1993, p. 266). In other words, if an item is put in a black box with some holes on the surface, and if enough holes are provided, the item can thus be recognised. Pang et al. (2002) define this model as:

> Let $\{f_1, . . . , f_m\}$ be a predefined set of *m features* that can appear in a document; examples include the word "still" or the bigram "really stinks". Let $n_i(d)$ be the number of times f_i occurs in document d. Then, each document d is represented by the document vector $\vec{d} : = (n_1(d), n_2(d), . . . , nm(d))$. (p. 81)

Therefore, the model converts word frequency into a vector to be summarised by computing the cosine similarity. Moreover, it is important to notice that either the unigram or the multi-gram model is based on the presence of the parsing result. Particularly the multi-gram model may ignore the meaning of individual segmentation parts.

Applying the bag-of-words model to sentiment analysis, keywords are classified into positive/negative categories, so each keyword is assigned a value representing either positive or negative sentiment. Some studies added a neutral category for unclassified keywords (Pang & Lee, 2008). Roughly, analysts can recognise the general positive/negative feature of the sample text based on the sum of features of the component words.

The main drawback of the bag-of-words approach is that it hypothesises that each of the "pattern" can work individually. This is rarely true in natural language. First, it ignores the linguistic connections of individual words. For example, in the sentence "this is not too bad", although it has two negative words – "not" and "bad", the polarity of the whole sentence is not entirely negative. One typical instance is mentioned by Pang et al. (2002) that "there are many words indicative of the opposite sentiment to that of the entire review" (p. 85), so the bag-of-words model cannot precisely identify the overall emotion in such a domain. In this case, they need sufficient context. If the scope focuses on a relatively restricted context, for example, one sentence or one tweet, the outcome is likely to be problematic. As Nigam and Hurst (2004) pointed out, "since a sentence is substantially shorter than the average document, there will be many fewer features in a sentence bag-of-words than in a document bag-of-words" (p. 5). They proposed this argument in their topic identification work, but it can be applied to sentiment analysis as well. Furthermore, tweet contents are often incomplete, making the extraction of useful bag-of-features much more difficult, as demonstrated in Chapter 7.

Though it has such obvious drawbacks, it is still a mainstream approach in sentiment analysis. Therefore, the main analysis first uses this approach as the baseline and then tests for whether classifying sentiment based on the linguistic features can compete with this approach or whether combining them can achieve a better result.

In practice, most of the previous work used a predesigned sentiment word list to build the bag-of-words model (see Table 3.6.2), and they claimed that these word lists worked well with their datasets.

Table 3.6.2 Sentiment word lists used in previous studies

Project	Word list
Sprenger & Welpe (2010)	Harvard-IV-4 classification dictionary (Ogilvie, Stone, & Kelly, 1982)
W. B. Zhang & Skiena (2010)	NA
Bollen et al. (2011)	OpinionFinder (Wilson et al., 2005a), Google Profile of Mood States (Bollen et al., 2011)
Mao et al. (2011)	Loughran and McDonald Financial Sentiment Dictionaries (Loughran & McDonald, 2011)
Oh & Sheng (2011)	a manually crafted lexicon of bullish and bearish keywords (Oh & Sheng, 2011)
Smailović et al. (2012)	NA
Oliveira et al. (2013)	Harvard-IV-4 classification dictionary (Ogilvie et al., 1982), MPQA Subjectivity Lexicon (Wilson, Wiebe, & Hoffmann, 2005b), SentiWordNet (Baccianella, Esuli, & Sebastiani, 2010), Opinion Lexicon (Hu & Liu, 2004), Macquarie Semantic Orientation Lexicon (Mohammad, Dunne, & Dorr, 2009)
Nann et al. (2013)	NA

However, the validity of these word lists is in doubt, because none of the above predesigned word lists were designed specifically for tweet data and previous studies of linguistics features of tweets highlighted many features distinct to other domains in tweets (as discussed in Section 3.4.1). Therefore, it is necessary to examine them, and if they do not work, a new word list should be designed. Hence, the fourth research question is as follows:

> Question 4. How can a robust sentiment word list for tweet sentiment classification be designed?

This question is addressed in Chapter 7.

3.6.2 Semantic orientation

Another main sentiment classification approach is semantic orientation. Introducing more mathematical tools and machine learning methods allows sentiment analysis to investigate more linguistic details of target texts. As Turney (2002) noted, semantic orientation "is calculated as the mutual information between the given phrase and the word *excellent* minus the mutual information between the given phrase and the word *poor*" (p. 417). This concept can be seen to derive from the idea of semantic field. Lehrer (1974) defined semantic field as "a group of words closely related in meanings, often subsumed under a general term" (p. 1), meaning different sets of words or phrases can be grouped as vocabulary according to different classification criteria. Thus, semantic field considers that there are different "sets which are related to conceptual fields" (p. 15) in language. The semantic orientation model hypothesises that there are two fields: positive and negative. Therefore, each words can be grouped into either field, or the distance to either field can be calculated, with the difference between the two fields can indicating the orientation of a word.

Unlike the bag-of-words model, this approach excludes meaningless parsing results. Another difference is that the result of semantic orientation depends on the calculation of the values assigned between the given phrase and the general term. Therefore, this result is not merely a statistical result of phrase presence, but a weighted result of selected occurrences. Finally, as indicated by Turney (2002), semantic orientation is an unsupervised learning approach, while the bag-of-words model is often used with supervised learning methods.

Nevertheless, one main shortcoming of semantic orientation is that the accuracy is subject to the target text type. The study by Turney (2002) suggests that the result is not stable across different types of texts, for example, the semantic field approach based on movie review and product review data can result in very different accuracy in sentiment analysis.

3.6.3 Emoticon model

Another approach being used in sentiment analysis uses emoticons. Emoticons or smileys, as Read (2005) defined them, are “glyphs constructed using the characters available on a standard keyboard, representing a facial expression of emotion” (p. 45), which often appear at the end of sentences in social media discourses. People prefer using these marks to express or reinforce their attitude; to some extent, this use has been accepted more and more widely, and in some domains, this has become an essential part of communication. Thus, some previous studies paid attention to this problem because they considered this to be an explicit signal of subjective information (Hogenboom et al., 2013; K. Liu et al., 2012; Read, 2005).

3.6.4 Stock tweets classification strategy

As introduced in Section 3.1, sentiment analysis has two classification strategies, either polarity classification or fine-grained classification.

Studies on using general tweets to predict the market trends tend to use the fine-grained approach. Started by Bollen et al. (2009), six sentiment categories have been used in their analysis: tension, depression, anger, vigour, fatigue, and confusion. Then, Bollen et al. (2011) used a different category to classify tweets: calm, alert, sure, vital, kind and happy. In their subsequent work, however, they used a completely different approach, where they considered “a tweet as bullish if it contains the term ‘bullish’, and bearish if it contains the ‘bearish’ ” (Mao et al., 2011, p. 3). As shown in Table 3.1.1, most of the Stanford final projects followed this approach (Chakoumakos et al., 2011; Hsu et al., 2011; Mittal & Arpit, 2011).

On the other hand, studies on using ticker-specific tweets to predict individual performance tend to use the polarity approach. Smailović et al. (2012), Tushar Rao & Srivastava (2012), and Oliveira et al. (2013) used the literally “positive” and “negative” categories to define the sentiment in tweets. However, Sprenger and Welpe (2010) used a slightly different definition of “positive” and “negative” in their research, in which they consider *bullish* as positive, and *bearish* as negative. Oh and Sheng (2011) and Nann et al. (2013) used the same strategy to classify their tweet data.

Using polarity within the stock tweet context is not as simple as in other contexts because there is a disagreement between the market sentiment and the linguistic definition of the polarity. Sprenger and Welpe (2010) pointed out that “buy and sell signals may carry very different information with respect to subsequent stock returns” (p. 10). This indicates that the relationship between investment action market trend does not match perfectly. They then explained different polarities respectively.

1. Positive emotions frequently occur in buy signals.
2. Negative emotions frequently occur in sell signals.
3. Positive and negative emotions equally frequently occur in hold signals (pp.26-27).

Many studies on this topic classify tweets into such a polarity categories, but few have discussed the classification criteria in detail. Does a “positive” stock tweet mean “buy”? Does a “negative” stock tweet mean “sell”? And does a “neutral” stock tweet mean “hold”? The question form the basis for the fifth research question:

> Question 5. Does a more precise definition of a positive, neutral, or negative stock tweet in accordance with market values help to improve the quality of stock tweet sentiment analysis?

This question is addressed in Chapter 4.

The disagreement between the relationship of the market trend and investment action increases the difficulty of designing a way of classifying ticker-related tweets, and it also highlights the difficulty of analysing the correlation between the market and the sentiment.

Another fundamental problem in the polarity approach is how to process the neutral sentiment. Some of the studies assigned unclassified data to the neutral category as noted by Pang and Lee (2008). This is confusing because the unclassified data probably contain some irrelevant contents: Thus, classifying them as neutral tweets and regarding them as relevant content can cause confusions. In addition, some of the analyses simply removed the neutral category when they compute the overall sentiment score (Oh & Sheng, 2011). This is also problematic, because if normalising the sentiment score is necessary, then the neutral data counts is an important part in the score. Therefore, removing them can generate a different, and potentially inaccurate, result. Because of this, another question must be addressed:

> Question 6. How can the neutral sentiment category of stock tweets be defined and processed?

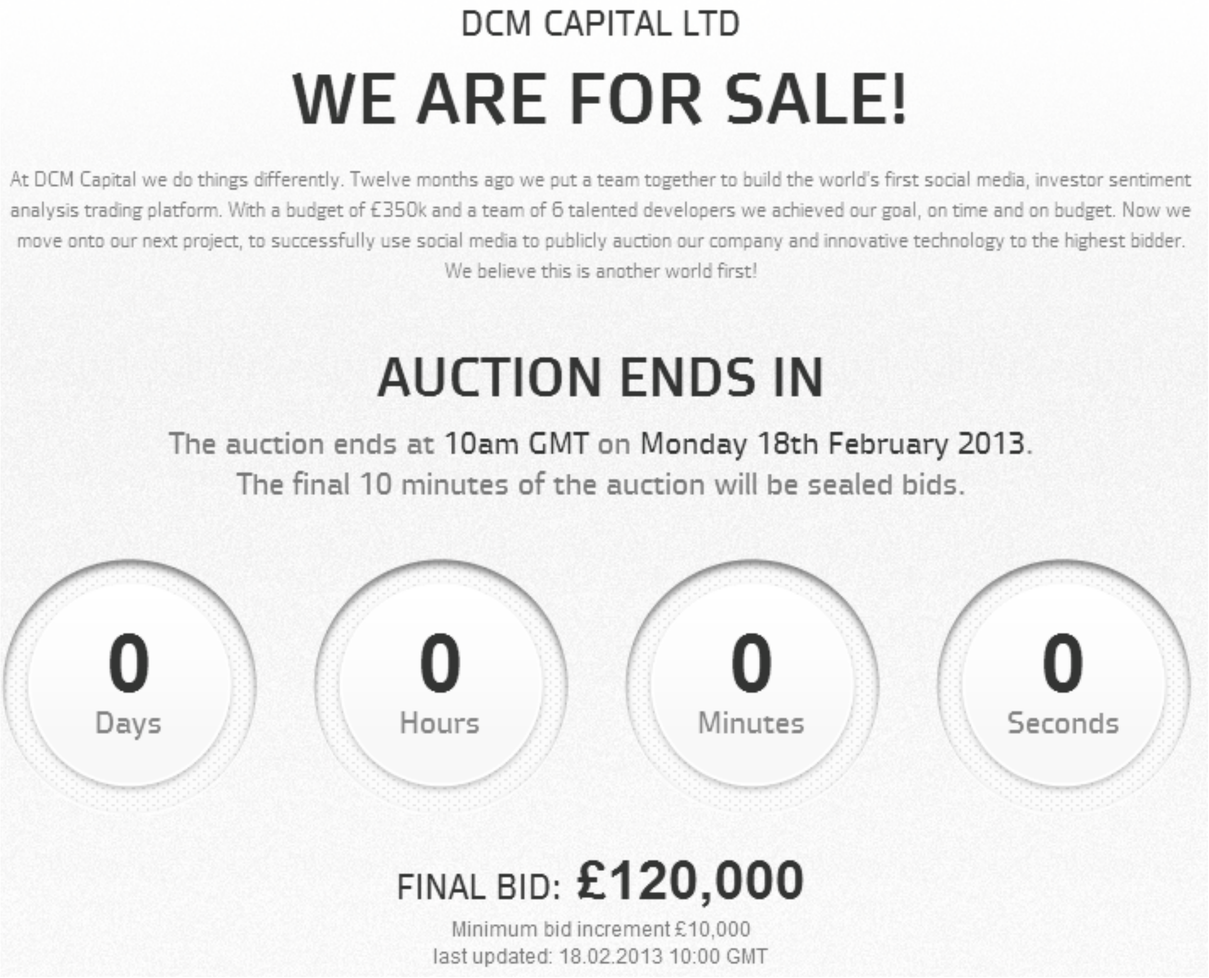
This question is addressed in both Chapters 5 and 8.

3.7 Summary

This chapter first introduced the background knowledge to sentiment analysis and stock prediction based on tweet data. As the previous literature shows, there are several unanswered questions in this area. First, what is a stock tweet? Second, do stock tweets have specific linguistic features? Third, what is a positive, neutral, or negative stock tweet? Fourth, can neutral sentiment exclude the unclassified data? The following analyses try to answer these questions.

Rather than understand the correlation between tweets and the market in general, this research is more interested in the correlation between tweets and the individual ticker's performance. Thus, it aims to develop an approach to improve the accuracy of sentiment analysis based on previous studies with this focus. In addition, current approaches to sentiment analysis make little use of linguistic features. For example, the bag-of-words approach neglects the semantic relationship in the sample. Thus, the present research tests whether better accounting for the linguistic features of stock tweets can improve the accuracy of sentiment classification.

As discussed above, Bollen et al. (2011) made the first attempt in this area, and soon after their paper was published, they received 25 million pounds to establish a hedge fund Derwent Capital Markets in London (Jordan, 2010). However, two years later, this hedge fund was for sale in an online auction (Sukumar, 2013). The final bid was only 120,000 pounds (see Figure 3.7.1). In an ironic statement, the owner of the hedge fund claimed, "[n]ow we move onto our next project, to successfully use social media to publicly auction our company and innovative technology to the highest bidder" (Sukumar, 2013).



DCM CAPITAL LTD

WE ARE FOR SALE!

At DCM Capital we do things differently. Twelve months ago we put a team together to build the world's first social media, investor sentiment analysis trading platform. With a budget of £350k and a team of 6 talented developers we achieved our goal, on time and on budget. Now we move onto our next project, to successfully use social media to publicly auction our company and innovative technology to the highest bidder. We believe this is another world first!

AUCTION ENDS IN

The auction ends at 10am GMT on Monday 18th February 2013.
The final 10 minutes of the auction will be sealed bids.

0 Days 0 Hours 0 Minutes 0 Seconds

FINAL BID: £120,000

Minimum bid increment £10,000
last updated: 18.02.2013 10:00 GMT

Figure 3.7.1 The screenshot of Derwent Capital Markets' final online auction

Chapter 4 Tweet Data Collection and Annotation

As data-driven research, the data collection is fundamental to this thesis: Not only does the research make aggressive investigations into the data, the enormous quantity of data also pose additional problems, so thoughtful considerations are necessary. Therefore, this chapter first briefly discusses the difficulties of undertaking this interdisciplinary project, which includes the sparseness found in massive tweet data sets, the frequent changes of Twitter's data releasing policy, the presence of non-standard expressions in tweets, the disagreement of the trinary relationship (i.e., the imperfect relationship between how tweets are classified as positive, negative, or neutral in terms of sentiment and how the market is defined as bullish, bearish or hold), and the different timings of when the stock market is open and when people post tweets. This chapter then introduces the collection procedure of stock tweets. This type of data is difficult to collect due to the tight restrictions of Twitter, so the manipulation of raw tweets is introduced and the scheme for the manual annotation of stock tweets is explained.

4.1 Difficulties of Data Preparation

Previous literature suggests that researchers face a number of difficulties with both the Twitter platform and tweet data. These difficulties include the following aspects: (a). Twitter has massive amounts of data, (b). its data releasing policy is not stable, (c). tweets contain a high frequency of the non-standard expressions and ironic statements, (d). the tweet sentiment does not correspond to market data well, and (e). the timing of stock tweets is different from the stock price data. This section briefly outlines these difficulties, and the following chapters discuss more details where relevant.

4.1.1 The paradoxical sparseness of massive tweet data sets

In theory, it is better to analyse a larger set of tweets. However, Twitter currently has enormous quantities of data, so retrieving the relevant data is extremely difficult. For instance, in 2010, more than 25 billion tweets were posted on Twitter (Costolo, 2010). As a result, it is neither possible to collect the entire dataset, nor realistic to analyse it. O'Connor, Balasubramanyan, Routledge, and Smith (2010) pointed out that it is technically simple to retrieve millions of tweets daily, but that few of them relate to the research purpose. Bollen et al. (2009) made a similar point, finding that only 10% of the collected data could be used for their later sentiment analysis.

In addition, tweet data are unbalanced. Oh and Sheng (2011) noticed that the top frequent tickers account for the majority of stock tweets in their collected data, as not all stocks are equally discussed on Twitter (Nann et al., 2013). The unbalanced tweet data create several difficulties to the automation of the analysis, including imposing high computational costs and overfitting the data (Yi, 2009).

For this research, relevant random stock tweets were collected and analysed as defined in Section 1.2. As shown later in Section 4.2, the method used could only retrieve dozens to hundreds of relevant data each day. Thus, the research focused on using a limited sample of the data to conduct the analysis. From April 2012 to February 2013, about 300,000 stock tweets were collected for this research. Compared to other stock prediction projects based on tweets, this quantity of data is relatively limited. However, to carefully analyse the linguistic features in detail, a small but well-sampled dataset was a better choice.

4.1.2 Frequent changes of Twitter's data releasing policy

Twitter changes its data releasing policy frequently and unexpectedly, which brings considerable difficulties to researchers. As Zappavigna (2012) noticed, "the way Twitter allows developers to access its data and the extent of the granted access is changing" (p. 24), so any currently available collecting method will soon become less applicable or even inapplicable. There are two aspects to this problem: Twitter forbidding the re-release of tweet data or restricting API usage.

For the first aspect, there have been a number of large scale tweet corpora developed for research purposes by different projects. For instance, the Edinburgh tweet corpus contained 97 million tweets when it was first released (Petrovic, Osborne, & Lavrenko, 2010), and the Stanford SNAP (Stanford Network Analysis Platform) tweet corpus contained 467 million tweets (J. Yang & Leskovec, 2011). However, soon after their respective releases, Twitter asked them to remove the public accesses to the data. Curtailing the data sharing within academia brings tremendous problems, because building a tweet corpus on such a scale needs many resources and much effort, making it beyond the capabilities of many researchers. Thus, restricting content sharing means that researchers from different institutions needs to constantly reinvent the wheel by building corpora themselves.

The second aspect is more complicated: A number of tweet corpora have used APIs to collect tweet data, but this is no longer an easy option to apply. As Twitter attempts to monetize their data release, they have curbed API requests. Generally, data can be retrieved from the Search API or Streaming API. The first one only randomly releases very limited and highly repeated data (see Section 4.2.6), and the second one requires a registration. Although the Steaming API releases more data than the Search API, it has an hourly request limit: For a normal Streaming API, the hourly limit is 150, and the "Whitelist" Streaming API can request as much as 20,000 times per hour (Twitter, 2012f). Furthermore, after registration, the Streaming API requires OAuth 2.0 protocol[4] to verify the user when they are retrieving data (Twitter, 2012a). Both APIs are not easy to use, so they create more obstacles to researchers, especially to those without a programming background.

[4] Oauth is an open sourced application protocol, which significantly simplifies the authentication process for a third-party application.

Gradually, Twitter has restricted access to free data by the public via its APIs. Below is a list of the major changes in the Twitter API in recent years:

1. 2006-09-20 Twitter began to introduce APIs to release data in JSON and XML format (B. Stone, 2006).

2. 2009-01-20 Twitter set the ceiling hourly request of each whitelisted API to 20,000 requests (Payne, 2009).

3. 2011-02-11 Twitter stopped providing the whitelisted API to developers, which has an hourly limit of 20,000 requests (Melanson, 2011).

4. 2011-02-11 Twitter began to forbid others to "sell, rent, lease, sublicense, redistribute, or syndicate access to the Twitter API or Twitter Content to any third party without prior written approval from Twitter" (Twitter, 2011a). This points out that, even for the research purpose, no one can redistribute tweet data.

5. 2012-08-16 Twitter planned to release the new Twitter API (version 1.1), in which "an application that only accesses one endpoint may be more restricted" (Sippey, 2012).

6. 2012-09-05 Twitter released the new Twitter API (version 1.1), which sets "the rate limit window into 15 minute chunks per endpoint, with most individual calls allowing for 15 requests in each window" (Twitter, 2011a).

7. 2013-02-05 Twitter planned to stop using version 1.0 of the Twitter API in March 2013, so all protocols based on API 1.0 will be invalid (Singletary, 2013).

Clearly, Twitter has restricted the release of data continually. Not only does reducing the API rate limit restricts the quantity of data that can be collected, but also the frequent changes require crawlers to change accordingly. Moreover, sometimes Twitter suddenly appears to stop releasing retweet data through the Search API without any explanation as observed during the data collection for this project (see Section 4.3). This occurred without the crawler violating the rate limit – it kept at the same rate as before. Another case of an apparent malfunction occurred during January, 2013: The Search API suddenly released 533 tweets posted during 2010 or 2011, about 0.44% of all tweets collected from the ticker tweets crawler in the previous 9 months.

These additional difficulties caused by Twitter make data collection even more challenging.

4.1.3 Non-standard expressions in tweets

The third problem relates to the linguistic features of tweet data. Unlike other domains observed, tweet data contain a high ratio of non-standard expressions:

1. They contain emoticons;
2. They have RT or @ signals;
3. They contain URL addresses;
4. They contain misspellings;
5. They mix lowercase and uppercase;
6. They omit subject;
7. They contain text speaks;
8. They contain non-standard verb forms;
9. They contain ironic statements (Go & Bhayani, 2010);
10. They contain additional html or Java escape characters to represent emoticons.

These features bring numerous challenges to data processing, and as Kumar and Sebastian (2012) pointed out that the unique challenge of twitter sentiment analysis is mainly due to the informal tone. For example, the last feature adds additional characters to tweets, as a result, some of them have more than 140 characters. Thus, they require extra attention to deal with, as outlined in Section 4.2.

4.1.4 Disagreement of the trinary sentiment relationship

This research followed a trinary model of sentiment classification that classifies sentiment as "negative", "neutral" and "positive". The intention is to echo the market changes as "bearish", "hold" and "bullish" respectively. However, as explained later in Section 4.4.6, there is another trinary relationship in the market: The investor's reaction as "selling", "holding" and "buying". These three relationships do not perfectly match each other, although they have large overlaps. These disagreements bring further difficulties to either the annotation, or the further temporal correlation analysis.

4.1.5 Different time scale of tweets

Tweet data are often regarded as "streaming data" as they present information in a sequence (Zappavigna, 2012), and this sequence is a never-ending sequence. However, the stock price data have certain breaks, because no market opens during holidays. This translates into huge differences between these two groups of data. Applying time series analysis on them becomes challenging. A further discussion of this can be found in Chapter 5.

4.2 Stock Tweet Collection

The main challenge of this entire data collection is to collect tweets from the Twitter Search API, so this section discusses extensive details of the collection procedure, including the collection criteria, raw data structure, Twitter Search API, collection keyword selection, server configuration, and collection frequency.

4.2.1 Collection criteria

Pettit (2011a, 2011b, 2011c, 2011d) gave a four-step guidance for collecting social media data, taking into consideration about data content.

1. Collect social media data that mentions cookies. Does the kind of cookie matter?

The first step points out that social media data often contain ambiguous or irrelevant contents, so extracting the relevant contents is extremely important.

2. Clean the social media data. Do we really want viral games in our dataset? It depends on the research objective.

The second step is a post-collecting procedure, because Twitter contains a large portion of spam. According to Finger and Dutta (2013), the percentage of spammed tweets was 11% in 2012, reducing to 1% in 2013. In a huge dataset, even 1% can be a very large amount, so cleaning is often a key step in data processing. Moreover, André, Bernstein, & Luther (2012) pointed out that only 36% of tweets are worth reading, which indicates that the topic on Twitter is limitedly relevant to readers' interests. Hence, for specific extraction tasks, the useful information from a tweet dataset is rather limited, which requires more careful data investigations.

3. Did you forget to code slang? Emoticons? Does it matter? YES!

The third step is also a key step in tweet processing due to the linguistic features of tweet contents, because Kaufmann (2010) suggested, tweets contain "an unusually high amount of repetition, novel words, and interjections" (p. 1). Indeed, the research results are not only affected by emoticons, code slang, novel words, repetitions or interjections, but also hashtags, retweets and replies.

4. Sample precisely. Or wherever you're allowed. Or wherever you remember to look.

The last step requires the most attention, because data representativeness is often the central question in corpus research (Hunston, 2002). Although it is not easy to understand how Twitter is used in trading, random sampling should be used. There is an increasing trend of applying Twitter in trading; for example, Bloomberg Terminal has integrated Twitter as an important information source as mentioned in Chapter 1 (Indvik, 2013).

Moreover, Escobar (2011) provided a detailed guide for selecting a proper tool for monitoring social networks. It has 22 main aspects, including time, content, size (see Table 4.2.1). Following his suggestions, the research used the relevant elements (in the first two rows of Table 4.2.1) to improve the selection criteria.

Table 4.2.1 Escober's (2011) suggestions for monitoring social network

Data coverage	**Data retrieval**	**Content**
Geographic scope	**Tool installation**	**Spam**
Reporting capabilities	Top USP's	Training
Data reach	Cost structure	Search
Value added services	Integration	Clients
Franchisee based features	Data presentation	Metrics
Minimum contract period	Term search speed	Engagement
Contract termination notification		

According to this guide, this research looked to include three types of tweets: Tweets containing financial news account names (including breaking news accounts), retweets of the first group, and random tweets containing any of 30 DJIA ticker hashtags. Accordingly, the keyword list was built as introduced in Section 4.2. In total, 220 unique search queries were in the keyword list. The configuration of these search queries is introduced in the following sections.

As discussed in Chapter 2, Twitter provides a charged service to access their data through different data reseller; however, this research was sensitive to cost and required specific data rather than general data. Thus, all data were collected via the Twitter Search API based on the specific criteria discussed below. The data collection has lasted for 10 months to retrieve sufficient data: It began at the end of April 2012, and ended in February 2013.

Instead of classifying the geographic information in the meta data, this research chose, the platform language, English to reduce the number of non-English tweets. In this research, geographic boundaries are not particularly relevant because traders usually move around, or they may have offices in different places. Also, there might be a relatively large portion of spam tweets in the collection, especially in the random tweet queries. This is discussed later in the following sections.

The whole data collection relies on an external Ubuntu server with 512MB RAM and 20GB hard disk. The tools used in this tasks are free command line tools, including Unix command line tools such as `wget`, `crontab`, and two other programming languages, `Python` and `R`.

4.2.2 Raw data structure

Sample 4.2.1 is a random tweet in JSON format retrieved from Twitter. JSON is a lightweight language for storing complicated data, and Twitter uses it as the default format to store data. Actually, The original JSON file is much larger, so only the main parts are introduced here: The metadata of the tweet, the entities included in the tweet, and the tweet content.

```
},
        "created_at": "Tue Oct 25 21:11:00 +0000 2011",
        "retweeted": false,
    "in_reply_to_status_id_str": null,
    "in_reply_to_status_id": null,
    "source": "web",
    "in_reply_to_user_id_str": null,
    "truncated": false,
    "id": 1289416549032427**,
    "entities": {
        "hashtags": [
            {
                "indices": [
                    81,
                    86
                ],
                "text": "CLSP"
            }
        ],
        "urls": [
            {
                "url": "http://t.co/XXXXXXXX",
                "expanded_url": "http://vimeo.com/clsp/XXXX",
                "indices": [
                    104,
                    124
                ],
                "display_url": "vimeo.com/clsp/fernando-…"
```

```
29              }
30          ],
31          "user_mentions": [
32              {
33                  "screen_name": "earnmytur**",
34                  "name": "Fernando Perei**",
35                  "id_str": "1973135**",
36                  "indices": [
37                      54,
38                      66
39                  ],
40                  "id": 1973135**
41              }
42          ]
43      },
44      "contributors": null,
45      "text": "I uploaded \"Are Linear Models Right for Language?\"by
        \@earnmyturns from the 2008 #CLSP seminar series.
        http://t.co/XXXXXXXX"
46  }
```

Sample 4.2.1 A tweet in the complete JSON format

Lines 2 to 9 provide the meta information of the tweet: Line 2 indicates the post time of one tweet, and the time is converted into Greenwich Mean Time (GMT-0), which is convenient for further sorting tweets. The sample, as suggested, was created at 21:11:00, Oct 25, 2011 according to GMT-0 time. Line 3 indicates whether a tweet is retweeted. This sample tweet is not a retweet.

Lines 10 to 43 are the *entities* according to Twitter: Line 11 to Line 19 indicate if the tweet contains a hashtag or not. If it does not contain a hashtag, then Line 12 to Line 19 will be omitted; if it has a hashtag, then Line 17 is the content of that hashtag. In this sample, it contains a hashtag "CLSP". Line 20 to Line 30 indicate whether it contains a URL. If the tweet does not have a URL, then Line 21 to Line 29 Line will be omitted. In this sample, the Twitter shortened URL is "http://t.co/XXXXXXXX", the original URL is "http://vimeo.com/clsp/XXXX". Lines 31 to 42 indicate if it mentions another user. In this sample, a user is mentioned, and its user ID, user name and user screen name are provided.

Line 45 is the tweet content, what can be seen on screen. The content here is:

> "I uploaded "Are Linear Models Right for Language?" by @earnmyturns from the 2008 #CLSP seminar series. http://t.co/XXXXXXX"

In this example, only one line, the tweet content, is displayed to general users; the rest is meta information for Twitter's processing and development. This research utilised only parts of this information, time information and the original tweet content. The following sections explain how to extract these two types of information.

4.2.3 Twitter Search API

As discussed in Chapter 2, this research used the REST Search API to collect data, and this section introduces the details of applying the REST Search API.

In the API version 1.0, the REST Search API has a rate limit based on "complexity and frequency", instead of the API request (Twitter, 2012b). This brings a problem because Twitter has not released any specific explanation of either complexity or frequency. There are only some implicit descriptions of Search API rate limit, because “not every tweet can be indexed in Twitter Search” (Twitter, 2012b). Hence, one month of the data collection period was spent on understanding the best rate of relevant results, which will be discussed in Section 4.2.6.

The basic format of a Search API URL looks like this:

> http://search.twitter.com/search.format

Different parameters can be used to substitute the *search.format* in the above URL. Twitter provides a compulsory parameter *#q* of search keyword and twelve optional parameters. To satisfy the analysis, these optional parameters were chosen: *lang* as the operating platform language to minimise the number of non-English tweets, *resulttype* as the filter to obtain the most recent tweets, and *rpp* as the number of tweets in each feedback.

Here is an example of the ticker search query URL:

> http://search.twitter.com/search.JSON?q=%24KO&result_type=recent &lang=en&rpp=50&

The keyword in this URL is $KO. The dollar sign is encoded in ASCII format as %24, because all parameters must be properly URL encoded (Carey, 2012). Accordingly, the hash sign # is encoded as %23, whereas the at sign, @ is encoded as %40. The optional parameters in this query include *resulttype* as recent, *lang* as English, and *rpp* as 50 results per page.

4.2.4 Keyword selection

To reach the primary goals of the data collection, this research carefully selected keywords and grouped them into three categories: ticker group, media group, and retweets group.

Considering the massive quantity of data, the research focused on the Dow Jones Industrial Average (DJIA), which contains 30 companies in total (see Table 4.2.2). As described in the Dow

Jones Index overview (2011), these 30 companies are “all major factors in their industries”, and the index “provides nearly complete coverage of the U.S. stock market”. Thus, due to the representativeness of these 30 tickers on the market, it is reasonable to assume that they are adequately discussed on Twitter, especially in the anglophone “Twittersphere”. Thus, theoretically, this data collection procedure retrieves sufficient data. In addition, focusing on 30 tickers significantly reduces the complexity of the research later.

Table 4.2.2 List of 30 companies and their tickers in the Dow Jones Industrial Average

Ticker	**Company**	**Ticker**	**Company**	**Ticker**	**Company**
MMM	3M Company	MSFT	Microsoft Corp.	CSCO	Cisco Systems Inc.
INTC	Intel Corp.	AA	Alcoa Incorporated	PG	Procter & Gamble Co.
IBM	IBM Corp.	JNJ	Johnson & Johnson	DD	E.I. DuPont de Nemours & Co.
PFE	Pfizer Inc.	KO	Coca-Cola Co.	AXP	American Express Co.
SBC	AT&T Inc.	SPC	Travelers Cos. Inc.	CHL	JPMorgan Chase & Co.
DIS	Walt Disney Co.	XON	Exxon Mobil Corp.	UTX	United Technologies Corp.
BA	Boeing Co.	GE	General Electric Co.	BEL	Verizon Cummunications Inc.
KFT	Kraft Foods Inc.	MCD	McDonald’s Corp.	HWP	Hewlett-Packard Co.
CAT	Caterpillar Inc.	MRK	Merck & Co. Inc.	WMT	Wal-Mart Stores Inc.
HD	Home Depot Inc.	CHV	Chevron Corp.	NB	Bank of America Corp.

The second group is the media group. Some breaking news accounts were selected to show the tweeting behaviours of news accounts as shown in Table 4.2.3. The second column indicates the created date of each account. The last column shows the average tweet count of an individual day. Though the present research intended to crawl tweets posted by these media accounts, using the Search API, this is not possible because the Search API (version 1.0) cannot provide tweets of a specific user. Thus, using these account names as keywords only enables the retrieval of tweets mentioning these users, but not tweets posted by them.

Table 4.2.3 Statistics of some breaking news Twitter accounts

Account	**Created date**	**Average daily tweets**
BreakingNews	13-05-2007	43.43
cnnbrk	02-01-2007	22.58
BBCBreaking	22-04-2007	18.14
msnbcbreaking	04-01-2008	3.57
wsjbreakingnews	09-03-2009	1.57
SkyNewsBreak	04-11-2009	9.43
BreakingNewz	20-07-2009	NA
CNBCbrk	25-03-2009	22.43
USABreakingNews	04-08-2008	0.43
ABSCBNBreaking	17-04-2009	1.43
newsoftheday	29-12-2009	84.57
breakingstorm	25-08-2011	3.14
aubrk	02-08-2009	40.71
BreakingNews	03-05-2008	63.43

The third group is the retweet group. Accordingly, the retweet format of the media groups above, “RT @account_name”, were selected as the keywords. Similarly, these tweets only contain the keywords, but are not the retweets posted by these media accounts.

In total, 220 keywords were selected, and all of their query URLs were written in the same format.

4.2.5 Server configuration

The main crawling scripts ran on a Ubuntu server. Ubuntu is a popular distribution of Linux, so configuring Unix script on it is fairly easy. All scripts use `wget`, a command-line download tool, to collect data according to a list of specific URL addresses. Different types of tweets were collected separately, and all of them were then stored according to the written time. Different crawlers ran according to different schedules, which were set by another command-line timer tool `crontab`. The four-level hierarchy folder structure that was used is shown in Sample 4.2.2:

> Sample 4.2.2 Type of tweets / written time / Main search URL / File named after the search query

Three types of tweets were collected from Twitter in three crawlers:

1. The *ticker tweet crawler* collected ticker tweets of the 30 DJIA companies,
2. The *media tweet crawler* collected tweets containing the names of the 95 media accounts names.
3. The *retweet crawler* collected retweets of these 95 media accounts.

Accordingly, three URL address lists, named ticker address list, media address list, and retweet address list, contain the specific URLs. In total, three URL address lists contain 220 unique URLs.

4.2.6 Collection rate

The primary goal of data collection was to maximise the retrieval efficiency. Retrieving as many tweets as possible and reducing the redundancy of irrelevant ones as far as possible. To balance these two tasks is not easy due to Twitter's strict data policies.

Some known difficulties are challenging. First, the quantity of data is tremendous. If every single tweet were crawled, the data could neither be stored nor be processed by an individual researcher. Second, Twitter constantly changes their data releasing policy as mentioned in Section 4.1.2, and they have strict request limits. Hence, it was proper to collect data according to a set schedule.

The biggest challenge was to find what the most efficient collection rate is. A proper collection rate needs to balance two conditions: maximising the number of unique tweets and avoiding the violation of the Twitter's collection restrictions. As discussed in Section 2.2.2, Twitter has not released any specific frequency restrictions of the Search API, so the only solution was to adjust the rate by experimentation.

The ticker tweet crawler started running from April 1, 2012. At first, the rate was set to collect every 15 minutes. Figure 4.2.1 shows that the first five days obtained about 200,000 tweets at each day, and each URL obtained about 6700 tweets. However, there were only about 2500 unique tweets in total, so each URL only had 83 tweets on average. This indicates that the redundancy is massive: Each tweet repeated about 80 times in the process of collection. The possible reason for this redundancy is that Twitter may have only released a very small portion of tweets on each calendar day, so the crawlers repeatedly collected data from the same proportion.

Thus, the collection rate was reduced to every 30 minutes. From April 6 to 26, the script collected over 70,000 raw tweets and 3,000 unique tweets everyday. This result showed a slight increase compared to the previous five days, but it is clear that reducing the collection rate did not affect the number of unique tweets very much. Instead, the number of unique tweets rose slightly. At this rate level, each individual URL collected about 2,300 raw tweets, with about 100 unique tweets.

To further test the collection rate, it was reduced to a 4-hour slot on April, 27. Obviously, from Figure 4.2.1, the result was not optimistic: The number of unique tweets decreased sharply. Thus, the rate is changed to every hour. From April, 28 to May 11, about 50,000 raw tweets were collected each day, and the number of unique tweets reminded at nearly 2,900. Considering the ratio of raw tweets to unique tweets, the one-hour rate appeared to be the best. Therefore, the ticker tweet crawler was set to collect every hour thereafter.

The media tweet crawler used an address list much longer than the ticker tweet crawler's, so it was

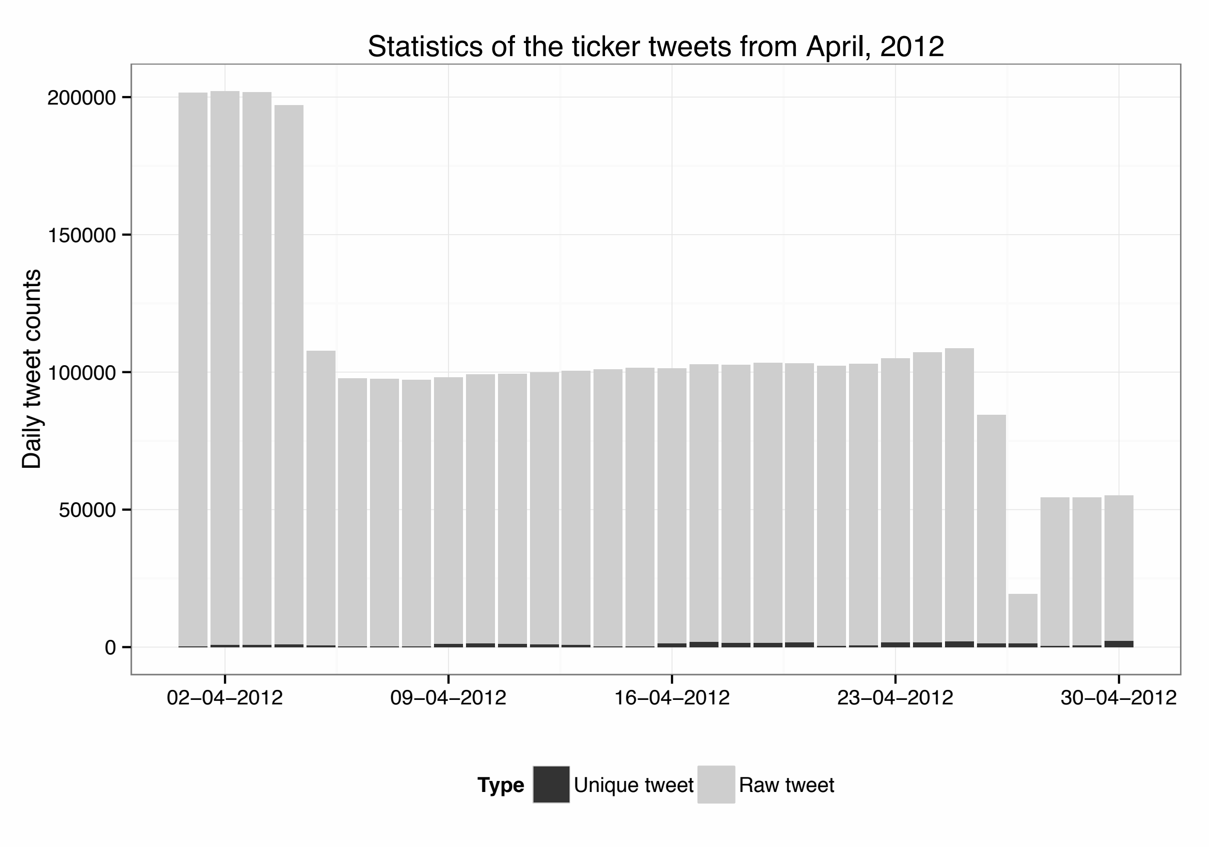


Figure 4.2.1 The statistics of ticker tweets in April, 2012

easier to violate Twitter's restriction. Thus, the collection rate was set to twice a day. Coincidently, one parameter of `crontab` for the media tweet crawler was set to the wrong number, so from April 5 to 20, the crawler kept running every minute for 2 hours each day. As presented in Figure 4.2.2, the number of raw tweets exceeds 400,000, while the number unique tweets is less than 8000. One thing that should be borne in mind is that the address list of this ticker contains 95 URLs, which means each of them obtains more than 4,200 raw tweets but only 84 unique tweets.

Interestingly, this unexpected mistake suggests that even an intensive crawling job like this did not violate the request restriction of the Search API. Rather, the tweets released from the Search API are extremely limited, meaning that each specific keyword would only obtain less than 100 results each day on average.

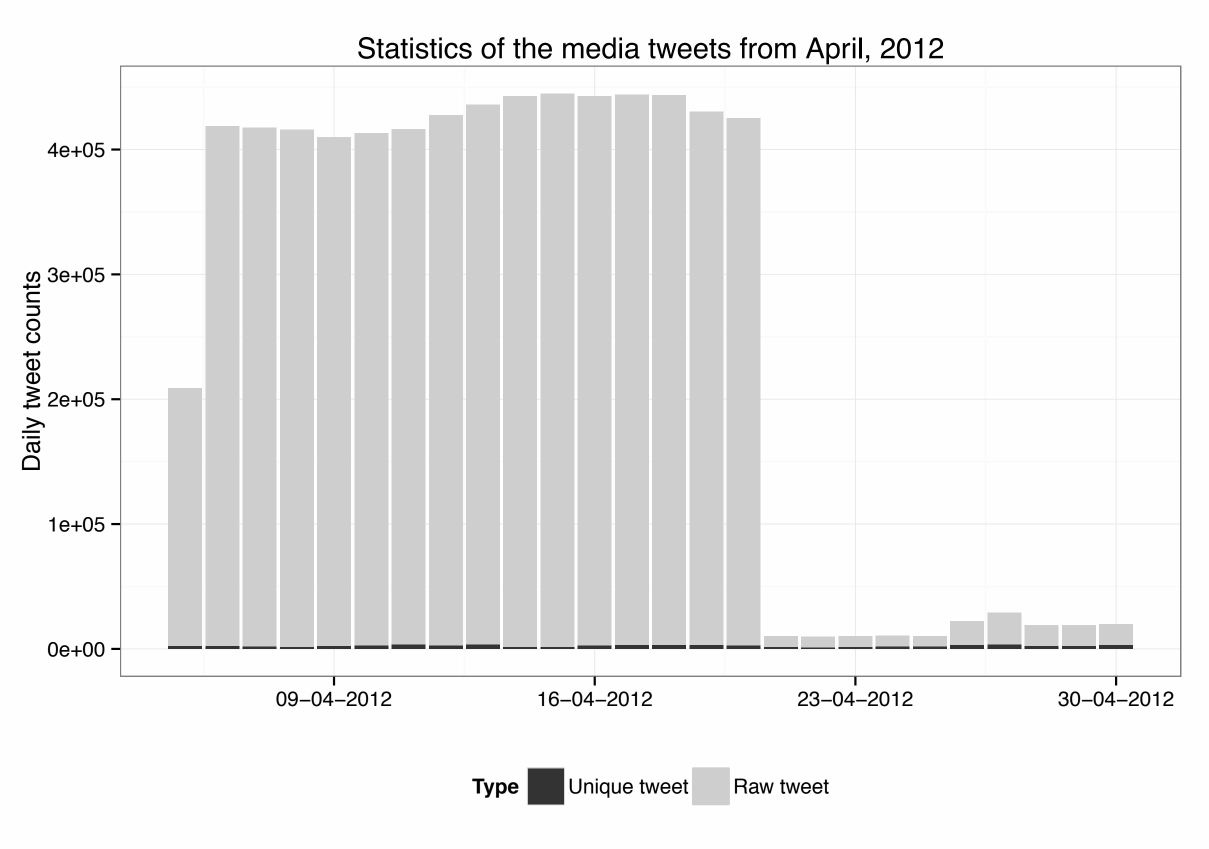


Figure 4.2.2 The statistics of media tweets in April, 2012

The crawlers was then set to collect twice a day. Therefore, from April 21 to 25, the crawler collected about 8500 raw tweets and 5500 unique tweets on each calendar day. At this rate, the number of unique tweets decreases remarkably. Hence, the rate was adjusted to four times, six times and twelve times per day. The best raw/unique tweet ration was at the six-hour rate in Figure 4.2.2. According to this, the collection rate for media mention tweets was set to four times per day.

The collecting rate of the retweet crawler was set to the same as the media mention tweet crawler: It collects tweets every six hours. After the configuration, each URL can obtain about 50 raw tweets and 20 unique tweets each day.

To summarise, the final collection rate of each crawler was set to the time intervals as shown in Table 4.2.4.

Table 4.2.4 Final collection rate of each crawler

Crawler type	**Final rate**
Ticker crawler	Every hour
Media crawler	Every 6 hours
Retweet Crawler	Every 6 hours

4.2.7 Summary

This section first discussed the criteria in tweet collection. It then introduced the raw tweet data structure and Twitter search API in depth. Also, it explained the design of the tweet crawlers in this research, including the selection of the query keywords, server configuration and collection rate of each crawler.

4.3 Raw Tweet Manipulation

After crawling data from Twitter, the next task was to extract the relevant information from the raw data. The aim of this stage remained as the same as that of the above data collection: to maximise the efficiency of data retrieval.

The tweet samples in the following sections are formatted so that the first line shows the original content, and the second lines shows the changed content.

4.3.1 Original folder structure

Although the four-level hierarchical directory works well for storing the crawling results, it was not easy to process due to the following reasons: First, the extreme sparseness of data was even greater than expected (although the frequency of the crawling job has been reduced, there is still a high proportion of repeated tweets); second, the multi-level hierarchic directory is not easy to manipulate; and third, the bottom level is named as the search query, so they became too long to read. Thus, the first goal for the data manipulation was to reorganise the folder structure to overcome the above shortcomings. Furthermore, in order to reduce the quantity of data, it was better to merge the crawling results to eliminate repeated contents.

The new folder structure was designed as outlined in Sample 4.3.1:

> Sample 4.3.1 Type of tweets and written date/Main search URL/Keywords as file names

This slight change significantly reduced the complexity of the original storing structure. The original data were stored according to the specific collection time, so there was more than one file for an individual crawling keyword each day. Also, at the beginning of the crawling task, the frequency was higher, so each day had more folders. These two problems generated multiple crawling result files to each individual keyword, so reorganising them made them easier to process for the time series analysis later. Most importantly, the new filenames were named based on the keyword instead of the search query, so they became more readable to humans.

4.3.2 New record format

As shown above, the original files are in the JSON format: It is convenient for machine to identify different levels of information but not for humans. The structure of JSON is different from the XML format, which is one of the traditional ways to store complicated data. XML is line-based, so each level of information has an individual line or multiple lines; but for the JSON format, each file has only one line, and different levels of information are labelled by different types of bracket marks. In section 4.2.2, Sample 4.2.2 showed a parsed JSON file, which is completely different from the original JSON file. Without parsing, a JSON file looks like a massive data string but in only one line. One thing that should be borne in mind is that each JSON file is a complete and individual file, and they cannot be simply merged together. If several JSON files are just copied and pasted together, a JSON parser cannot read the merged file. However, using the built-in JSON parsing tool of Python, all JSON files were treated as line-based files; thus, the later cleaning procedures were easy to apply.

In addition, the original JSON file provides complete information of one tweet as shown in the Sample 4.2.1. However, the analysis only requires limited information: the tweet category, crawling keyword, post time, and tweet content. Thus, part of this phase was to extract the needed information and store them as a tab-separate file, providing convenience for later analyses.

The first step was to extract relevant information from the JSON files, and store them as a tab-separate file. An example of this is shown in Sample 4.3.2, where the first column indicates the specific post time of one tweet, and the second column is the tweet content.

> Sample 4.3.2 30 May 2012 03:59:35 RT @ABSCBNBreaking: http://t.co/HwPqhjot will be down for a short while due to maintenance. Please check back, it will be up again soon. Thanks for reading!

Then, as shown in Sample 4.3.3, the second step was to extract tweets collected in the same month to an individual file. The format was slightly changed: The first part indicates the collection date, and the rest remains as same.

> Sample 4.3.3 M01-06-12 @ABSCBNBreaking 30-05-2012 03:59:35 RT @ABSCBNBreaking: http://t.co/HwPqhjot will be down for a short while due to maintenance. Please check back, it will be up again soon. Thanks for reading!

Additionally, the untagged file is more human-readable and economic in terms of size. Therefore, at this stage the raw data are half-prepared, but several problems remain, and these are explained below.

4.3.3 Escape characters

Twitter uses the XML character entity to store some of the punctuation marks and emoticon marks, which brings extra difficulties. For example, the apostrophe ' is stored as *&#39*. This issue is important to Twitter data, because Twitter has a length limit. As shown in Chapter 8, a very large percent of tweets have a length of nearly or exactly 140 characters. Hence, in this phase, these XML character entities were converted into the UTF-8 (Unicode Standard Transformation Format—8-bit) encoding by a command line tool `iconv`.

4.3.4 Retweets

Retweets are a unique feature of Twitter, but they also make it difficult to analyse tweets because of two problems: First, Twitter assigns a new tweet ID and ID string to a retweet, even if it is retweeted by the original author; next, if the retweet contains a URL link, then this link will be converted to a new Twitter's shortened link that is different from the original post (compare Samples 4.3.4 and 4.3.5).

> Sample 4.3.4 R15-06-12 RT@breakingstorm 12-06-2012 03:48:52 Interesting: BreakingNews: RT @breakingstorm: 3,000 forced to evacuate homes in Taiwan because of flas... http://t.co/p6GmMMKd Please RT

> Sample 4.3.5 R15-06-12 RT@breakingstorm 12-06-2012 03:48:52 Interesting: BreakingNews: RT @breakingstorm: 3,000 forced to evacuate homes in Taiwan because of flas... http://t.co/2MDMT58M Please RT

The first problem makes it impossible to count the total number of unique tweets by the tweet ID string or tweet ID; the second brings obstacles to merging the same retweet contents. Ruiz et al. (2012) also noticed this problem that "a single URL can be referred to as several different short URLs" (p. 3), so they found all original URLs to replace the shortened URLs. Because this research does not take the URL-referred contents into account, replacing URLs with a special mark was more efficient. All tweet content was extracted as shown above, and then all URL links were substituted within tweets. To avoid extra problems, the research used a special mark *_THIS_IS_A_URL_LINK_* to label and substitute original URL links. This mark has the same length of twenty characters as the original shorten links, and more importantly, there is no such expression in the content of any of the tweets. Thus, it was safe to replace URL links in this way, and replaced retweets were then regarded as the same tweet (compare Samples 4.3.7 and 4.3.8).

Sample 4.3.7 R15-06-12 RT@breakingstorm 12-06-2012 03:48:52 Interesting: BreakingNews: RT @breakingstorm: 3,000 forced to evacuate homes in Taiwan because of flas... _THIS_IS_A_URL_LINK_ Please RT

Sample 4.3.8 R15-06-12 RT@breakingstorm 12-06-2012 03:48:52 Interesting: BreakingNews: RT @breakingstorm: 3,000 forced to evacuate homes in Taiwan because of flas... _THIS_IS_A_URL_LINK_ Please RT

4.3.5 Repeated tweets

As mentioned in Section 4.2.6, there is a large amount of repetition in the collected tweets due to the data releasing policy of the Twitter Search API. Therefore, to improve the efficiency of the analysis, eliminating repeated contents is a key step. However, this leads to another important question: How should repeated tweets be defined? The following experiments and samples (Samples 4.3.6-4.3.13), collected in May, 2012, illustrate the complexity of this problem.

For **Experiment 1**, the raw data from the crawler number 2,305,173 tweets in total, but simply merging all repeated contents, there are only 424,561 tweets left. In Sample 4.3.9 and 4.3.10, both tweets are the same retweet of a tweet posted by @ABSCBNBreaking, but at different times. These two are repeated content. Thus, 18.42% of the raw contents were extracted.

Sample 4.3.9 M01-05-12 @ABSCBNBreaking 01-05-2012 01:41:20 RT @ABSCBN-Breaking: WEATHER 05/01/2012: PAGASA said the rest of PH will be partly cloudy-cloudy w/ isolated rain showers/thunderstorms, mostly in p.m. (2/2)

Sample 4.3.10 M01-05-12 @ABSCBNBreaking 01-05-2012 04:31:29 RT @ABSCBN-Breaking: WEATHER 05/01/2012: PAGASA said the rest of PH will be partly cloudy-cloudy w/ isolated rain showers/thunderstorms, mostly in p.m. (2/2)

Samples 4.3.11 and 4.3.12 are an extract of **Experiment 2**. They show that the process merged all repeated content, and replaced all URL links with the special mark *_THIS_IS_A_URL_LINK_*. After this, there were 423,404 tweets left.

Sample 4.3.11 M01-05-12 @ABSCBNBreaking 01-05-2012 09:49:10 @ABSCBNBreaking: Hot weather to end mid-May: PAGASA http://t.co/RGrt5lpl2̆01d

Sample 4.3.12 M01-05-12 @ABSCBNBreaking 01-05-2012 09:49:10 @ABSCBNBreaking: Hot weather to end mid-May: PAGASA _THIS_IS_A_URL_LINK_

Experiment 3 removed the post time and replaced URL links (compare Samples 4.3.13 and 4.3.14). After this, 293,078 tweets were left. This large drop is due to that many tweets are retweeted at different times during the days.

Sample 4.3.13 M01-05-12 @ABSCBNBreaking 01-05-2012 09:49:10 @ABSCBNBreaking: Hot weather to end mid-May: PAGASA http://t.co/RGrt5lpl2̆01d

Sample 4.3.14 M01-05-12 @ABSCBNBreaking 01-05-2012 @ABSCBNBreaking: Hot weather to end mid-May: PAGASA _THIS_IS_A_URL_LINK_

Experiment 4 removed the post time and written date of the crawler, and replaced URL links (compare Samples 4.3.15 and 4.3.16). After this, there were 151,476 tweets left. This is because that the Twitter Search API provides content from previous days.

Sample 4.3.15 M01-05-12 @ABSCBNBreaking 01-05-2012 09:49:10 @ABSCBNBreaking: Hot weather to end mid-May: PAGASA http://t.co/RGrt5lpl2̆01d

Sample 4.3.16 M @ABSCBNBreaking 01-05-2012 @ABSCBNBreaking: Hot weather to end mid-May: PAGASA _THIS_IS_A_URL_LINK_

Experiment 5 removed all time-related information, leaving 142,732 tweets left. An extract of such tweets is shown in Samples 4.3.17 and 4.3.18. The amount reduced further, because some users post tweets from previous days, such as retweets.

Sample 4.3.17 M01-05-12 @ABSCBNBreaking 01-05-2012 09:49:10 @ABSCBNBreaking: Hot weather to end mid-May: PAGASA http://t.co/RGrt5lpl2̆01d

Sample 4.3.18 M @ABSCBNBreaking @ABSCBNBreaking: Hot weather to end mid-May: PAGASA _THIS_IS_A_URL_LINK_

Experiment 6 removed the category information of tweets but kept the post date as shown in Samples 4.3.19 and 4.3.20. This left 151,476 tweet. There was no difference in Experiment 4 and 6 in terms of the number of tweets that remained. This shows that there is no cross-posting among different categories of collected tweets.

Sample 4.3.19 M01-05-12 @ABSCBNBreaking 01-05-2012 09:49:10 @ABSCBNBreaking: Hot weather to end mid-May: PAGASA http://t.co/RGrt5lpl2̆01d

Sample 4.3.20 @ABSCBNBreaking 01-05-2012 @ABSCBNBreaking: Hot weather to end mid-May: PAGASA _THIS_IS_A_URL_LINK_

Experiment 7 kept only the post time and tweets (compare Samples 4.3.21 and 4.3.22), leaving 132,004 tweets.

Sample 4.3.21 M01-05-12 @ABSCBNBreaking 01-05-2012 09:49:10 @ABSCBNBreaking: Hot weather to end mid-May: PAGASA http://t.co/RGrt5lpl2̆01d

Sample 4.3.22 01-05-2012 @ABSCBNBreaking: Hot weather to end mid-May: PAGASA _THIS_IS_A_URL_LINK_

Experiment 8 only kept tweets (compare Samples 4.3.13 and 4.3.24). This left the same number of tweets as Experiment 7. This demonstrates that removing the post time does not affect the contents.

Sample 4.3.23 M01-05-12 @ABSCBNBreaking 01-05-2012 09:49:10 @ABSCBNBreaking: Hot weather to end mid-May: PAGASA http://t.co/RGrt5lpl2̆01d

Sample 4.3.24 @ABSCBNBreaking: Hot weather to end mid-May: PAGASA _THIS_IS_A_URL_LINK_

While Experiment 7 and 8 retained the lowest number of tweets, in lacking of time information, they have limited value to the analyses required later. Also, removing the keywords brings a number of difficulties to the research. Thus, to balance the need of minimising the amount of tweets and keeping relevant information for further research, the merging task followed Experiment 4. The cleaned results keep the category, keyword, post time and tweet content as shown in Sample 4.3.13. Table 4.3.2 shows the statistics of the merging results, and makes it apparent that the number of tweets significantly dropped.

Table 4.3.2 Statistics of processed stock tweets

Month	Unique tweets	Raw tweets	Month	Unique tweets	Raw tweets
4/2012	100327	9,938,607	5/2012	151,476	2,305,173
6/2012	145,666	2,165,005	7/2012	134,576	2,186,225
8/2012	109,154	2,054,606	9/2012	104,515	2,028,752
10/2012	87,650	1,532,027	11/2012	111,986	1,952,689
12/2012	101,676	2,072,317	1/2013	122,408	2,127,668
2/2013	111,920	1,913,464			

4.4 Manual Annotation of Stock Tweets

After cleaning tweets, the next phase was to annotate a certain number of tweets manually. To achieve the best accuracy, tweets to be annotated were first selected carefully, the annotation criteria was designed, and the annotation results were double checked. Although the data collection lasted for 11 months from April, 2012 to February, 2013, the manual annotation started in September, 2012, so it only focused on the tweet data from the months prior to starting (April to August, 2012).

4.4.1 Introduction of the tweets from the ticker tweet crawler

As demonstrated in Section 4.3, on each individual day, the three crawlers (the ticker tweet crawler, retweet crawler and media tweet crawler) collected a large portion of repeated content from the Twitter search API. Table 4.4.1 shows the results after the initial data manipulation, from April to August, 2012.

Table 4.4.1 Statistics of the collected tweets from three crawlers, April - August, 2012

Month	**Ticker tweets**	**Media tweets**	**Retweets**
04/2012	29,954	65,580	2,832
05/2012	26,257	84,792	38,700
06/2012	26,475	81,196	36,254
07/2012	29,915	83,384	19,310
08/2012	24,274	83,284	0
Total	136,874	398,236	97,096

As Twitter has rapidly changed their data releasing policy, the Search API stopped releasing any retweets. The retweet crawler stopped working from July 18th, and thereafter did not receive any retweets. Thus, tweets from the retweet crawler were discarded from the research. Moreover, the average number of tweets from the media tweet crawler was much lower than the ticker tweet crawler, so the following data analyses focus only on tweets collected from the ticker tweet crawler.

The ticker crawler started collecting tweets from April 1, 2012, so theoretically, by August 31, 2012, each crawler should have worked for 153 days. However, only 16 ticker crawlers worked for 153 days (hereafter, 153-DAY crawlers), and the rest did not collect tweets as frequently as designed (see Figure 4.4.1). Possible reasons are that these tickers were less likely to be discussed on Twitter, or that Twitter's algorithm caused these unbalanced results.

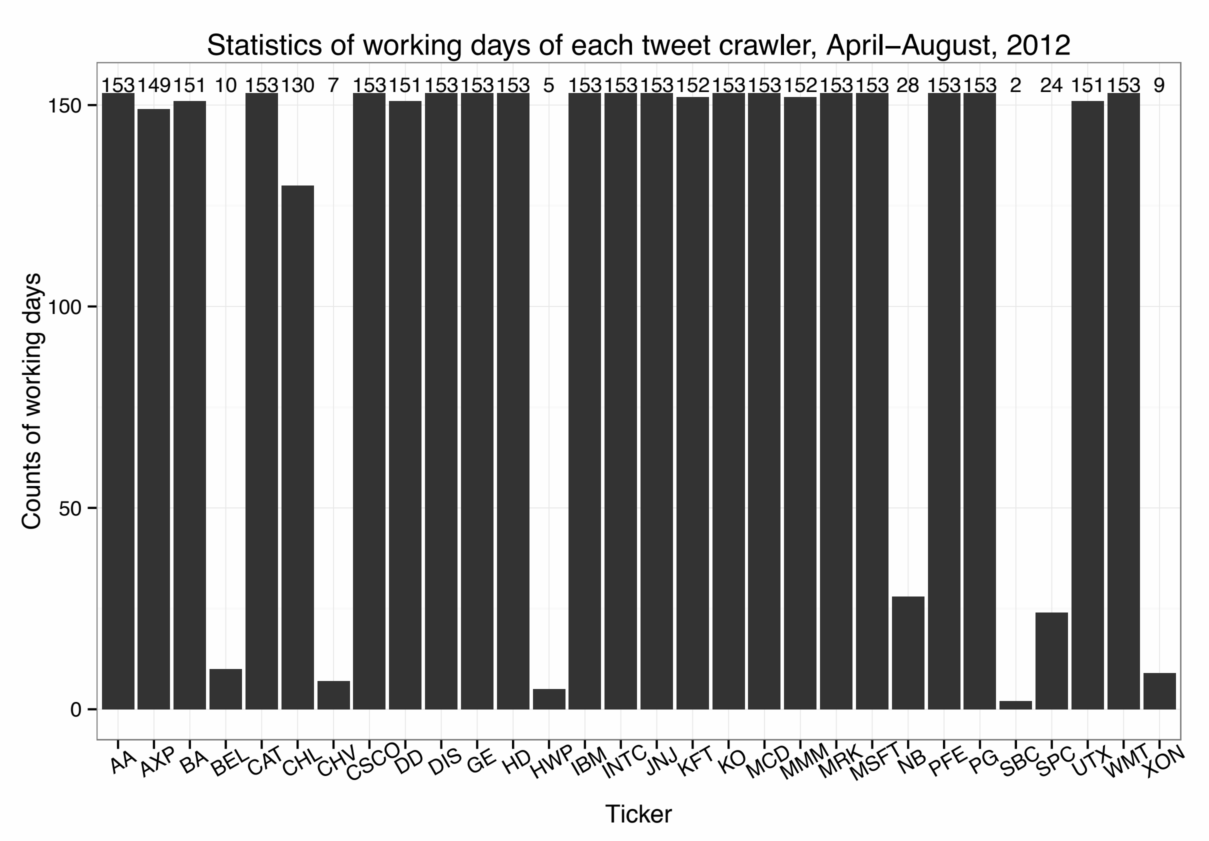


Figure 4.4.1 Statistics of working days of each tweet crawler, April-August, 2012

The average number of working days for all 30 ticker tweet crawlers is 119 days. In these 5 months, these 30 ticker tweet crawlers collected 136,874 tweets, and the sixteen 153-DAY crawlers collected 125,710 tweets, which is about 91.84% of all tweets. This shows how unbalanced the crawling result was (see Figure 4.4.2). The average number of tweets of these sixteen 153-DAY crawlers is 7,856.9 tweets, so the daily average of each ticker is 51 tweets.

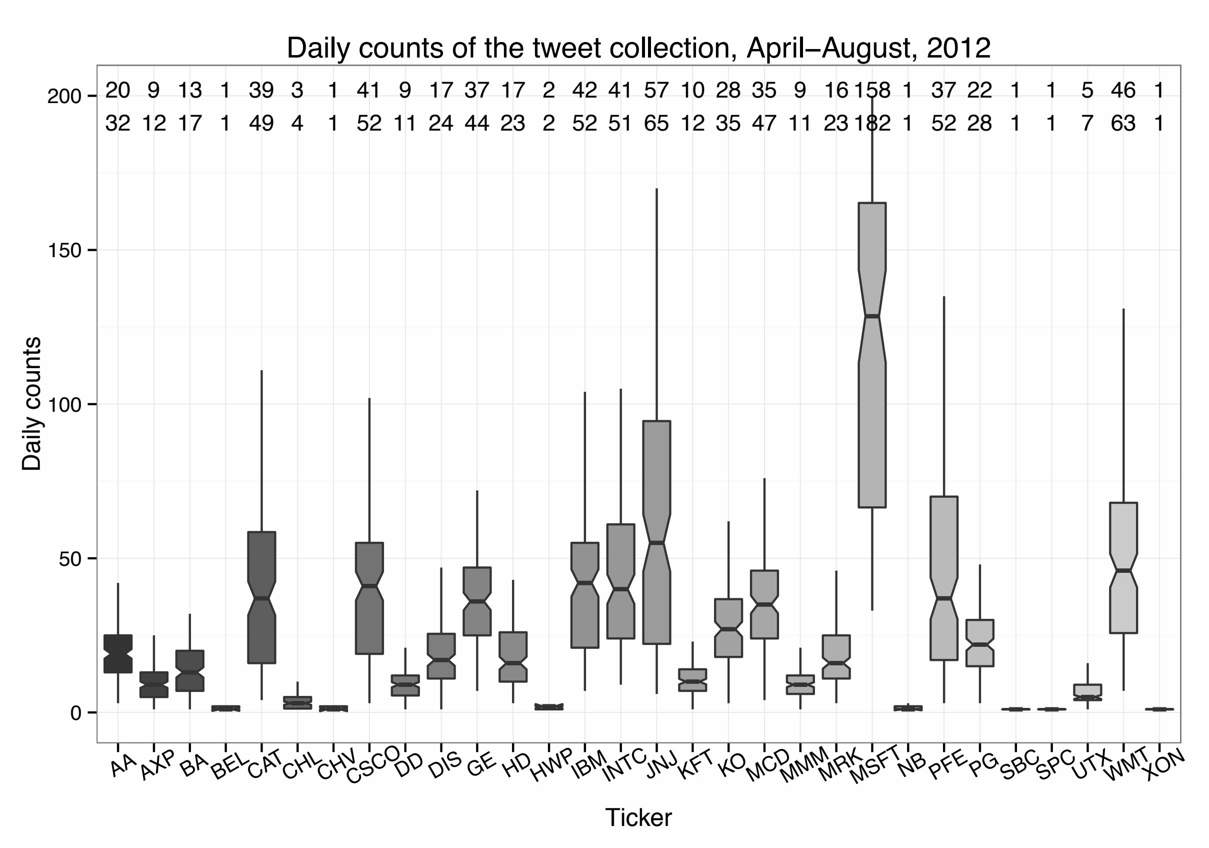


Figure 4.4.2 Daily counts of the tweet collection in logarithm scaling, April-August, 2012

Despite the average, there is a big difference in the means of these sixteen 153-DAY crawler's results (see Figure 4.4.3); thus, the crawling results with a daily average of over 35 tweets were chosen. Only 10 tickers matched this condition. These 10 tickers (hereafter 10-MOST tickers) numbered 100,487 tweets in total, which is about 73.42% of all collected ticker tweets.

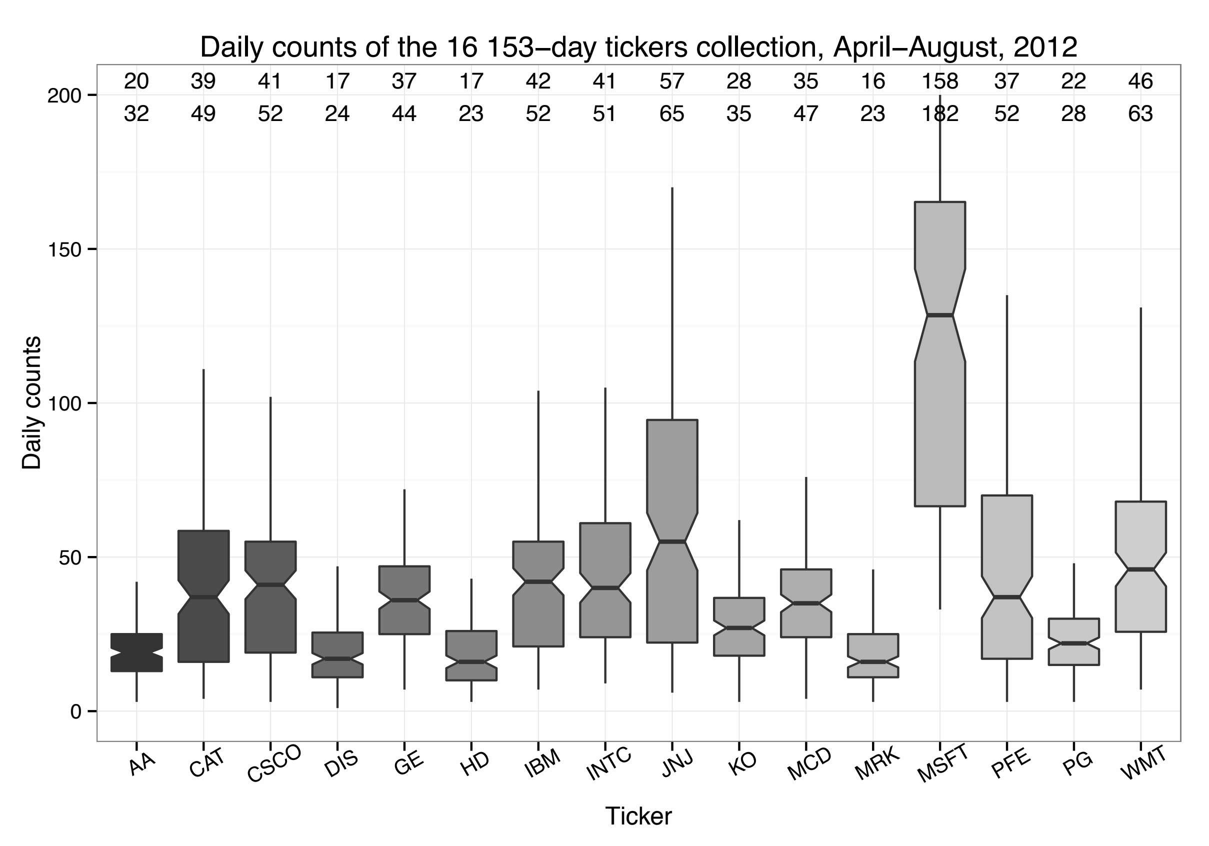


Figure 4.4.3 Daily counts of the 153-DAY tickers in logarithm scaling, April-August, 2012

These 10-MOST tickers had 10,049 tweets in 153 days on average, and the daily average is 65.68 tweets. As shown in Figure 4.4.4, despite the fact that the ticker $MSFT has the most contents, 9 other tickers have similar quantities of data: The total number vary from 7000 to 10,000, the medians range from 35 to 56, and the means range from 40 to 65. These 10-MOST tweets were selected for later analyses.

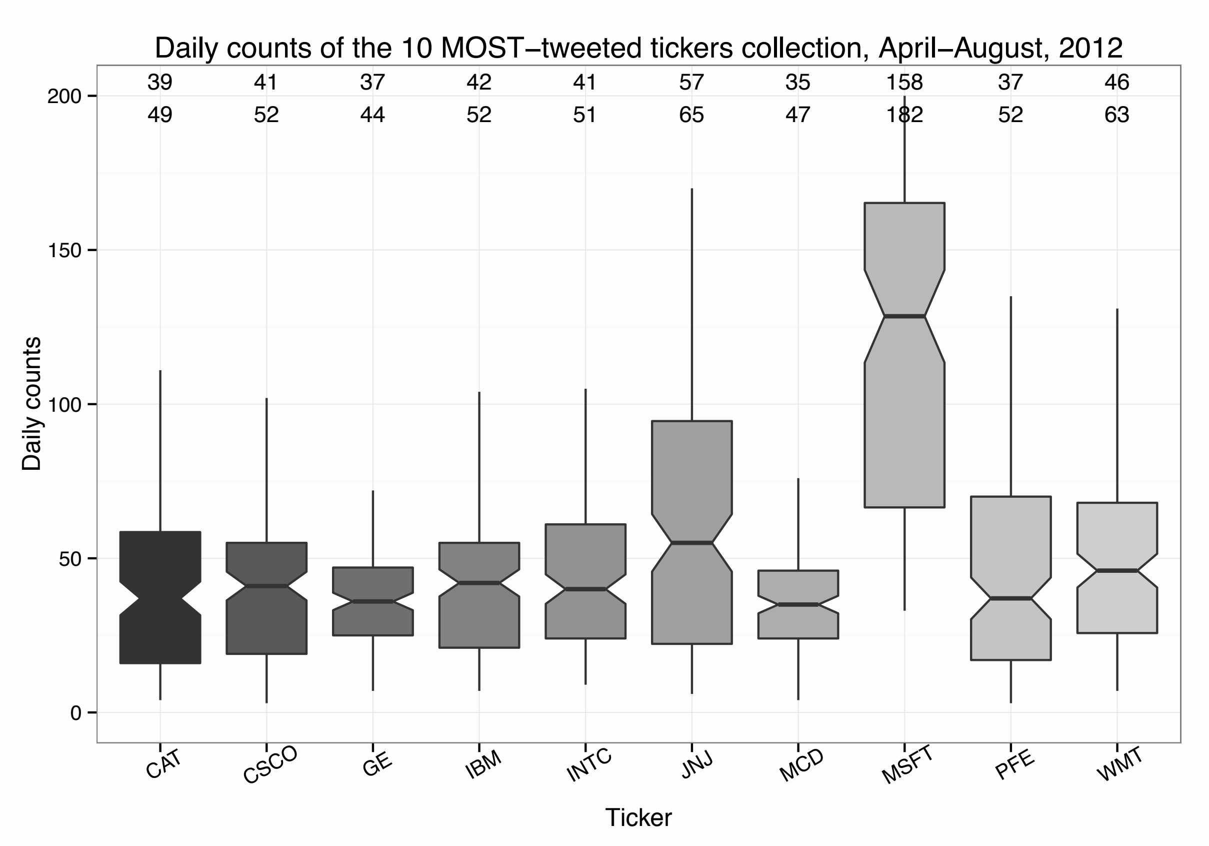


Figure 4.4.4 Box plot of the 10-MOST tickers tweets in logarithm scaling, April-August, 2012

4.4.2 Manual annotation

To improve the understanding of the characteristics of the stock-related tweets, the manual annotation was applied to the collected tweets. This is an important part of the research because all the subsequent analyses relied on the quality of this phase. Thus, it needed attention.

To reduce the possibility of human errors, the manual annotation selected General Electric ($GE) tweets as the sample data. It had the fewest tweets among the 10-MOST tickers, so it required the least amount of effort to annotate, which is beneficial because using the smallest group can reduce the human error in the manual annotation procedure. In total, 6735 GE tweets were collected using the keyword *$GE* over 153 days, and the average daily result is 43.09 tweets. The largest number of tweets collected on any one day was 254 tweets on July 31, while the smallest number of tweets collected was on July 7, where only 7 tweets were collected.

One main shortcoming of manual analysis is incoherence, so another solution to minimise the human error is to recheck the annotation. Repeating the annotation procedure can reduce the number of incoherent annotations and enables the correction of any the annotation errors. Double-checks of the manual annotation were undertaken in order to guarantee its quality. To further minimise the human error, the codes used to represent each category of tweets were designed to be simple (see Section 4.4.3). In taking these steps, (i.e., choosing the fewest tweet group and repeating annotation), the manual annotation kept the human errors to a minimum.

4.4.3 The multi-level classification criteria

Previous studies have rarely discussed sentiment classification criteria (see Chapter 3). Using either polarity or fine-grained classification, they tend to put unclassified content into a neutral category, which acts as "a label for the objective class ('lack of opinion')" (Pang & Lee, 2008, p. 18). The following sections use specific samples from the annotated tweets to critique this classification approach.

Bar-Haim et al. (2011)'s system intended to find the professional traders amongst Twitter users. It classified tweets into two categories, fact and opinion, which each comprised of four subcategories. The fact category contained four subcategories: News, chart pattern, trade and trade outcome, while the opinion category had speculation, chart prediction, recommendation and sentiment.

Following this approach, for the sentiment analysis task, this thesis developed a hierarchy system to categorise stock tweets into different levels, and used different tags to annotate them.

1. Stock level. This first level classifies tweets based on whether they relate to stock topics for the reason that non-stock-related tweets have very limited connection with the price changes of market.
2. Ticker level. This second level classifies tweet based on their relation to the ticker topic because non-ticker-related tweets may only have an indirect connection to the changes.
3. Polarity level. This third level classifies the sentiment of ticker-related tweets because they have a direct connection with the price changes.

For clarity, it is useful to define the concepts of "stock-related tweet", "ticker-related tweet" and "non-stock-related tweet" (see Figure 4.4.5). For this analysis, stock-related tweets indicate tweets discussing the topics related to the stock market. They discuss either the investment in stock or a related topic such as a product produced by the ticker company. In this type of tweets, the cashtag was used as the identification of the ticker name, even if the main topic was not necessarily the ticker company. On the other hand, ticker-related tweets refer to tweets discussing relevant events because the main topic sticks to the ticker: It discusses either news or investment of the ticker. Furthermore, non-stock-related tweets are tweets that use the same cashtag, but the topic has nothing to do with the market. Non-ticker-related tweet are those stock-related tweets with a different main topic to the ticker keyword. The non-ticker-related tweets and ticker-related tweets are included in the stock-related category.

In addition, to classify the sentiment of the ticker-related tweets, the polarity model was adopted to simplify the process. It is clear that stock price change is also a trinary model: It is either bullish, or bearish, or no changes happen, whereas it is labeled as "hold". Thus, using a polarity model in classifying tweets echoes prices changes in the market. A more detailed discussion of this is presented in Section 6.2.6.

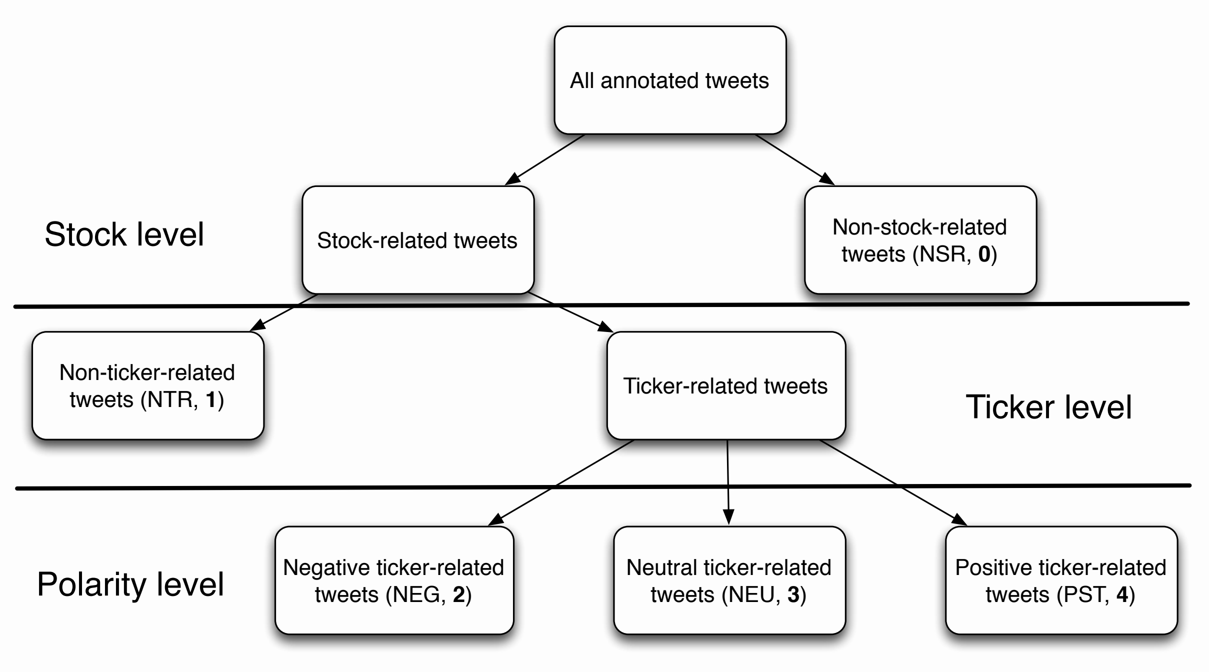

All annotated tweets
Stock level
Stock-related tweets
Non-stock-related tweets (NSR, **0**)
Non-ticker-related tweets (NTR, **1**)
Ticker-related tweets
Ticker level
Polarity level
Negative ticker-related tweets (NEG, **2**)
Neutral ticker-related tweets (NEU, **3**)
Positive ticker-related tweets (PST, **4**)

Figure 4.4.5 The relationship of different types of tweets

The relationship between these types of tweets in this diagram can be summarised as follows:

1. Stock level: All annotated tweets = stock-related tweets + non-stock-related tweets (NSR)
2. Ticker level: Stock-related tweets = ticker-related tweets + non-ticker-related tweets (NTR)
3. Polarity level: Ticker-related tweets = negative tweets (NEG)+ neutral (NEU) + positive (PST)

Also presented in the above figure, using a single digit tag can simplify the annotation, in order to avoid mistaken typing, also it is easier to extract and convert after the annotation. The tags used are outlined below:

1. "0" for non-stock-related tweets;
2. "1" for non-ticker-related tweets;
3. "2" for negative ticker-related tweets;
4. "3" for neutral ticker-related tweets; and
5. "4" for positive ticker-related tweets.

The following discussion explains these different categories in more depth with specific samples. Note that the tweet samples include the keyword and post date at the beginning of the complete tweet, but general users of Twitter would not experience these two types of meta information at the beginning. They are only presented here to show the additional information.

4.4.4 Non-stock-related tweets (NSR)

Although the crawler used the cashtag *$GE* as the search query keyword, a high proportion of the collected tweets are not relevant to the stock topic. They would make a limited contribution to the further analysis, so they were grouped in a separated category. The NSR tweets can be divided by the three main criteria described below:

NSR 1 Tweets borrow the cashtag, but do not discuss any relevant topic related to the concept of the crawling keyword. For example, Sample 4.4.1 discusses a music group, so it is not relevant to General Electric; however, this type of tweets often combines the hash mark with *$GE,* so the crawler still collected them.

> Sample 4.4.1 $GE 2012-05-02 I thought just my niggas go ham for team but #$GE girls bout that Life #$GE

NSR 2 Tweets use the cashtag according to its original purpose, but the main topic of these tweets is not about the stock market, the ticker company, or any related concepts. For instance, with the cashtag *$GE*, Twitter uses *$GE* as a sample to introduce the new function of cashtags, and many tweets discuss this topic (see Sample 4.4.2). In these tweets, the main topic is Twitter's new function and not General Electric or any of its related concepts. These tweets have very limited contribution to the movements of GE's price.

> Sample 4.4.2 $GE 2012-08-01 Twitter is rolling out a new feature that makes tweeted stock symbols clickable and searchable on _THIS_IS_A_URL_LINK_ Try it - $FB $GE

NSR 3 Tweets contain no text, other than a single cashtag (see Sample 4.4.3). It is hard to identify whether they were stock-related tweets or not. Also they contain no sentiment information, so they were filtered out.

> Sample 4.4.3 $GE 2012-08-01 $GE

4.4.5 Non-ticker-related tweets (NTR)

The NTR tweets belong to stock-related tweets, but the main topic is not closely relevant to the ticker's concept. They do not have a focus on any specific ticker, so they were excluded from further analyses. This type of tweets can be divided into the five types outlined below:

NTR 1 Tweets advertise stock-related products such as investment reports. Some of them add a string of cashtags at the end to attract more attention (see Sample 4.4.4). However, they have limited information because they do not focus on any specific ticker. The URL links connect to external content, which may have explicit sentiment, but because the tweet itself does not have any explicit sentiment, any sentiment is unlikely to be recognised.

> Sample 4.4.4 $GE 2012-08-01 $AA, $BAC, $BK, $CSX, $EMC, $F, $GE, $GLW, $HPQ, $KR, $LNC, $MS, $OI, $TER, $VLO: Watchlist Aug 2, 2012 _THIS_IS_A_URL_LINK_

NTR 2 Tweets discuss a similar topic, but the foci are not the keyword. This type of tweet discusses other tickers but lists the cashtags at the end. The main focus can be either a competitor, or a partner of the main ticker. However, sometimes, the relationship between them is not clear. In most cases, adding the cashtag of the main ticker is only to catch more attention, so they contribute in a limited way to the main ticker's changes. Generally, this type of tweet is a news tweet: They present the title of the news and include a URL link and cashtags at the end. The first two samples below (Sample 4.4.5 and 4.4.6) show information about General Electric's partners or competitors, and the last two (Sample 4.4.7 and 4.4.8) show the general market information; however, none of them explicitly discusses General Electric.

> Sample 4.4.5 $GE 2012-08-02 First Solar: Q2 Results Show Its Financial Strength And Market Potential Are Unmatched _THIS_IS_A_URL_LINK_ $GE $LDK $SPWR $SRE

> Sample 4.4.6 $GE 2012-08-30 NEWS FLASH: Tim Cook is a pretty good CEO. Apparently Steve Jobs didn't read a single word of Jack Welch. Thank goodness. $GE $AAPL

> Sample 4.4.7 $GE 2012-05-04 $GE News: Most active New York Stock Exchange-traded stocks _THIS_IS_A_URL_LINK_ #Active #Exchangetraded #Most

> Sample 4.4.8 $GE 2012-04-10 $GE News: U.S. Stocks Decline as Employment Report Misses Estimates _THIS_IS_A_URL_LINK_ #daytrading #Decline #Employment #Estimates

NTR 3 Tweets do not contain a precisely identifiable cashtag. Although they were collected by a crawler indexing a cashtag as the keyword, the tweet only contains part of the cashtag. As shown in

Sample 4.4.9, the cashtag "$GE_F" includes "$GE" but has no relationship with General Electric. Thus, they do not contribute to the main ticker's movements.

> Sample 4.4.9 $GE 2012-08-01 @HansaTrading no idea, that's all programmed in. $GE_F traders used to put up 1000 lots to get 10. Depends on a lot of factors

NTR 4 Tweets contain the cashtag, but are automatically generated by a website. Nowadays, more and more websites have integrated a convenient sharing feature as discussed in Chapter 2, so the tweets generated by these services usually only contain the title and the URL link of the post. This type of tweets also includes some cashtags, but they may not have a specific focus on this ticker (see Sample 4.4.10). Thus, they also have a limited association with the main ticker.

> Sample 4.4.10 $GE 2012-08-02 Commented on: "2 Short-Term Bearish Option Ideas On General Electric And Facebook" _THIS_IS_A_URL_LINK_ $FB $GE

NTR 5 Retweets of the previously listed four types of non-ticker-related tweets. Many users reposted other's tweets, especially news tweets. They may add comments or just simply retweet (see Sample 4.4.11). If the retweet does not include any relevant comments to the main topic, or if it is just a repost, it was classified as an NTR tweet, too.

> Sample 4.4.11 $GE 2012-08-29 RT @PaulLoete: Miracle Improvement In ALS Patient Could Force Big Pharma To Get Serious On Stem Cell Therapy _THIS_IS_A_URL_LINK_ $ATHX $GE $GSK

In previous studies as discussed in Chapter 3, tweets matching the five criteria above are classified as neutral tweets, because they have neither explicitly negative nor explicitly positive sentiment on the main topic. However, it is important to notice that these tweets do not have any explicit association with the keyword as ticker-related tweets. Furthermore, if they are classified as neutral tweets, the different degrees of influence would be neglected. Although a number of studies consider “neutral” as neither negative nor positive, this does not mean that neutral is equivalent to “no association”; in other words, unclassified tweets have no association to the stock changes, so they cannot be considered as neutral tweets. As discussed above, these types of tweets may discuss about a competitor, a partner or the market, meaning they can have some indirect associations with the main topic. Thus, these tweets were grouped for the further analysis, instead of being classified into the neutral category.

4.4.6 Ticker-related tweets

The following sections discuss the ticker-related tweets. As explained above, this type of tweet has a specific focus on the cashtag keyword, and they were categorised into positive, negative and neutral groups as discussed. It seems simple to echo negative, neutral and positive sentiments with the market trends as bearish (downward-moving), hold (steady) and bullish (upward moving) respectively, but the reality is not: In addition to the relationships of the market trend, there is another binary relationship – the investment action as buying, holding and selling (see Table 4.4.2).

Table 4.4.2 The relationship of the positive and negative relationship between the market trend and the investment action

Relationship	**Negative**	**Neutral**	**Positive**
Market trend	bearish	hold	bullish
Investment action	selling	holding	buying

Sprenger and Welpe (2010) pointed out that “buy and sell signals may carry very different information with respect to subsequent stock returns" (p. 10), and this indicates that the relationship of investment action does not well match that of the market trends. To explain the disagreements of these two relationships, it is better to take some concrete examples from the annotated tweets. Sample 4.4.12 first expresses a positive thought about the GE ticker, so it indicates that the market trend is bullish. However, it soon comments that it did not reach the author's expectation, so the user's reaction was to hold. This is an obvious difference between the market trend and the investor's evaluation, so this tweet was classified as an NEU tweet.

> Sample 4.4.12 $GE 2012-04-16 $GE Doesn't look bad, but it isn't good enough for me. I"ll check on it tomorrow. Maybe one more down day for $GE

In the stock market, the term *pullback* indicates a situation where the ticker's price drops after it reaches a peak, and investors may consider this as a long-term positive opportunity. In Sample 4.4.13 the author sees a pullback of the ticker, which is a bearish signal, but the author seems to regard it as a very good chance, using the word *great* in uppercase letters to express the intensity of the evaluation. Thus, the sentiment of the author is positive – this is opposite to the market trend. Considering the author's mood, it classified as a PST tweet.

> Sample 4.4.13 $GE 2012-05-09 $GE - With todays 1.30% pullback, GREAT buying opportunity for all long term investors. Also with future div increases - it's a bargain.

The disagreement between the relationship of the market trend and the investment action increases the difficulty of designing a classification of the ticker-related tweets and as well as the difficulty of analysing the correlation between the market and sentiment. To avoid ambiguities, this research defines sentiment as following the relationship of the investment action: Positive sentiment as the investor's willingness to buy the ticker, negative sentiment as the investor's willingness to sell, and neutral sentiment as willingness to hold. This is to simplify the annotation and analysis undertaken later. An in-depth discussion of each category is presented in the following sections.

4.4.7 Negative tweets (NEG)

The NEG group reflects the negative sentiment about the keyword ticker.

NEG 1 Tweets express negative thoughts about the ticker (see Sample 4.4.14 and 4.4.15). This type of tweets often contains the word *bearish*, *bearishly* or similarly negative expressions. With such an explicit expression, their sentiment is self-evident.

> Sample 4.4.14 $GE 2012-08-02 $GE - Rolling over. $19.40 bearish target. _THIS_IS_A_URL_LINK_

> Sample 4.4.15 $GE 2012-05-06 $GE - General Electric Stock Analysis - RSI is bearish and falling - _THIS_IS_A_URL_LINK_

NEG 2 Tweets express the idea of "selling the ticker". Although in Sample 4.4.16, there is no explicitly negative sentiment associated with the main ticker, it demonstrates that the user shortened (sold) the GE stock. Though selling is not a complete expression of negative sentiment and neither is buying entirely an expression of positive sentiment, the convention often identifies selling as a negative signal, and buying as a positive one. Thus, this research followed the convention to extend the concept of NEG and PST tweets.

> Sample 4.4.16 $GE 2012-08-03 Shorted $GE with a three cents stop. Looking for it to make a lower high.

NEG 3 Tweets discuss negative news of the ticker (see Sample 4.4.17). Bad news often brings bad changes (Naveed, Gottron, Kunegis, & Alhadi, 2011), so this type of tweets could have a direct influence on the price movements.

Sample 4.4.17 $GE 2012-08-03 General Electric Is Overvalued _THIS_IS_A_URL_LINK_ $GE

NEG 4 Tweets express disappointments or other negative emotions about events of relevance of the ticker. In Sample 4.4.18, NBC (National Broadcast Cable) has been acquired by General Electric before, and they cancelled the London Olympics' live broadcast, which caused widespread disappointment. Hence, the price performance of General Electric might be affected.

Sample 4.4.18 $GE 2012-08-01 NBC #fail for replacing Mad Money with @jimcramer with #Olympics boxing... Why didn't they make a special boxing channel? #tv $GE $CMCSA

4.4.8 Neutral tweets (NEU)

The NEU group contains tweets having either neutral or uncertain sentiment on the relevant topics of the ticker cashtag.

NEU 1 Tweets probably have a positive or negative tendency, but they ask a question, making them more ambiguous. Samples 4.4.19 and 4.4.20 are concrete examples: The first tweet is an open question that may have different answers and the second is a typical question about revenue.

Sample 4.4.19 $GE 2012-08-29 $ge will it ever trade above 21 again?

Sample 4.4.20 $GE 2012-08-01 @jimcramer what % of $GE's earnings come from India?

NEU 2 Tweets express the idea of "holding" the ticker. For example, Sample 4.4.21 indicates that the author is holding the GE ticker, without any apparent wish to buy or sell:

Sample 4.4.21 $GE 2012-05-30 Got long $ASPS today and $GE long still holding up

NEU 3 Tweets report the ticker news, but contain no explicit sentiment. Notice, this differs from Criteria 2 of the NTR tweets, because this type of tweets exclusively focuses on the GE ticker. Sample 4.4.22 demonstrates this. It is a news tweet about GE, without any other topic. However, it is also notable that this incomplete tweet contains limited sentiment as it only shows the title of a news report.

Sample 4.4.22 $GE 2012-08-03 CNN Money: How a top $GE exec engineered himself out of a job - _THIS_IS_A_URL_LINK_ #finance

NEU 4 Tweets advertise the ticker, focus on the ticker exclusively, but do not contain any obvious sentiments. This type of tweets differs from NTR 1, because it has a specific focus on the ticker. Though they are advertising tweets, they show that the ticker is well discussed within this community. For example, Sample 4.4.23 is an advertising tweet that only reports GE ticker.

Sample 4.3.23 $GE 2012-05-05 Check General Electric Co. ($GE) stock technical analysis based on closing price: _THIS_IS_A_URL_LINK_ .

NEU 5 Tweets contain conflicting opinions in an single tweet: The following examples below show that the users have different views of the performance, so it is difficult to summarise the sentiment. In Sample 4.4.24, the author first praises the GE ticker, then claimed that he is not satisfied; in Sample 4.2.25, the user is not completely satisfied with the performance of GE; and in Sample 4.2.26, although GE performs well, it has not reached the author's expectation.

Sample 4.4.24 $GE 2012-04-16 $GE Doesn't look bad, but it isn't good enough for me. I"ll check on it tomorrow. Maybe one more down day for $GE

> Sample 4.4.25 $GE 2012-05-16 $GE looks attractive with recent acquisition news. However, a conglomerate doesn't seem to have much growth potential. #stockaction

> Sample 4.4.26 $GE 2012-07-20 Multi-speed world: @GeneralElectric industrial #sales down 7% in Europe, up 6% in U.S., climb 24% in China: _THIS_IS_A_URL_LINK_ $GE

4.4.9 Positive tweets (PST)

The PST group includes tweets expressing positive thoughts about the ticker.

POS 1 Tweets repeat positive news related to the ticker, including new investments, new product launches, or a new leader's good performance. For instance, Sample 4.4.27 reports the launch of a new product and new technology, and Sample 4.4.28 is about an interview with GE's CEO.

> Sample 4.4.27 $GE 2012-06-19 RT @FinancialTimes: $GE launches new carbon capture technology _THIS_IS_A_URL_LINK_

> Sample 4.4.28 $GE 2012-04-19 CEO of General Electric on Sparking an American Manufacturing Renewal - Lean is in- Harvard Business Review: _THIS_IS_A_URL_LINK_ #in $GE

POS 2 Tweets report positive news of the ticker's movements as shown in Sample 4.4.29:

> Sample 4.4.29 $GE 2012-08-22 $ge clinging to major support at 20.80ish.Needs to get quick pop away from it off open or heading to 20.56.

POS 3 Tweets mention an investor's interest in or confidence about the ticker. For example, Sample 4.4.30 is a comment of one GE ticker buyer, in which the buyer expresses the expectation of obtaining the GE ticker.

> Sample 4.4.30 $GE 2012-08-21 I sold my shares of $PCS today. I"m thinking about throwing the money into $GE. I've been itching to buy some more shares of it for a while.

POS 4 Tweets express a positive sentiment explicitly, even when the market displays an opposite motion. Sample 4.4.31 provides a good illustration: The price of GE dropped 1.3%, but the tweet reports it as a buying opportunity, which is a positive sentiment.

> Sample 4.4.31 $GE 2012-05-09 $GE - With todays 1.30% pullback, GREAT buying opportunity for all long term investors. Also with future div increases - it's a bargain.

4.4.10 Further discussion on classification criteria

The above examples demonstrated the most typical samples found for each category, but in the annotating process, some particular cases appeared that are worth discussing.

Using the retweet function of Twitter has two common formats as discussed in Chapter 2. Sometimes, a tweet omits or curtails the sentiment information contained in the original tweet. Sample 4.4.32 shows that the original tweet includes a clearly positive sentiment, while Sample 4.4.33, a quoted tweet contains no explicit sentiment, so it was classified as a NTR tweet.

> Sample 4.4.32 $GE 2012-08-16 Retirement Strategy: General Electric Just Might Be The 2013 Stock Of The Year (Part 32) _THIS_IS_A_URL_LINK_ $GE

> Sample 4.4.33 $GE 2012-08-16 Commented on: "Retirement Strategy: General Electric Just Might Be The 2013 Stock ... _THIS_IS_A_URL_LINK_ $GE

Another typical phenomenon of retweets is that they share similar content. Using some Twitter clients might automatically add or cut some content of the original tweet, and some users prefer adding or cutting some contents when they repost. The following 7 tweets (Samples 4.4.34 - 4.4.39) contain the exactly same topic, but the contents are slightly different. The first three (Samples 4.4.34 - Sample 4.4.36) could be tweets posted by different users, so they add different cashtags at the end, or they could be posted from clients that automatically cut the end of original tweets. Although the last three tweets (Samples 4.4.37 - 4.4.39) are quoted retweets, they do not contain any comments. This type of repeated content is difficult to detect and distinguish, so they were regarded as different tweets.

Sample 4.4.34 $GE 2012-08-27 Miracle Improvement In ALS Patient Could Force Big Pharma To Get Serious On Stem Cell Therapy _THIS_IS_A_URL_LINK_ $ATHX $GE $GSK

Sample 4.4.35 $GE 2012-08-27 Miracle Improvement In ALS Patient Could Force Big Pharma To Get Serious On Stem Cell Therapy _THIS_IS_A_URL_LINK_ $ATHX $GE $GSK #ALS

Sample 4.4.36 $GE 2012-08-27 Miracle Improvement In ALS Patient Could Force Big Pharma To Get Serious On Stem Cell Therapy _THIS_IS_A_URL_LINK_ $ATHX $GE $GSK #stemcells

Sample 4.4.37 $GE 2012-08-27 RT @ALSChicago: Miracle Improvement In ALS Patient Could Force Big Pharma To Get Serious On Stem Cell Therapy _THIS_IS_A_URL_LINK_ $ATHX $GE $GSK #ALS

Sample 4.4.38 $GE 2012-08-27 RT @ALSChicago: Miracle Improvement In ALS Patient Could Force Big Pharma To Get Serious On Stem Cell Therapy _THIS_IS_A_URL_LINK_ $ATHX ...

Sample 4.4.39 $GE 2012-08-27 RT @nicoforraz: Miracle Improvement In ALS Patient Could Force Big Pharma To Get Serious On Stem Cell Therapy _THIS_IS_A_URL_LINK_ $ATHX $GE $GSK #stemcells

Some tweets are very similar to a report digest, because they contain information on multi-topics in one tweet. This brings difficulties in detecting the main topic. For example, Sample 4.4.40 introduces three topics in one tweet: the IPO of Burger King, the fine paid by JP Morgan, and Moody's changes of evaluation of General Electric. These three topics are independent of each other, and only the last relates to the topic.

Sample 4.4.40 $GE 2012-04-04 @ReutersInsider: Get a bite of #BurgerKing's IPO, $JPM to pay $20 mln fine & Moody's downgrades $GE.: _THIS_IS_A_URL_LINK_

4.4.11 Summary

This section discussed the classification criteria used in the manual annotation of GE ticker tweets. It used a three-level hierarchy system to classify tweets in order to differentiate the degree of influence to the market movements of tweets. Particularly, it put forward a critique of the popular approach that classifies unclassified tweets into NEU tweets with the argument that they may have some indirect influence on the market changes. Also, it explained the disagreement of the market trend relationship and the investment reaction relationship. This section provided detailed criteria of each of the categories of tweets, with an aim to disambiguate them as far as possible.

4.5 Relationship of Annotated Tweets

After the manual annotation, as shown in Figure 4.5.1, making up more than 30% of the tweets, the PST group dominates the collection of 6,735 GE tweets. The next most frequent are the NTR (27.97%) and NSR tweets (22.88%); which together account for 50.85% of all the GE tweets, so they account for a larger proportion of the data than the ticker-relevant content. On the other hand, the NEG and NEU tweets only share 8.23% and 10.42% of the entire dataset respectively. Specifically, in the ticker-related group, the PST group occupies 62.05% of the 3,310 ticker-related tweets, where the NEG and NEU group share 16.73% and 21.21% respectively. This roughly matches the overall positive trend of the GE share price during this period as shown in Chapter 5.

It is noticeable that the ratio between the positive tweets and the negative tweets is nearly 4:1. This is more balanced than Dewally's result of 7:1 (2003) and Antweiler and Frank's 5:1 (2004). It is similar to Sprenger and Welpe's 3:1 (2010), but less balanced than Rao and Srivastava's 3:2 (2012).

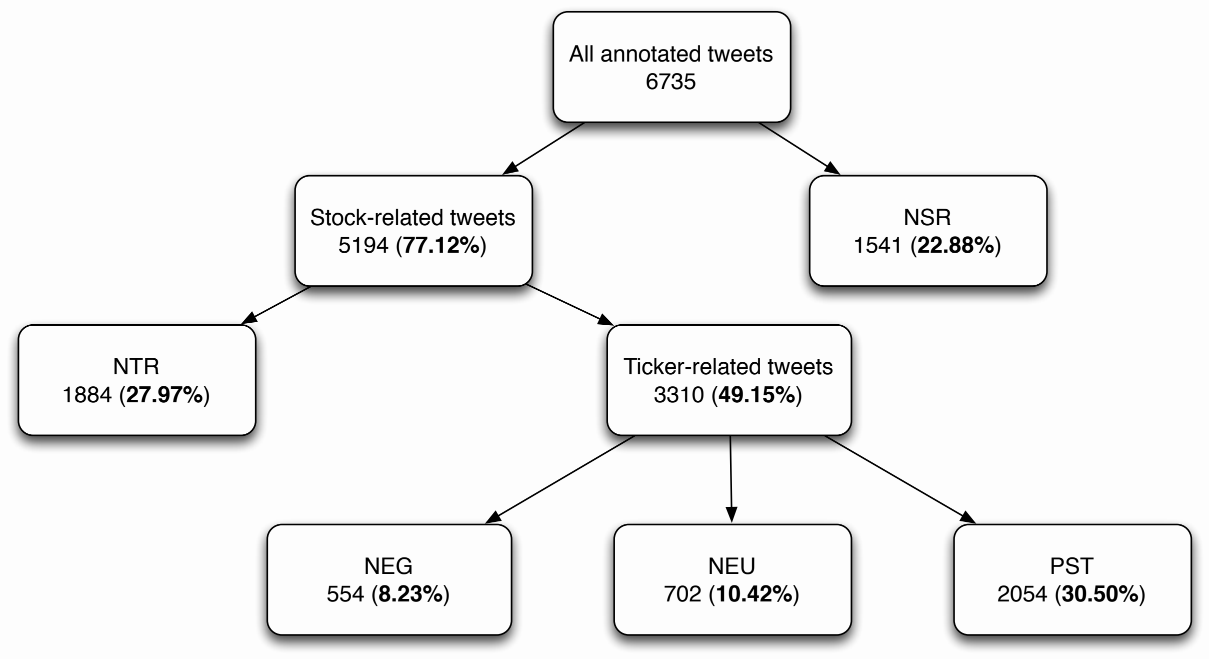


Figure 4.5.1 Distribution of different types of GE tweets after the manual annotation

4.6 Summary

This chapter introduced the procedure of collecting tweets discussing stock from the Twitter APIs. It presented the details of how to collect and manipulate data from the Twitter APIs: After reorganising the directory, converting JSON to TSV files, cleaning HTML encoding, and replacing URL links, the data are ready for further analysis. The next part discussed details of the manual annotation of GE ticker tweets. It designed a three level hierarchy classification category, namely the stock level, ticker level and polarity level. During the manual annotation, each tweet was grouped according to specific criteria into one of five categories of tweets, including the non-stock-related tweets, non-ticker-related tweets, negative ticker-related tweets, neutral ticker-related tweets, and positive ticker-related tweets. With this hierarchical classification system, the irrelevant contents in the tweet data was efficiently reduced . The following chapters use these processed annotated tweet data to explore their temporal and linguistic features.

Chapter 5 Temporal Relationship of Stock Price and Tweet Sentiment

The temporal feature of stock tweets is particularly important in the stock prediction context. If the temporal correlation between tweet sentiment and the stock market is weak, or non-existent, then the possibility of using tweets to predict the market will be bleak. Therefore, this chapter focuses on exploring the temporal correlation between the annotated GE ticker-related tweets and GE stock price. The chapter first briefly introduces the stock price data, and gives an overview of time series analysis in detail. Then, it reports on analyses of the temporal correlation of the raw data, including the correlation between stock price, tweet sentiment, and stock price and tweet sentiment. Next, the chapter reports on analyses of the temporal correlations between the converted data, which used different time series data manipulation techniques. Finally, it presents tests of the differences in the temporal correlations based on different definitions of neutral sentiment.

5.1 Stock Price Data Collection

The following analyses used the Dow Jones Industrial Average (DJIA) prices from the New York Stock Exchange (NYSE), which provides four different prices for each ticker: the open, high, low and close price. In the stock market, these four prices indicate different statuses of each stock on a given day. The high and low price are the highest and lowest price of a given stock within an individual day, and the open and close prices are the prices of a given stock at the opening and closing time of the market that day.

These data can be fetched from Google Finance (`http://www.google.co.uk/finance`). It offers the above four types of price data as well as the daily volume for each ticker in the Historical

Prices page. For example, the raw data of GE's prices can be found at (`http://www.google.co.uk/finance/historical?q=NYSE%3AGE`). These data can be directly downloaded as a CSV (comma-separated value) file, or be extracted by any relevant software, such as R.

As shown in Table 5.1.1, relevant price data from Google Finance were retrieved and converted them into TSV (tab-separated value) files. The ticker price data contain five columns, and the first row indicates the names of each column. Therefore, each row represents the price data of an individual day with same four intervals, and each price figure is the price generated by the market. In the sample, the last row does not contain any numeric data for the four prices as outlined above, because April 6, 2012 was a Saturday, and no stock market opens on weekends or holidays. In such instances, the mark *NA* is used to indicate missing data, because the following analyses relied on `R`, which regards *NA* as a logical constant of "Not available/missing data" (Venables & Ripley, 2002, p. 35). In total, there were 45 holidays (including weekends) during the period from April 1 to August 31, 2012, so 180 *NA* marks are assigned. Therefore, 29% of the data for this period do not include any price information. The difficulties brought by this problem were discussed in Section 4.1.

Table 5.1.1 Sample of converted GE price data

Date	**Open**	**High**	**Low**	**Close**
2012-04-01	20.07	20.13	19.95	20.07
2012-04-02	20.03	20.11	19.90	20.02
2012-04-03	19.97	20.02	19.80	19.96
2012-04-04	19.65	19.81	19.62	19.74
2012-04-05	19.61	19.71	19.46	19.49
2012-04-06	NA	NA	NA	NA

5.2 Introduction of Time Series Analysis

Time series analysis is a statistical analysis method designed for understanding temporal relationships in data, and this analysis used time series analysis to investigate the relationship between changes in stock price and tweet sentiment changes. According to Chatfield (1980), time series analysis mainly tests two types of relationship in the data: univariate processes and bivariate processes. In univariate analysis, forecasts can be based entirely on past observations in a given time series, by "fitting a model to the data and extrapolating" (p. 82). In multivariate analysis, forecasts can be made by "taking observations on other variables into account" (p. 83).

In brief, the first approach can be considered as a historical relationship at different time points within one group of time series data and the second one as a correlation between different groups of data within a certain time range. In this research, interpreting the correlation between stock price and tweet sentiment at the corresponding time is more important, so it focuses on the bivariate processes. This section first discusses what features of time series data have, and then introduces the main time series analysis models and methods used in financial data.

5.2.1 Time series data

As stated in Shumway and Stoffer (2006), time series data are "a collection of random variables indexed according to the order they are obtained in time" (p. 11). This is supported by Shasha (2000), who summarised time series data as "a sequence of values usually recorded at regular increasing intervals". The importance of the regularity is also emphasised. According to these, the time series data should meet four characteristics:

1. It is a group of data.
2. The data are random.

3. The data are stored according to specific time stamps.
4. The data are stored regularly.

Financial data, particularly stock return data are carefully indexed according to the market opening time. As introduced above, Table 5.1.1 shows the price data of the GE ticker for given days, and it matches these four characteristics. As such, time series analysis can be applied to the stock price data.

Moreover, the tweet sentiment data incorporate similar features. Table 5.2.1 shows the counts of each sentiment categories of annotated GE ticker tweets collected as described in Chapter 4. As with the stock price data, the first column illustrates the date of the data, and the remaining columns represent the counts of different categories of tweet. Each row contains an individual day's data with the same intervals, which were generated based on the annotated results in Chapter 4. Therefore, this group of data can also be regarded as a set of time series data.

Table 5.2.1 Sample data of GE sentiment

Date	NSR	NTR	NEG	NEU	PST
2012-03-28	1	11	2	2	12
2012-03-29	3	7	3	5	15
2012-03-30	0	12	0	1	9
2012-03-31	0	6	1	3	7

Based on the fact that they meet the four characteristics outlined above and are similar in format, it is reasonable to apply time series analysis on these two groups of data. The primary goal is to test the bivariate time series relationship between these two groups of data.

5.2.2 Test of correlation between different time series data

As discussed above, time series analysis can test two types of temporal processes. To test the correlation of univariate processes, the auto-covariance and the autocorrelation function (ACF) are widely applied. The auto-covariance function is used to measure "the *linear* dependence between two points on the same series observed at different times" (Shumway & Stoffer, 2006, p. 20), while the ACF is for "the linear predictability of the series at time t, say, X_t, using only the value X_s" (p. 24). In other words, autocorrelation is used to test the correlation of the changes at two different time points in the same data, where the ACF can be considered as a normalised auto-covariance function. In addition, the difference between time t and time s is called lag in time series analysis literature.

Furthermore, to test the correlation of bivariate processes, the cross-covariance function and the cross-correlation function (CCF) have been developed. Chatfield (1980) defined the cross-covariance function as follows:

> For a bivariate process, the moments up to second order consist of the mean and auto-variance function for each of the two components plus a new function. (p. 170)

To be precise, the cross-covariance function is used to measure "the predictability of another series Y_t from series X_s" (Shumway & Stoffer, 2006, p. 22), where cross-correlation function is the normalised version of the cross covariance, and both of them can be scaled up to "multivariate time series with r components" (p. 23).

In brief, the auto-covariance and the autocorrelation functions measure the temporal correlation within one group of time series data, while the cross-covariance and the cross-correlation functions measure the temporal correlation between two or more groups of time series data. More importantly, the cross-correlation function is widely used in the analysis of the correlation between stock activities and tweets, for example, Ruiz et al. (2012) used the CCF test to estimate the correlation between stock price, trading column, and tweeting behaviour.

5.2.3 Introduction of the ACF and the CCF functions in R

As a statisical computing language, R provides various tools to conduct time series analysis. Either the built-in functions or third-party packages offer this possibility. The built-in functions of `acf` and `ccf` in R were used in this research. A brief explanation of these functions is presented here. According to the R Team (2013), `acf` and `ccf` in the **stats** package are implanted from `S-plus` language. `S-plus` is the origin of R language, and R is similar to it. The built-in `acf` and `ccf` functions are slightly different from their mathematical definitions. According to the R Team (2013), the R documentation defines that the function `acf` "computes (and by default plots) estimates of the auto-covariance or autocorrelation function (p. 1014)". On the other hand, the function `ccf` "computes the cross-correlation or cross-covariance of two univariate series" (p. 1014). According to these, the function `ccf` can only deal with two groups of univariate time series data at once, which is slightly different from its mathematical definition that can compute two or more groups of time series data. However, the `acf` function is more robust in that it can be applied to more than two groups of univariate data at once:

> Our definitions are easily extended to several time series observed over the same interval. (Venables & Ripley, 2002, p. 390)

To better understand the relationship of the `acf` and `ccf` functions in R, the following example is taken from the GE price data used in Section 6.4. The result of using the `ccf` function to test the correlation between the open and close price of the GE ticker is shown in Table 5.2.2:

Table 5.2.2 Autocorrelations of series 'X', by lag

Lag	-4	-3	-2	-1	0	1	2	3	4
Correlation value	-.018	.022	.210	-.350	-.391	.956	-.151	-.520	.243

Figure 5.2.1 shows the cross-correlation result computed by the `acf` function in R. As indicated in the two circled frames, the correlation values of the close price in the matrix of open price shows the same results as the correlation values of the negative lags in Table 5.2.2, while the correlation values of the open price in the facet of the close price shows the same results as the correlation values of the positive lags. Combining the correlation values in these two frames together, they show exactly the same results as the `ccf` function does in Table 5.2.2.

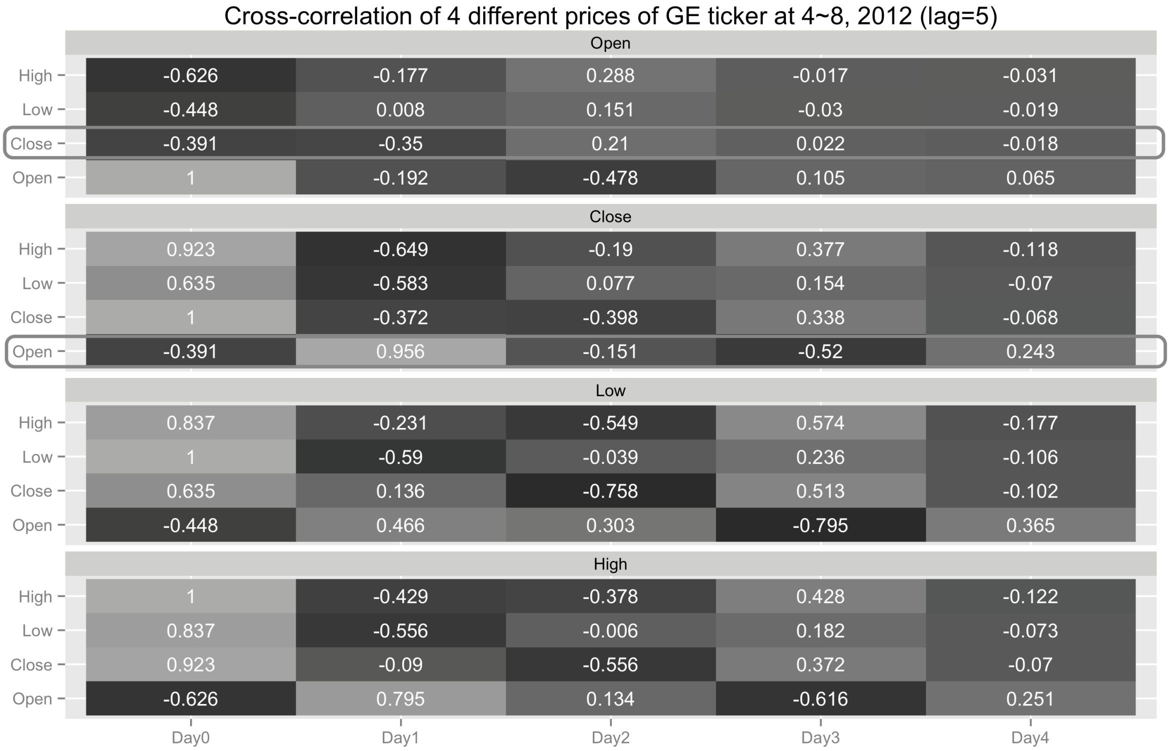


Figure 5.2.1 Cross-correlation result of GE price as shown in Section 5.3

Thus, the `acf` function in `R` can be used to test the correlation of two or more groups of time series data. It is much more robust than the `ccf` function in R, because it can compute the cross-correlation for more than two groups of data at the same time. Thus, the following analysis will use `acf` in `R` to test the cross-correlations of the time series data.

5.2.4 Correlation direction of the ACF and the CCF function

Another key problem is how the `acf` and `ccf` functions correlate data. The time series data in the following analysis consist of two groups: price data and sentiment data. In principle, using the current sentiment data to correlate with the future price data is preferable, because this is the way to predict the price movement based on the sentiment information. The R Team (2013) gives the following definition of the lag of `ccf` functions:

> The lag *k* value returned by the CCF (x,y) estimates the correlation between $\mathrm{x}_{[t+k]}$ and $\mathrm{y}_{[t]}$ (p. 1015).

In other words, the lag *k* value can be either positive or negative. The `ccf` function moves the time series *x* to the future time lags if *k* is positive, or to the past time lags if *k* is negative, but keeps the time series *y* unmoved.

However, in the `acf` function, any lag *k* value is positive, but the movement is bilateral: it first moves the time series *x* to the future by lag *k*, also keeps time series *y* unmoved; then it moves time series *y* to the future by lag *k* too, but keeps time series *x* unmoved. Because the lag *k* is always positive, combining these two movements together is equivalent to the movement in the `ccf` function. Thus, if the prediction of prices based on tweet sentiments is desired, it is reasonable to regard the price data as the time series *x*, and the sentiment data as the time series *y*.

One problem of using the cross-correlation test to compare two groups of univariate data is that it can produce redundant data. Considering that the research aim is to predict the price based on the sentiment, the following analysis only presents the necessary the CCF results in the plotting.

5.2.5 Plotting the ACF test results in R

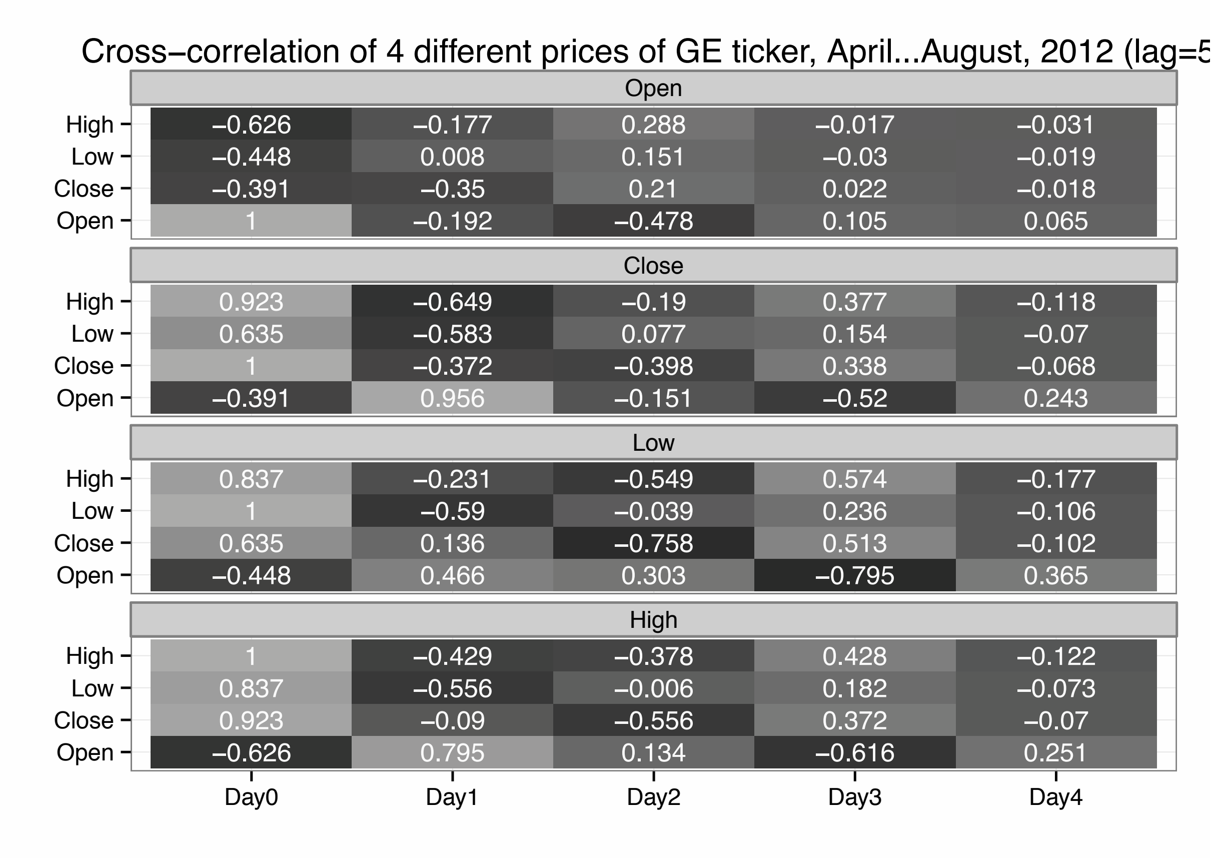
> While current automatic text-processing techniques mean that we are largely restricted to considering lexis rather than more complex discourse semantic features, the visualisation offers some initial guidance to the discourse analyst attempting to understand the unfolding complexity of streams of microposts. (Zappavigna, 2012, p. 42)

> Note that it might be more appropriate to produce a visualisation of sentiment data rather than a textual summary of it. (Pang & Lee, 2008, p. 4)

This research involved much quantitative analysis. A problem with this kind of analysis is that it tends to generate vast amounts of numerical data, which can be difficult to present or summarise in a concise and intelligible way. To assist the reader's understanding of data, therefore, the present study makes use of data visualisation techniques. The research mainly used `R`'s **ggplot2** package (Wickham, 2009) to visualise data.

Heat mapping is a convenient approach to display matrix data. Basically, it uses different colours to illustrate the values of a matrix: The more similar the colours are, the closer the values are; the darker the colours are, the smaller the values are. It is common to change the time lags in the ACF test to see if there is any correlation between different groups of time series data within different time stamps. Thus, a multi-matrix is generated by the ACF test, and plotting it on a heat map is an easy way to display the result. Figure 5.2.2 is a typical heat map of multi-matrix plotting. In this heat map, it shows four 5 by 5 matrices generated by an ACF test. Each row is an individual matrix, so these four matrices show the ACF test results of four groups of time series data.

In each matrix, each column shows the correlation in different time lags, and each row shows different groups of time series data. For instance, in the first matrix in Figure 5.2.2, it is the ACF test result between open price and the other three prices. The first column presents the same information as shown in the first row of Figure 5.2.2. The second column moves all four groups of data to the next time lag, as do the rest of the columns. From top to bottom, the rows show the time series data of the high, low, close and open price respectively.



Open	Day0	Day1	Day2	Day3	Day4
High	–0.626	–0.177	0.288	–0.017	–0.031
Low	–0.448	0.008	0.151	–0.03	–0.019
Close	–0.391	–0.35	0.21	0.022	–0.018
Open	1	–0.192	–0.478	0.105	0.065

Close	Day0	Day1	Day2	Day3	Day4
High	0.923	–0.649	–0.19	0.377	–0.118
Low	0.635	–0.583	0.077	0.154	–0.07
Close	1	–0.372	–0.398	0.338	–0.068
Open	–0.391	0.956	–0.151	–0.52	0.243

Low	Day0	Day1	Day2	Day3	Day4
High	0.837	–0.231	–0.549	0.574	–0.177
Low	1	–0.59	–0.039	0.236	–0.106
Close	0.635	0.136	–0.758	0.513	–0.102
Open	–0.448	0.466	0.303	–0.795	0.365

High	Day0	Day1	Day2	Day3	Day4
High	1	–0.429	–0.378	0.428	–0.122
Low	0.837	–0.556	–0.006	0.182	–0.073
Close	0.923	–0.09	–0.556	0.372	–0.07
Open	–0.626	0.795	0.134	–0.616	0.251

Figure 5.2.2 Example of a heat map of a set of multi-matrix data

5.2.6 Pearson product-moment correlation coefficient

To understand the robustness of the cross-correlation coefficients, the Pearson product-moment correlation coefficient *r* is used. It is "a purely descriptive measure of degree of linear relationship between two variables" (Cohen, 1988, p. 75). The CCF correlation ranges from -1 to 1, which matches the limit of in Pearson's r. According to (Cohen, 1988, pp. 79–80), the robustness of *r* can be defined as shown in Table 5.2.3, and the effect size of Pearson's *r* can be described as *none*, *small*, *medium*, or *strong*.

Table 5.2.3 The effect size of Pearson's *r*

Correlation	**Negative**	**Positive**
None	−0.09 to 0.0	0.0 to 0.09
Small	−0.3 to −0.1	0.1 to 0.3
Medium	−0.5 to −0.3	0.3 to 0.5
Strong	−1.0 to −0.5	0.5 to 1.0

This evaluation might be "biased as in a 'soft' direction" (Cohen, 1988, p. 79), but considering the complicated nature of the data in linguistics and economics, this can be regarded as a proper evaluation. The analysis presented below was particularly interested in the strong correlations from the CCF test results.

5.2.7 Imputation

As presented in Section 5.1, from April 1 to August 31, 2012, there were 153-day continuous tweet data, but only 108-days of stock price data. This 29% margin makes the correlation analysis difficult. To solve this problem, statistics has different data manipulation methods, known as imputation. This analysis used one imputation method to overcome the missing value problem, namely the last observation carried forward. Enders (2010) defines the last observation carried forward method as one in which "the procedure imputes missing repeated measures variables with the observation that immediately precedes dropout" (pp. 51 – 52). In other words, it reduces the impact of the missing value by replacing the missing value with the last non-missing value. Moreover, This method is "specific to longitudinal designs" (2010, p. 51). In R, this imputation can be done with the na.contiguous function. Figure 5.2.3 shows that replacing the missing values with the last non-missing values makes the trends much smoother. Thus, the imputation method to overcome the missing value problem encountered in this research.

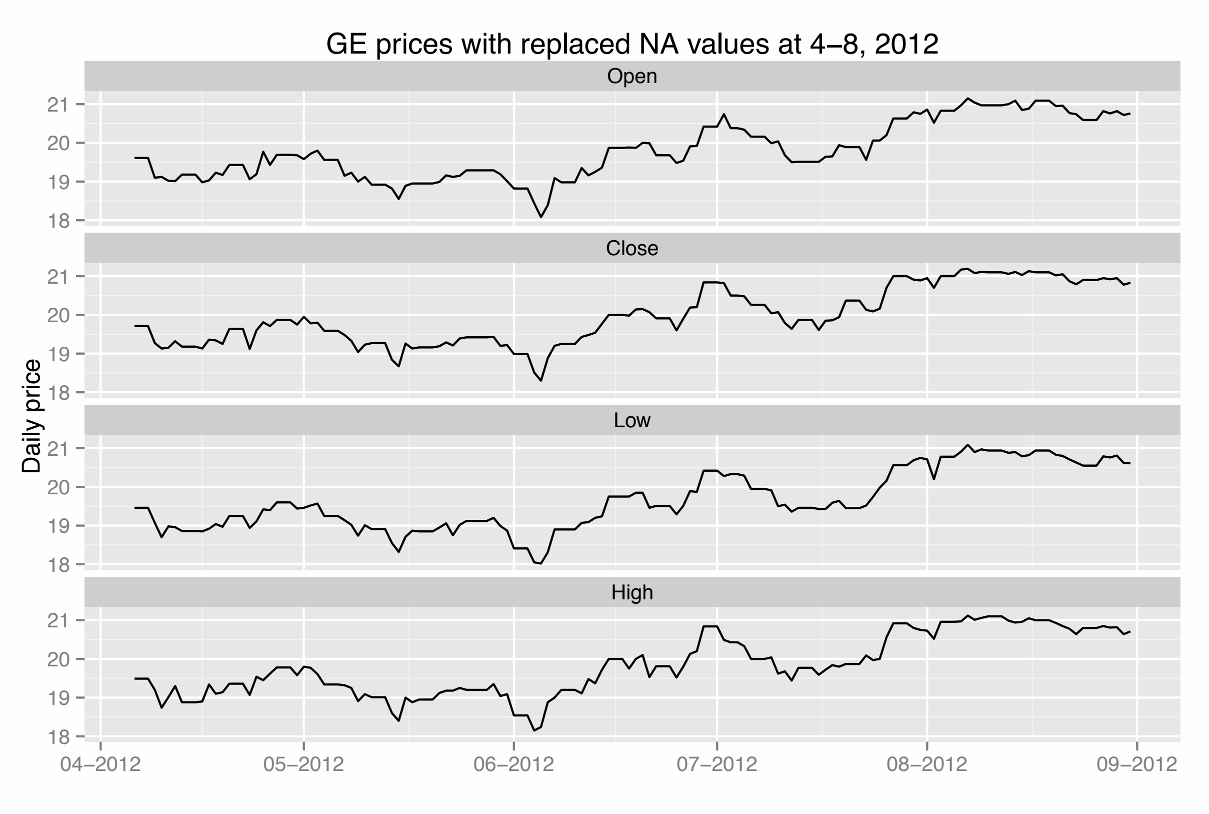


Figure 5.2.3 Plot of GE prices with replaced NA values, April-August, 2012

Two other imputation methods, the window function and the moving average smoothing, are introduced in detail in Section 5.4.2 and 5.4.3.

5.3 Time Series Analysis of Raw Data

This section reports on temporal correlations within raw stock price data, within tweet sentiment data, and between stock price and tweet sentiment.

5.3.1 Time correlation of the raw GE ticker's prices

The first test focused on the GE price data collected as described in Section 5.1. In the upper part of Figure 5.3.1, from April to August, 2012, the overall trend of the GE price was to increase. In other words, these four prices had a bullish trend. From top to bottom, they are the open, close, high, and low price of the GE ticker respectively, and all plotting lines are continuous. This is a typical plot of price trends in real applications of financial analysis.

However, public holidays do not have any price information. Therefore, if the plot includes the 45-day missing data, the result becomes completely different as displayed in the lower part of Figure 5.3.1. All four plotting lines were cut into 24 pieces. As shown by the vertical dashed lines in the lower part of Figure 5.3.1, the longest break were for three days (April 6 to 8, and May 26 to 28), and in the 9 days from June 30 to July 8, there were 3 breaks of 5 holidays. This required extra attention in the analysis reported below (see Section 5.3.4).

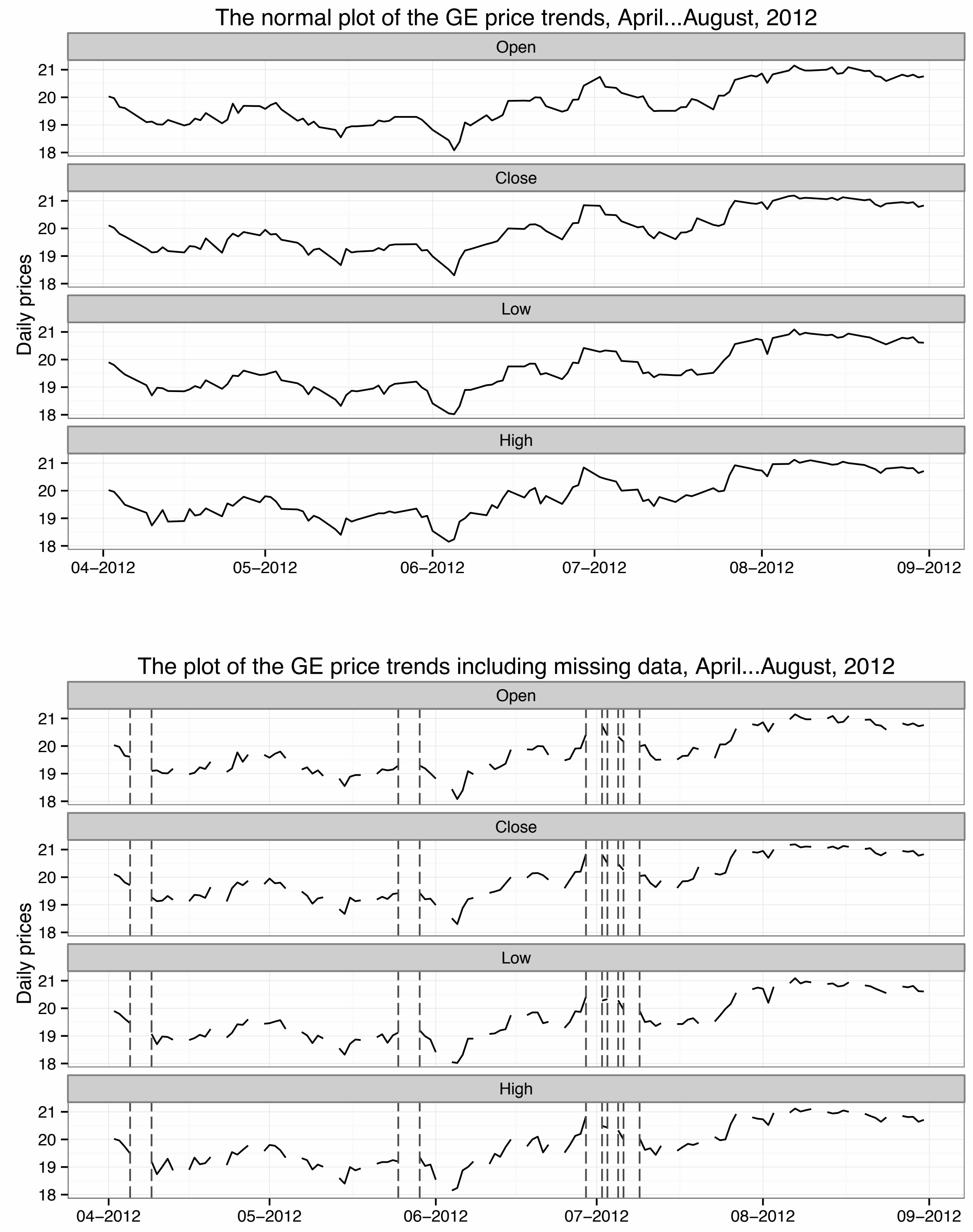
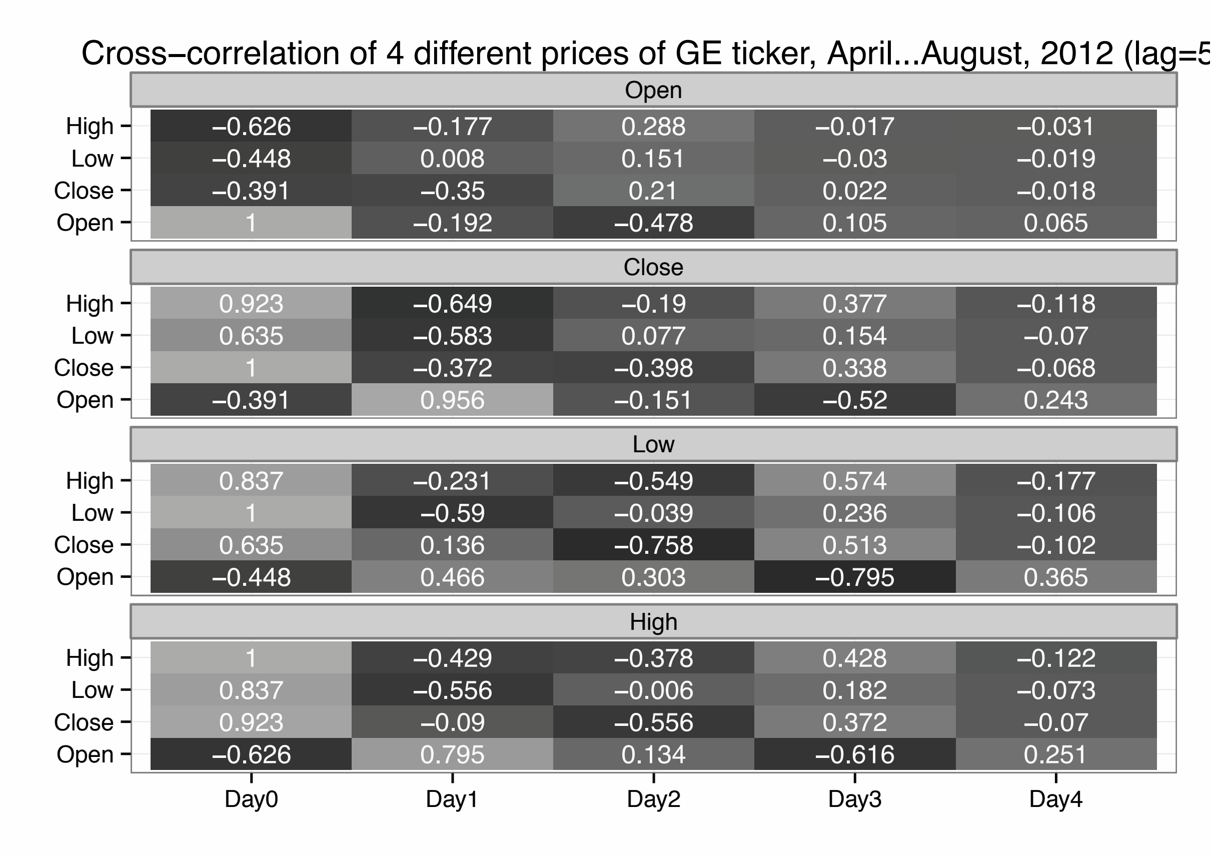


Figure 5.3.1 The normal plot of price trends of the GE ticker, April-August, 2012

Figure 5.3.2 shows the correlation result of the CCF test. The first column indicates the correlation at Day$_0$: Clearly, the trend of the open price differs most from the other three, which negatively correlates to the others, because the open price has a stronger correlation with the previous day's price trend than the current day's value. Besides, the high price closely correlates to the low and close price; but the correlation between the low price and close price is not highly positive. In addition, moving the open price to the Day$_1$ lag, it becomes positively correlated to the close and high price, and the correlation changes in low price is also obvious: from -.448 to 0.466 in Pearson's *r*. Finally, the Day$_4$ open price becomes the most negatively related to the other three, but the other three fourth-day prices become the second most positively correlated to each other.



Cross–correlation of 4 different prices of GE ticker, April...August, 2012 (lag=5

Open

	Day0	Day1	Day2	Day3	Day4
High	–0.626	–0.177	0.288	–0.017	–0.031
Low	–0.448	0.008	0.151	–0.03	–0.019
Close	–0.391	–0.35	0.21	0.022	–0.018
Open	1	–0.192	–0.478	0.105	0.065

Close

	Day0	Day1	Day2	Day3	Day4
High	0.923	–0.649	–0.19	0.377	–0.118
Low	0.635	–0.583	0.077	0.154	–0.07
Close	1	–0.372	–0.398	0.338	–0.068
Open	–0.391	0.956	–0.151	–0.52	0.243

Low

	Day0	Day1	Day2	Day3	Day4
High	0.837	–0.231	–0.549	0.574	–0.177
Low	1	–0.59	–0.039	0.236	–0.106
Close	0.635	0.136	–0.758	0.513	–0.102
Open	–0.448	0.466	0.303	–0.795	0.365

High

	Day0	Day1	Day2	Day3	Day4
High	1	–0.429	–0.378	0.428	–0.122
Low	0.837	–0.556	–0.006	0.182	–0.073
Close	0.923	–0.09	–0.556	0.372	–0.07
Open	–0.626	0.795	0.134	–0.616	0.251

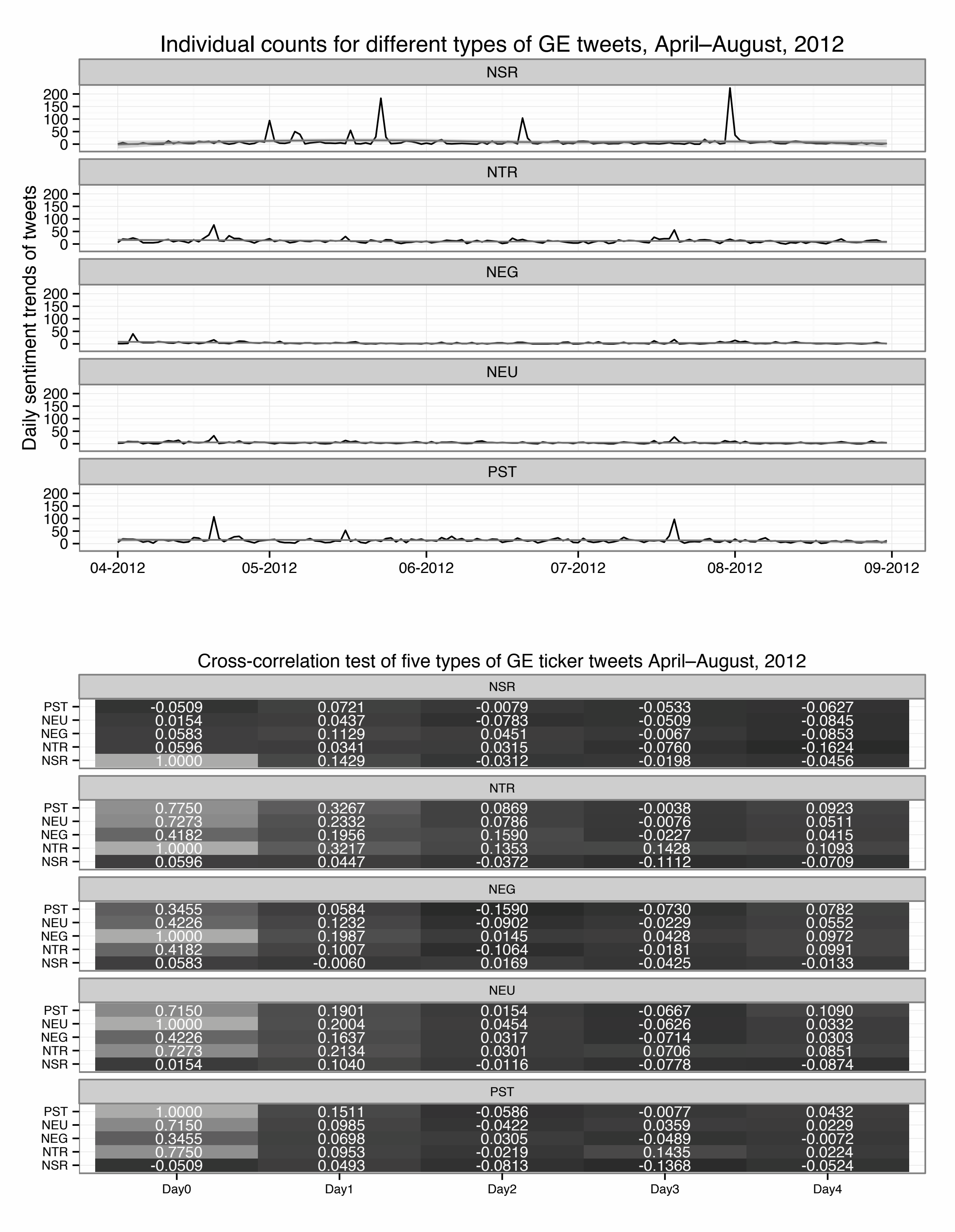
Figure 5.3.2 The CCF test of four prices of the GE ticker, April-August, 2012

5.3.2 Time correlation of the raw GE tweet sentiments

The second test focuses on the raw GE tweet count. Without any manipulation, it uses the counts of the different types of annotated GE tweets as the record of each day.

Plotting the actual counts of five types of tweets individually is clearer (see the upper part of Figure 5.3.3). In these five trends, the NSR trend is the least stable. It has six peaks, and at the end of July (July 31), it has more than 200 daily tweets. This unique trend shows that the NSR tweets are less relevant to the overall tweet sentiment of the GE ticker in terms of the temporal relationship. Within the stock-related tweets, more similarities exist than differences. There are fewer fluctuations in the entire period, and during late April and late July, there are two echoing climbs in all four types of tweets. Moreover, the climbs in PST tweets are most apparent because the overall trend of GE price in this period is bullish as shown in the last section.

Applying the CCF test to these five groups of time series data (see the lower part of Figure 5.3.3), it is clear that there are positive correlations between the NTR tweets and the PST and NEU tweets only at Day_0. Apart from this, there is no strong positive correlation between any types of tweets at any other time lags. These show that, among different types of sentiments, there is less correlation than found with the different types of price data aforementioned.



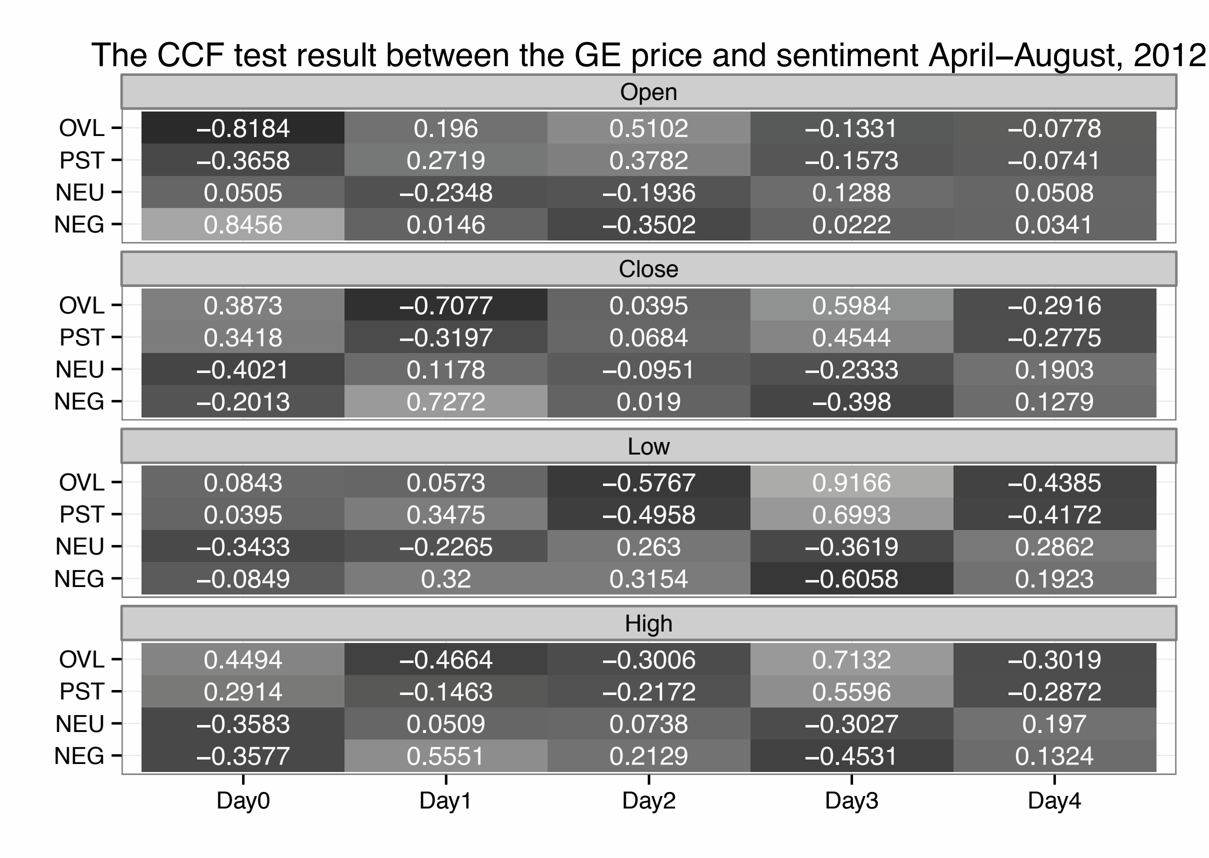
Panel	Row	Day0	Day1	Day2	Day3	Day4
NSR	PST	-0.0509	0.0721	-0.0079	-0.0533	-0.0627
NSR	NEU	0.0154	0.0437	-0.0783	-0.0509	-0.0845
NSR	NEG	0.0583	0.1129	0.0451	-0.0067	-0.0853
NSR	NTR	0.0596	0.0341	0.0315	-0.0760	-0.1624
NSR	NSR	1.0000	0.1429	-0.0312	-0.0198	-0.0456
NTR	PST	0.7750	0.3267	0.0869	-0.0038	0.0923
NTR	NEU	0.7273	0.2332	0.0786	-0.0076	0.0511
NTR	NEG	0.4182	0.1956	0.1590	-0.0227	0.0415
NTR	NTR	1.0000	0.3217	0.1353	0.1428	0.1093
NTR	NSR	0.0596	0.0447	-0.0372	-0.1112	-0.0709
NEG	PST	0.3455	0.0584	-0.1590	-0.0730	0.0782
NEG	NEU	0.4226	0.1232	-0.0902	-0.0229	0.0552
NEG	NEG	1.0000	0.1987	0.0145	0.0428	0.0972
NEG	NTR	0.4182	0.1007	-0.1064	-0.0181	0.0991
NEG	NSR	0.0583	-0.0060	0.0169	-0.0425	-0.0133
NEU	PST	0.7150	0.1901	0.0154	-0.0667	0.1090
NEU	NEU	1.0000	0.2004	0.0454	-0.0626	0.0332
NEU	NEG	0.4226	0.1637	0.0317	-0.0714	0.0303
NEU	NTR	0.7273	0.2134	0.0301	0.0706	0.0851
NEU	NSR	0.0154	0.1040	-0.0116	-0.0778	-0.0874
PST	PST	1.0000	0.1511	-0.0586	-0.0077	0.0432
PST	NEU	0.7150	0.0985	-0.0422	0.0359	0.0229
PST	NEG	0.3455	0.0698	0.0305	-0.0489	-0.0072
PST	NTR	0.7750	0.0953	-0.0219	0.1435	0.0224
PST	NSR	-0.0509	0.0493	-0.0813	-0.1368	-0.0524

Figure 5.3.3 Plot and the ACF test result of different types of GE tweet raw counts, April-August, 2012

5.3.3 Time correlation between the raw GE price and tweet sentiments

To understand the correlation between price and tweet sentiment of the GE ticker, a CCF test was carried out between them. The overall sentiment (OVL) is simply calculated as the PST sentiment subtracted by the NEG sentiment. It considers the open, high, low, and close price trends as the price trends, and the NEG, NEU, PST and OVL sentiment trends as sentiment trends.

In the CCF test result, (see Figure 5.3.4), at Day$_0$, the open price has a strong negative correlation between the OVL sentiment (*r* = -.8184), and a strong positive correlation with NEG tweets (*r* = .8456).



The CCF test result between the GE price and sentiment April–August, 2012

		Day0	Day1	Day2	Day3	Day4
Open	OVL	−0.8184	0.196	0.5102	−0.1331	−0.0778
	PST	−0.3658	0.2719	0.3782	−0.1573	−0.0741
	NEU	0.0505	−0.2348	−0.1936	0.1288	0.0508
	NEG	0.8456	0.0146	−0.3502	0.0222	0.0341
Close	OVL	0.3873	−0.7077	0.0395	0.5984	−0.2916
	PST	0.3418	−0.3197	0.0684	0.4544	−0.2775
	NEU	−0.4021	0.1178	−0.0951	−0.2333	0.1903
	NEG	−0.2013	0.7272	0.019	−0.398	0.1279
Low	OVL	0.0843	0.0573	−0.5767	0.9166	−0.4385
	PST	0.0395	0.3475	−0.4958	0.6993	−0.4172
	NEU	−0.3433	−0.2265	0.263	−0.3619	0.2862
	NEG	−0.0849	0.32	0.3154	−0.6058	0.1923
High	OVL	0.4494	−0.4664	−0.3006	0.7132	−0.3019
	PST	0.2914	−0.1463	−0.2172	0.5596	−0.2872
	NEU	−0.3583	0.0509	0.0738	−0.3027	0.197
	NEG	−0.3577	0.5551	0.2129	−0.4531	0.1324

Figure 5.3.4 The CCF test result of between the price and sentiment of GE ticker, April-August, 2012

The close price has a strong negative correlation with the next day (Day$_1$) OVL sentiment (*r* =

-.7077), and a strong positive correlation with the next-day NEG (r = .7272). Moreover, there is a strong positive correlation between the low price and Day$_3$'s OVL sentiment (r = .9166), and strong positive correlations exist between the low price and Day$_3$'s PST sentiment (r = .6993), and the high price with Day$_3$'s OVL sentiment (r = .7132). Apart from these, there is no strong correlation between the price and the sentiment.

5.3.4 Discussion

Overall, this section reported on the analysis of three groups of temporal relationships: within the raw price data, within the raw tweet sentiment data, and between the raw stock price and tweet sentiment data.

This first part analysed the overall trend of the four prices of the GE ticker from April to August, 2012. The four prices showed a high overall similarity regardless of some minor local differences. The open price had an opposite trend with the other three price trends, and the non-stock and non-ticker related tweet trends showed fewer correlations with the other ticker-correlated tweets. To summarise, the open price negatively correlated to the other three prices at the Day$_0$, but the high price highly correlates to the close price and low price at the Day$_0$. There was a 1-day lag between the open price and the other three prices: The next-day open price highly correlated to the high and low price. At the Day$_3$ lag, there was an opposite trend between the open price and the other three prices: The Day$_3$'s open price was the most negatively correlated to the other prices. However, there were many holiday breaks in this group of data, which results in a 29% data margin.

In the second part, the time series analysis of the sentiment data reported above showed that the five groups of GE sentiments exhibited a very different trend, especially the NSR tweets. This suggests that the NSR tweets temporally differ from stock-related tweets. The CCF results showed that some positive correlations exist between the NTR and the PST and NEU groups at the Day_0, but no strong correlation existed between the other groups at the Day_0. Moreover, some much weaker correlations were found within ticker-related tweets and between the NTR and three groups of ticker-related tweets on the other days.

Finally, the third part demonstrated that, between the raw price and tweet sentiment data, there was a strong correlation between the NEG trend and the open price at the Day_0, and the OVL sentiment trend had the most correlations with the price trends. The strongest correlation occurred at the Day_3 lag, where the correlation between the OVL sentiment and the Day_3 low price is 0.9166. The result of the raw GE price and tweet sentiment was then considered as the baseline of the correlation between these two groups of temporal data, and the following experiments compared the results with this baseline.

5.4 Time Series Analysis of Transformed Data

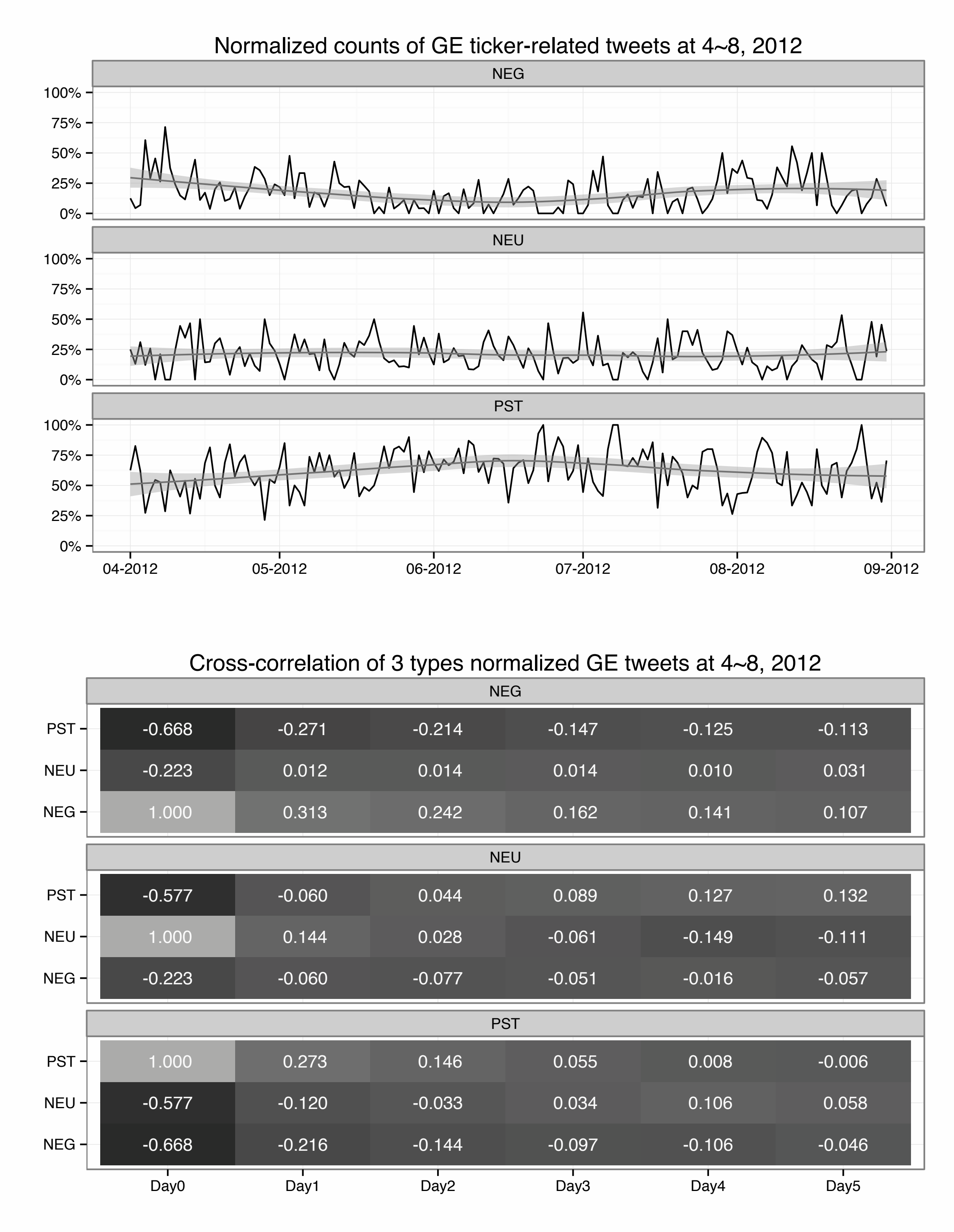
From the above tests, there was another major obstacle in this case study apart from the missing values: the unstable total daily counts of collected tweets. In April 2012, the crawler was set with different collecting rates to find the most efficient crawling rate, which resulted in frequent changes to the collection rate. After this testing period, the collection rate was set to a fixed rate, but the daily amount of collected tweets still varied greatly (see Figure 4.2.1 in Chapter 4). Moreover, this problem was compounded by the fact that the number of tweets being posted to Twitter on any individual day varies. The overall trend of the number of tweets being posted is increasing, but breaking events can bring peaks at certain times. This again may have an impact on the unbalancedness of the data. All of these observations suggest that comparing the absolute counts of tweets could be misleading. Several solutions can be used to reduce the possibility of the misleading results. Either excluding the irrelevant tweets, or using normalised data can be helpful, and combining these two strategies could reduce the noise further. This section focuses on the merits of using percentage normalisation to tackle this problem. The rest of this section then reports on how the window function and moving average smoothing were used to deal with the missing value problem as discussed in Section 5.2.7.

5.4.1 Percentage normalisation

Percentage normalisation is a widely applied technique to reduce the impact of different data sizes at each time point. It uses the total data count at a specific time point to divide the sample data count at that time point. Considering the daily counts of tweets presented in Figure 4.2.2, each day had different counts of these 5 types of tweets, so comparing the actual counts of the tweets can be misleading. Thus, percentage normalisation is introduced to reduce the drastic changes at any given time point, and the analysis used this normalised data to compare daily changes. For example, the normalised PST is calculated as follows:

Normalised PST tweets = PST tweets / (NEG + NEU + PST tweets)

where the denominator is equivalent to the total count of ticker-related tweets. The first experiment was to normalise the data according to the percentage of the daily count of GE ticker-related tweets (see the upper part of Figure 5.4.1): All the trends become more dynamic. Applying a CCF test to these normalised data, the correlations among them are still weak (see the lower part of Figure 5.4.1). Comparing the result with the correlation of the raw counts in Figure 5.3.5, most correlations drop greatly. The main reason is that these three types of sentiment are independent events, but using the percentage normalisation, they become dependent on each other. Hence, this approach weakens the correlations in the previous test.



Cross-correlation of 3 types normalized GE tweets at 4~8, 2012

NEG

	Day0	Day1	Day2	Day3	Day4	Day5
PST	-0.668	-0.271	-0.214	-0.147	-0.125	-0.113
NEU	-0.223	0.012	0.014	0.014	0.010	0.031
NEG	1.000	0.313	0.242	0.162	0.141	0.107

NEU

	Day0	Day1	Day2	Day3	Day4	Day5
PST	-0.577	-0.060	0.044	0.089	0.127	0.132
NEU	1.000	0.144	0.028	-0.061	-0.149	-0.111
NEG	-0.223	-0.060	-0.077	-0.051	-0.016	-0.057

PST

	Day0	Day1	Day2	Day3	Day4	Day5
PST	1.000	0.273	0.146	0.055	0.008	-0.006
NEU	-0.577	-0.120	-0.033	0.034	0.106	0.058
NEG	-0.668	-0.216	-0.144	-0.097	-0.106	-0.046

Figure 5.4.1 Plot and the CCF test result normalised GE ticker-related tweets at, April-August, 2012

However, if adding the percentage counts of the three ticker-related tweets together, the trend becomes different. The overall trend of the transformed percentage normalisation is calculated as follows:

OVL = (PST - NEG) / (NEG + NEU + PST tweets)

where the denominator is equivalent to the total count of ticker-related tweets. This takes the daily percentage of the NEG sentiment as negative value, the NEU sentiment as zero, and the PST sentiment as positive. The upper part of Figure 5.4.2 shows the result of this summary: Generally, most points cluster above 0%; the three vertical lines indicate the -50%, 0%, and 50% range from bottom to top respectively; each point indicates the value of each individual day.

The above approach may ignore the influence of neutral tweets because it considers the value of each neutral tweet to be zero. To improve this, it recalculates the percentage of the positive and negative tweets as follows:

Weighted overall sentiment (WOS) = (PST - NEG) / (PST + NEG)

Clearly, more points cluster at the top of the lower part of Figure 5.4.2, because the denominator (PST + NEG) is smaller than the denominator (PST + NEU + NEG) in the last transformation.

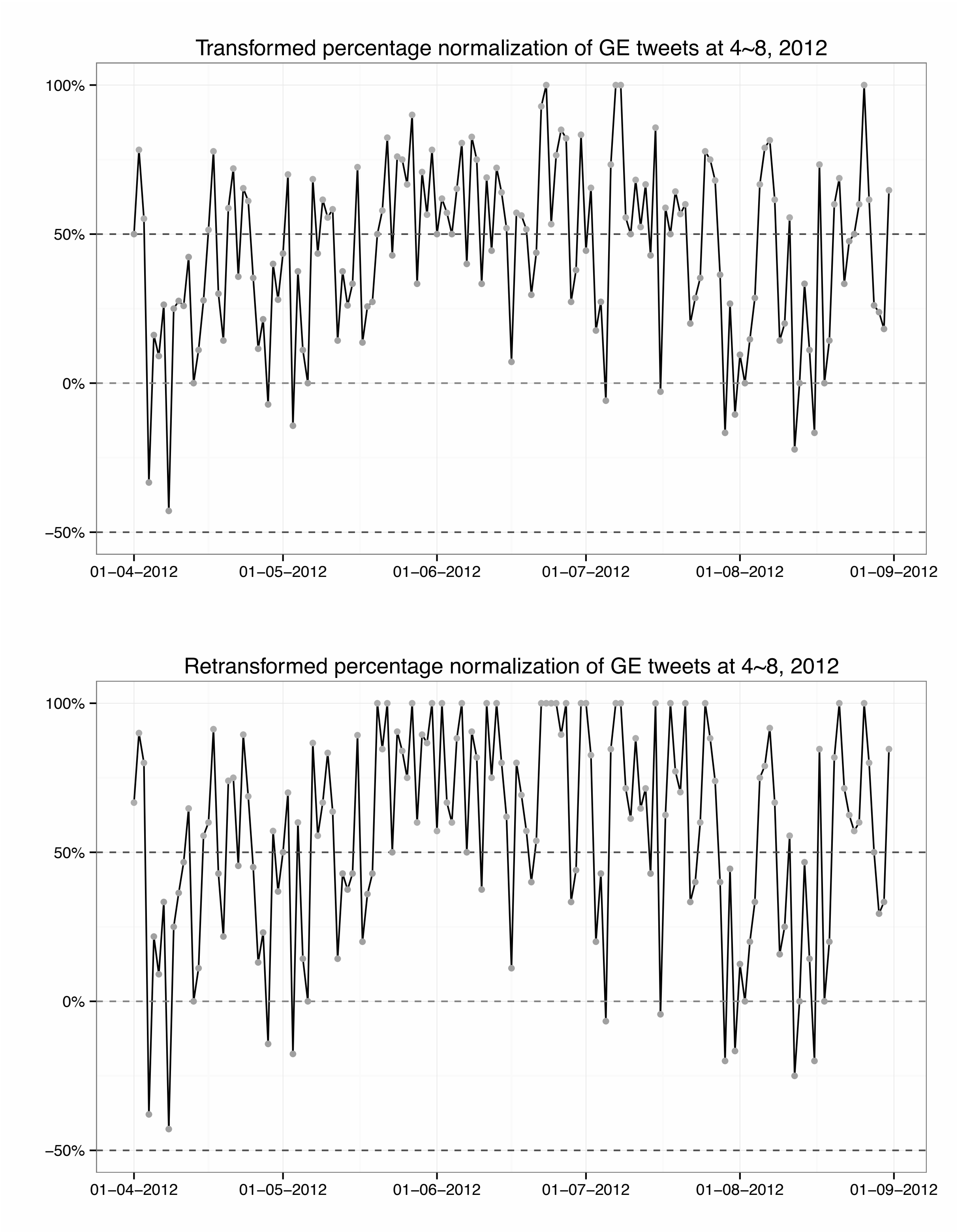


Figure 5.4.2 Transformed and retransformed percentage normalisation of GE tweets, April-August, 2012

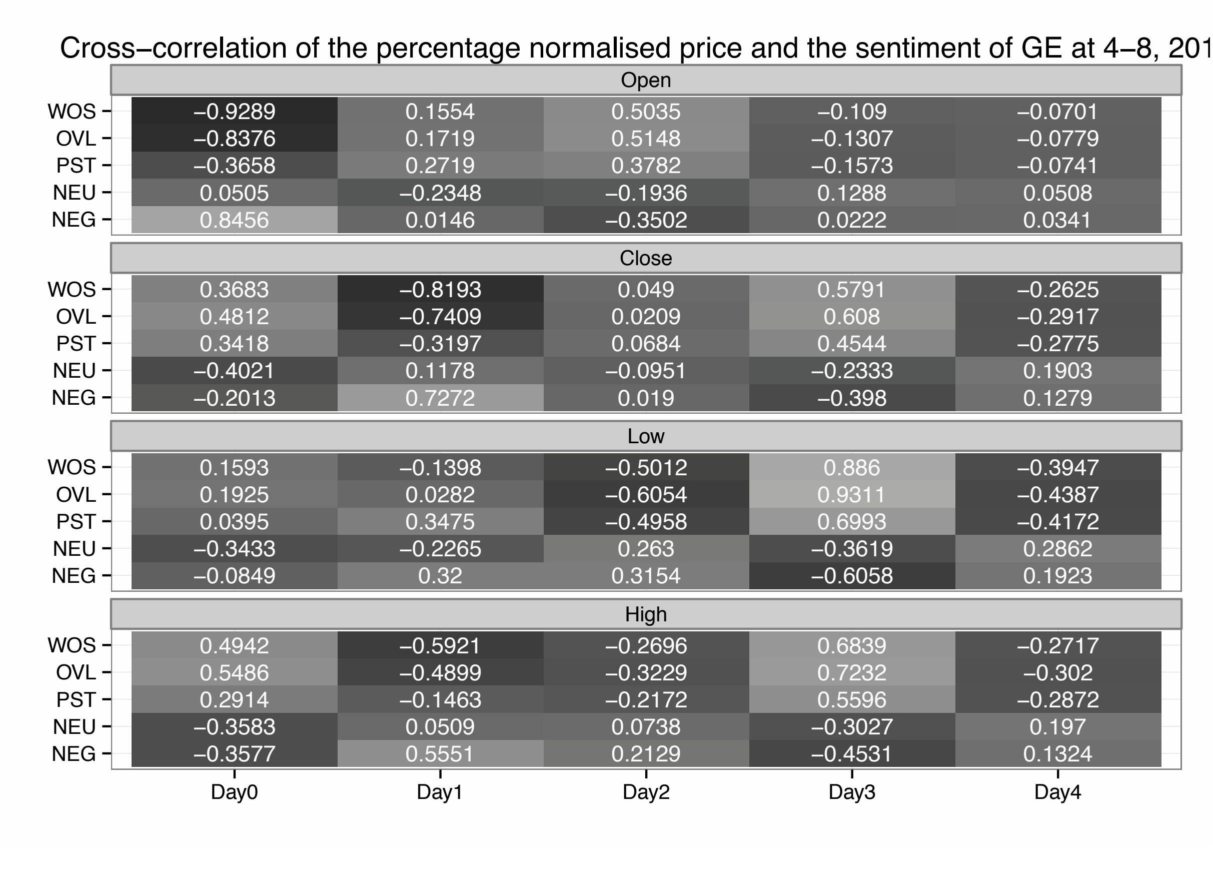
Table 5.4.1 shows that, in using both transformations, most days show a positive score. In the transformed result, nearly half of the days reach a score of more than 50%. This suggests that the overall trend of the GE sentiment is positive based on the percentage normalisation. The four points in the 100% category is due to the absence of negative or neutral tweets on these four days. In the retransformed result, the entire positive group moves upward, and the 100% category shows an obvious difference because the 19 days do not have any NEG tweets.

Table 5.4.1 Comparison of the day counts of the transformed and retransformed percentage normalisation of GE tweets at 4-8, 2012

Method	**-50% - 0**	**0**	**0 - 50%**	**50% - 100%**	**100%**
Transformed	10 (6.54%)	5 (3.27%)	67 (43.79%)	67 (43.79%)	4 (2.61%)
Retransformed	10 (6.54%)	5 (3.27%)	48 (31.13%)	67 (43.79%)	23 (15.03%)

To improve the understanding of the correlation between the price and the normalised sentiment, the CCF test was also applied (see Figure 5.4.3), taking the result of the first normalisation as the OVL trend, and the second normalisation as the WOS trend in the graph. As shown in Figure 5.4.3, at Day$_0$, both the OVL and the WOS trends display a strong negative correlation with the open price (r = -.8376 and r = -.9289 respectively). The strongest correlation occurs at the Day$_3$ lag, where the r between the OVL sentiment and the Day$_3$ low price is 0.9311. Apart from this, there are two strong negative correlations between the close price and the Day$_1$'s OVL and WOS trends on the one-day lag (r = -.7409 and r = -.8193 respectively). Also, at the Day$_3$, there is a strong correlation between the low price and WOS trends (r = .886), and a strong positive correlation between the high price and the OVL trend (r = .7232). There is no other strong correlation between the normalised overall sentiment and the price data.

Comparing the CCF result with the baseline discussed in Section 5.3.3, the three individual sentiments, NEG, NEU and PST sentiments keep exactly the same correlations, but the OVL sentiment has a slight improvement. At the Day_0 lag, the OVL sentiment becomes more positively correlated to the close, low and high prices, but more negatively correlated to the open price. Compared with the peak correlation occurring in the Day_3 lag in the base line, the OVL sentiment becomes more positively correlated to the low and high prices. In addition, compared with the OVL sentiment in the baseline, the correlations of the WOS sentiment become stronger at the same lag except for the correlation between the close price and the WOS sentiment, but slightly weaker at the other time lags.



Open	Day0	Day1	Day2	Day3	Day4
WOS	−0.9289	0.1554	0.5035	−0.109	−0.0701
OVL	−0.8376	0.1719	0.5148	−0.1307	−0.0779
PST	−0.3658	0.2719	0.3782	−0.1573	−0.0741
NEU	0.0505	−0.2348	−0.1936	0.1288	0.0508
NEG	0.8456	0.0146	−0.3502	0.0222	0.0341

Close	Day0	Day1	Day2	Day3	Day4
WOS	0.3683	−0.8193	0.049	0.5791	−0.2625
OVL	0.4812	−0.7409	0.0209	0.608	−0.2917
PST	0.3418	−0.3197	0.0684	0.4544	−0.2775
NEU	−0.4021	0.1178	−0.0951	−0.2333	0.1903
NEG	−0.2013	0.7272	0.019	−0.398	0.1279

Low	Day0	Day1	Day2	Day3	Day4
WOS	0.1593	−0.1398	−0.5012	0.886	−0.3947
OVL	0.1925	0.0282	−0.6054	0.9311	−0.4387
PST	0.0395	0.3475	−0.4958	0.6993	−0.4172
NEU	−0.3433	−0.2265	0.263	−0.3619	0.2862
NEG	−0.0849	0.32	0.3154	−0.6058	0.1923

High	Day0	Day1	Day2	Day3	Day4
WOS	0.4942	−0.5921	−0.2696	0.6839	−0.2717
OVL	0.5486	−0.4899	−0.3229	0.7232	−0.302
PST	0.2914	−0.1463	−0.2172	0.5596	−0.2872
NEU	−0.3583	0.0509	0.0738	−0.3027	0.197
NEG	−0.3577	0.5551	0.2129	−0.4531	0.1324

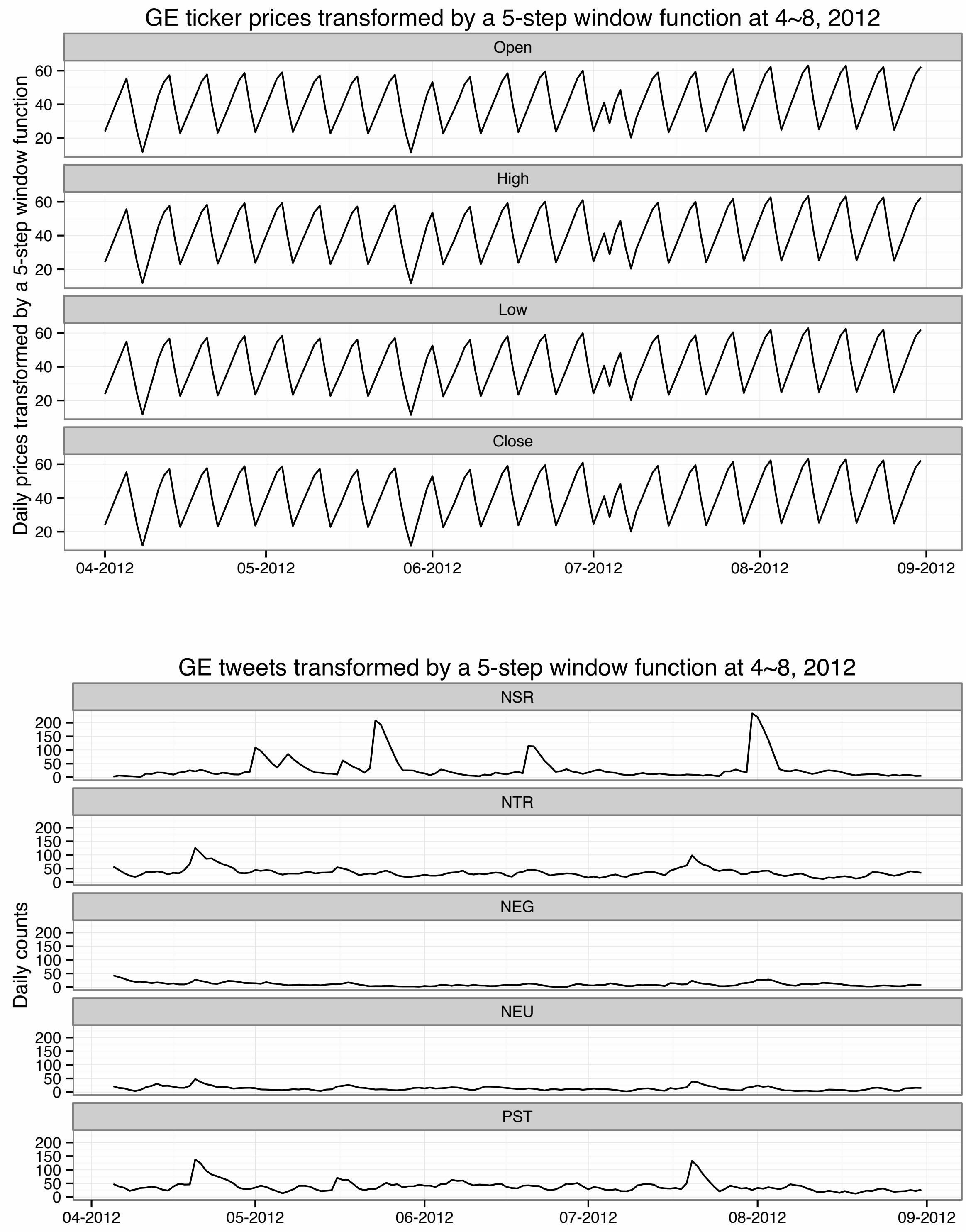
Figure 5.4.3 The CCF test result of the percentage normalised price and the sentiment of the GE ticker, April-August, 2012

5.4.2 Window function Transformation

Window function, a common approach in signal processing, "uses a gradual time-window to minimise the effects of sharp frequency-domain transitions" (Cavicchi, 2000, p. 614). According to this, the transformed value is a weighted sum of the previous values. Considering the longest break in the price trends is three days, the step of the window function is designed as five days:

> Transformed value of Day_{n} = value of Day_{n} * 100% + value of Day_{n-1} * 80% + value of Day_{n-2} * 60% + value of Day_{n-3} * 40% + value of Day_{n-4} * 20%
>
> where Day_{n-1} is the first day before Day_{n}, Day_{n-2} is the second day before Day_{n}, and so on.

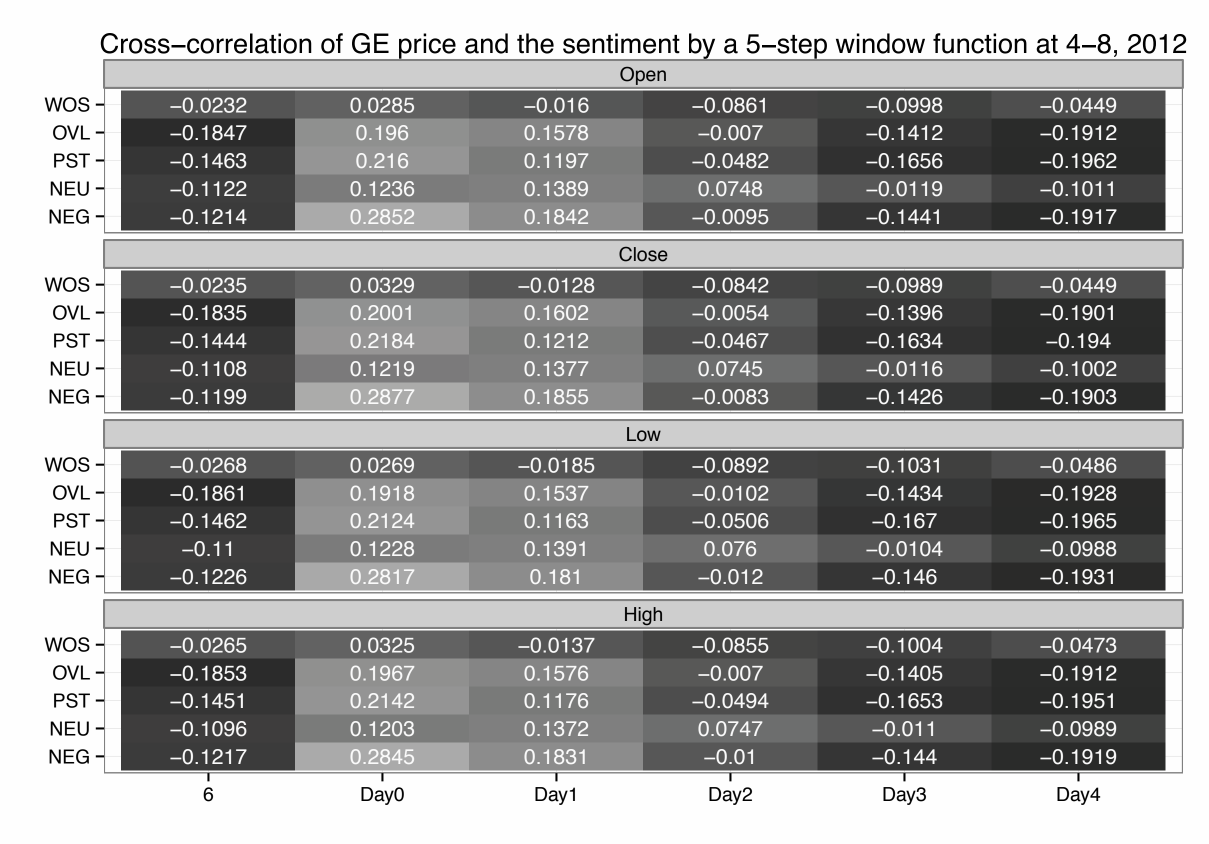
Thus, this function can reduce the missing value problem in time series data, helping to deal with the stock price data as outlined above. The upper part of Figure 5.4.4 shows the result of applying the window function above to the four types of GE ticker price. As might be expected, they become similar. Transforming the GE tweet sentiment data with the same procedure, the trends for each type of tweets become smoother (see the lower part of Figure 5.4.4). However, this is completely different from the price data transformation.


GE ticker prices transformed by a 5-step window function at 4~8, 2012

GE tweets transformed by a 5-step window function at 4~8, 2012

Figure 5.4.4 Plot of GE price normalised by a 5-step window function, April-August, 2012

Applying the CCF test between the price and sentiment data showed a different trend (see Figure 5.4.5). Compared to the baseline of the GE sentiment and price, the CCF test showed that the window function weakens correlations, and even worse, no strong correlation could be found in this result. Thus, this approach is not helpful in this case.



Cross–correlation of GE price and the sentiment by a 5–step window function at 4–8, 2012

Open

	6	Day0	Day1	Day2	Day3	Day4
WOS	−0.0232	0.0285	−0.016	−0.0861	−0.0998	−0.0449
OVL	−0.1847	0.196	0.1578	−0.007	−0.1412	−0.1912
PST	−0.1463	0.216	0.1197	−0.0482	−0.1656	−0.1962
NEU	−0.1122	0.1236	0.1389	0.0748	−0.0119	−0.1011
NEG	−0.1214	0.2852	0.1842	−0.0095	−0.1441	−0.1917

Close

	6	Day0	Day1	Day2	Day3	Day4
WOS	−0.0235	0.0329	−0.0128	−0.0842	−0.0989	−0.0449
OVL	−0.1835	0.2001	0.1602	−0.0054	−0.1396	−0.1901
PST	−0.1444	0.2184	0.1212	−0.0467	−0.1634	−0.194
NEU	−0.1108	0.1219	0.1377	0.0745	−0.0116	−0.1002
NEG	−0.1199	0.2877	0.1855	−0.0083	−0.1426	−0.1903

Low

	6	Day0	Day1	Day2	Day3	Day4
WOS	−0.0268	0.0269	−0.0185	−0.0892	−0.1031	−0.0486
OVL	−0.1861	0.1918	0.1537	−0.0102	−0.1434	−0.1928
PST	−0.1462	0.2124	0.1163	−0.0506	−0.167	−0.1965
NEU	−0.11	0.1228	0.1391	0.076	−0.0104	−0.0988
NEG	−0.1226	0.2817	0.181	−0.012	−0.146	−0.1931

High

	6	Day0	Day1	Day2	Day3	Day4
WOS	−0.0265	0.0325	−0.0137	−0.0855	−0.1004	−0.0473
OVL	−0.1853	0.1967	0.1576	−0.007	−0.1405	−0.1912
PST	−0.1451	0.2142	0.1176	−0.0494	−0.1653	−0.1951
NEU	−0.1096	0.1203	0.1372	0.0747	−0.011	−0.0989
NEG	−0.1217	0.2845	0.1831	−0.01	−0.144	−0.1919

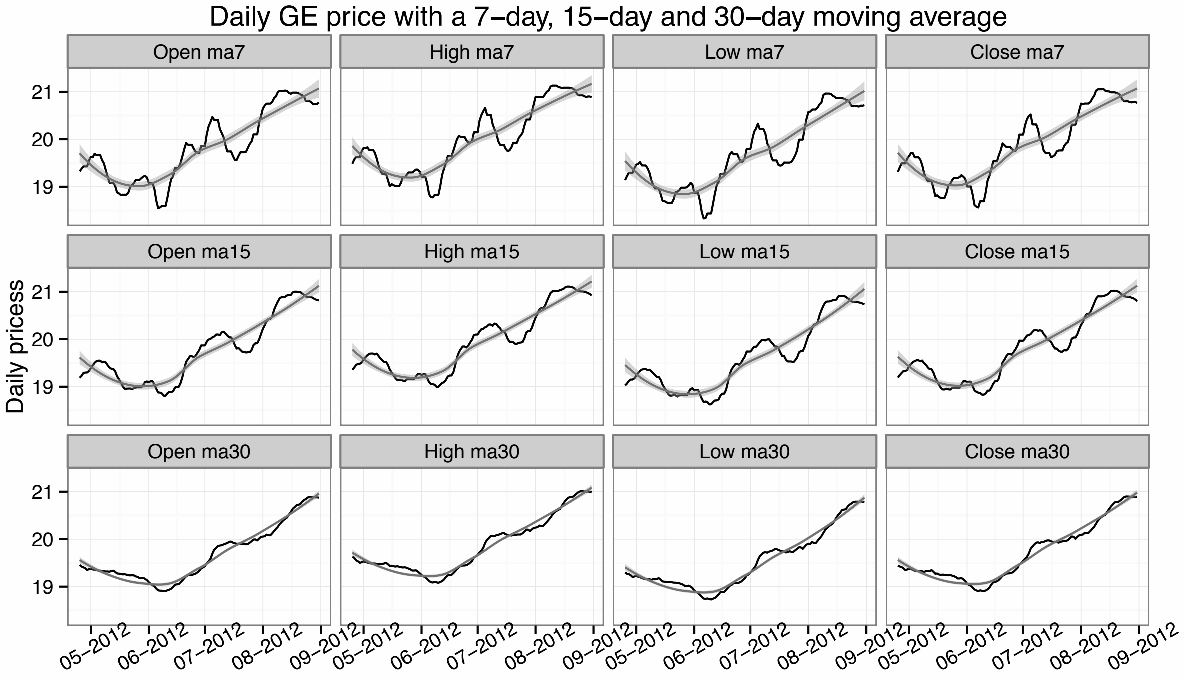
Figure 5.4.5 Cross-correlation result of GE sentiment and price by a 5-step window function, April-August, 2012

5.4.3 Moving average smoothing

Moving average smoothing is another common approach for dealing with temporal data because it is “useful in discovering certain traits in a time series, such as long-term trend and seasonal components” (Shumway & Stoffer, 2006, p. 75). It systematically takes the mean of the neighbours of the target item to replace the actual value. Using this technique, the noise in the temporal data can be effectively reduced, and therefore, the overall trend is displayed.

In particular, O’Connor et al. (2010, p. 125) pointed out that smoothing is critical in that “it causes the sentiment ratio to respond more slowly to recent changes, thus forcing consistent behaviour to appear over longer periods of time”. Furthermore, according to their study, applying 7-, 15-, and 30-day windows to the Gallop poll data, the correlation reach 71.6%, 76.3%, and 79.4% respectively.

Similarly, the same windows for the moving average smoothing were applied in this research applies. The upper part of Figure 5.4.6 displays the moving average smoothing results of GE price and sentiment with 7-, 15-, and 30-day windows: from the top to bottom, they are the 7-, 15- and 30-day moving average results respectively. It shows a gradually smoothing trend from top to bottom and an overall increasing trend of the price movement. Meanwhile, the curves in the GE sentiment graphs (see the lower part of Figure 5.4.6) show gradually stable trends too. These suggest that using moving average smoothing is a useful approach to reducing the drastic daily changes in tweet sentiment.



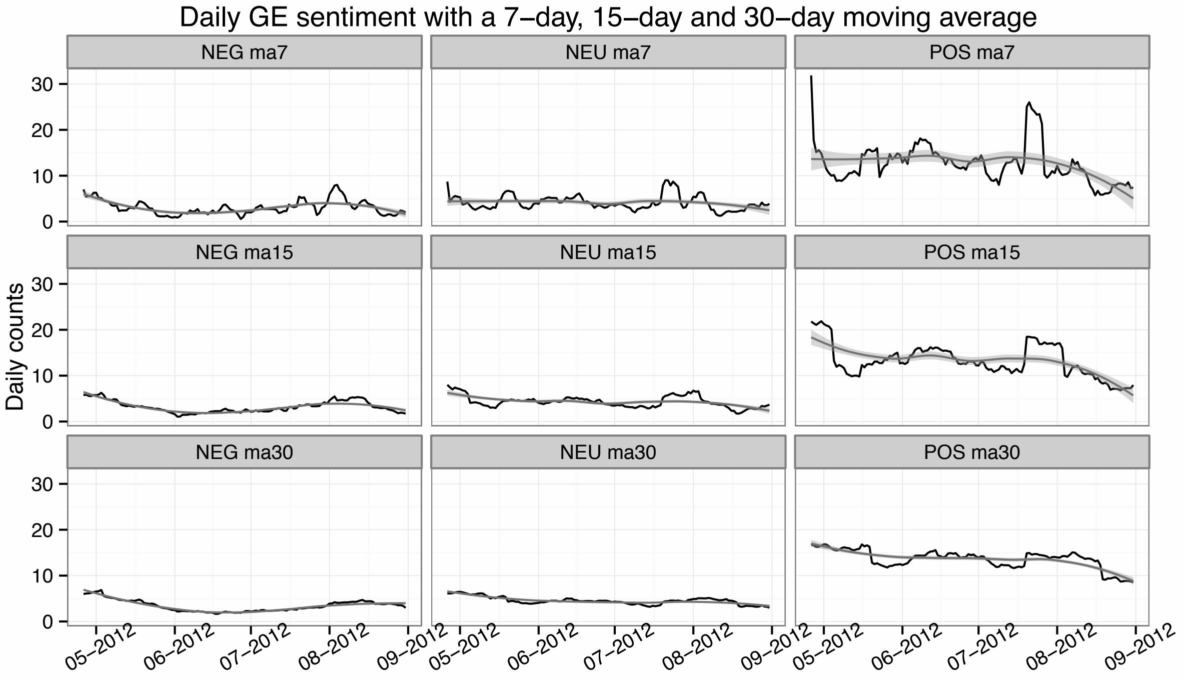
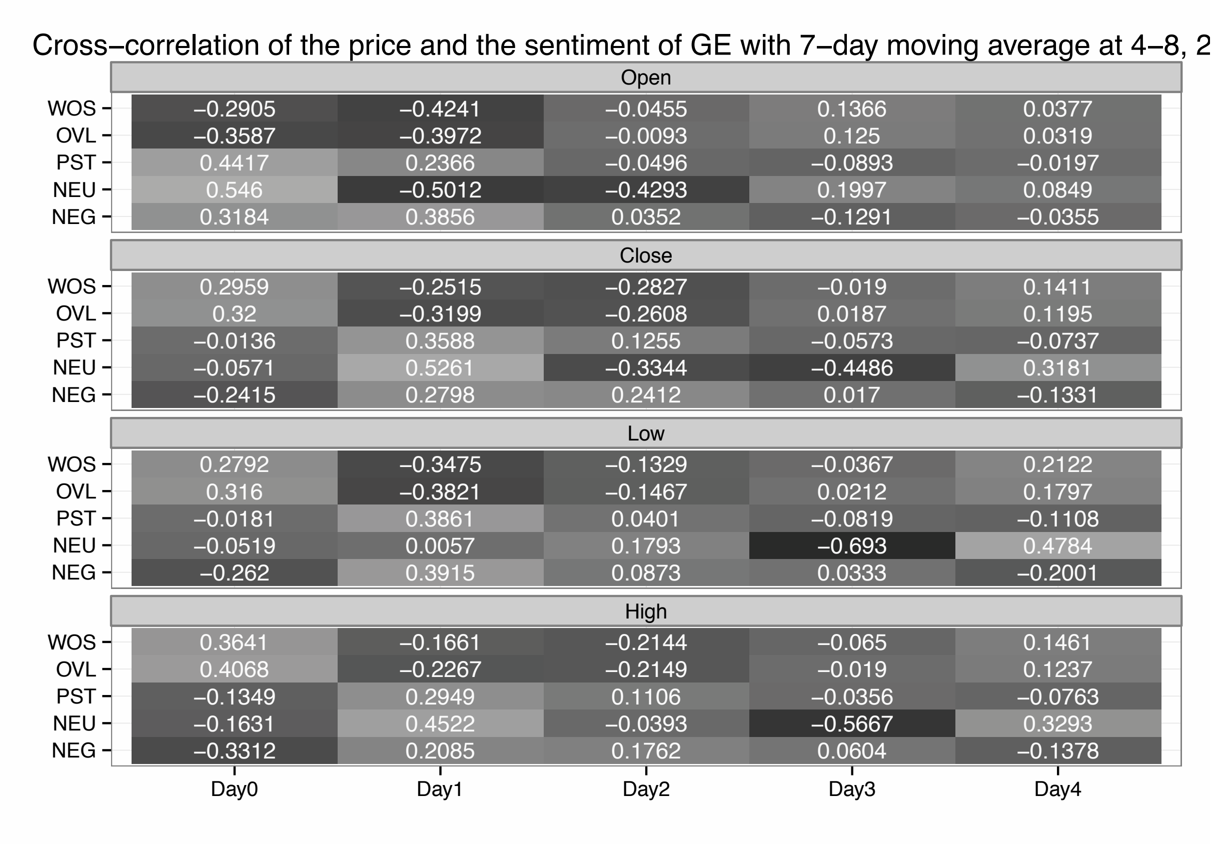


Figure 5.4.6 Daily GE price trend with a 7-day moving average

In addition, the CCF test showed some different phenomena. As Figure 5.4.7 illustrates, for the 7-day moving average between the NEU sentiment and open price, there was a strong positive correlation at Day$_0$ only (r = .546), and a strong negative correlation at the Day$_1$ lag (r = -.5012). Apart from these, at the Day$_3$ lag, the NEU sentiment had a strong negative correlation between the low and high price respectively (r = -.693 and r = -.5667 respectively). Compared the CCF test result of the 7-day moving average manipulation with the baseline, there were more positive correlations, but most correlations became weaker than the baseline. The OVL and WOS sentiments in this manipulation did not outperform the OVL sentiment in the baseline.



Cross–correlation of the price and the sentiment of GE with 7–day moving average at 4–8, 2

Open

	Day0	Day1	Day2	Day3	Day4
WOS	–0.2905	–0.4241	–0.0455	0.1366	0.0377
OVL	–0.3587	–0.3972	–0.0093	0.125	0.0319
PST	0.4417	0.2366	–0.0496	–0.0893	–0.0197
NEU	0.546	–0.5012	–0.4293	0.1997	0.0849
NEG	0.3184	0.3856	0.0352	–0.1291	–0.0355

Close

	Day0	Day1	Day2	Day3	Day4
WOS	0.2959	–0.2515	–0.2827	–0.019	0.1411
OVL	0.32	–0.3199	–0.2608	0.0187	0.1195
PST	–0.0136	0.3588	0.1255	–0.0573	–0.0737
NEU	–0.0571	0.5261	–0.3344	–0.4486	0.3181
NEG	–0.2415	0.2798	0.2412	0.017	–0.1331

Low

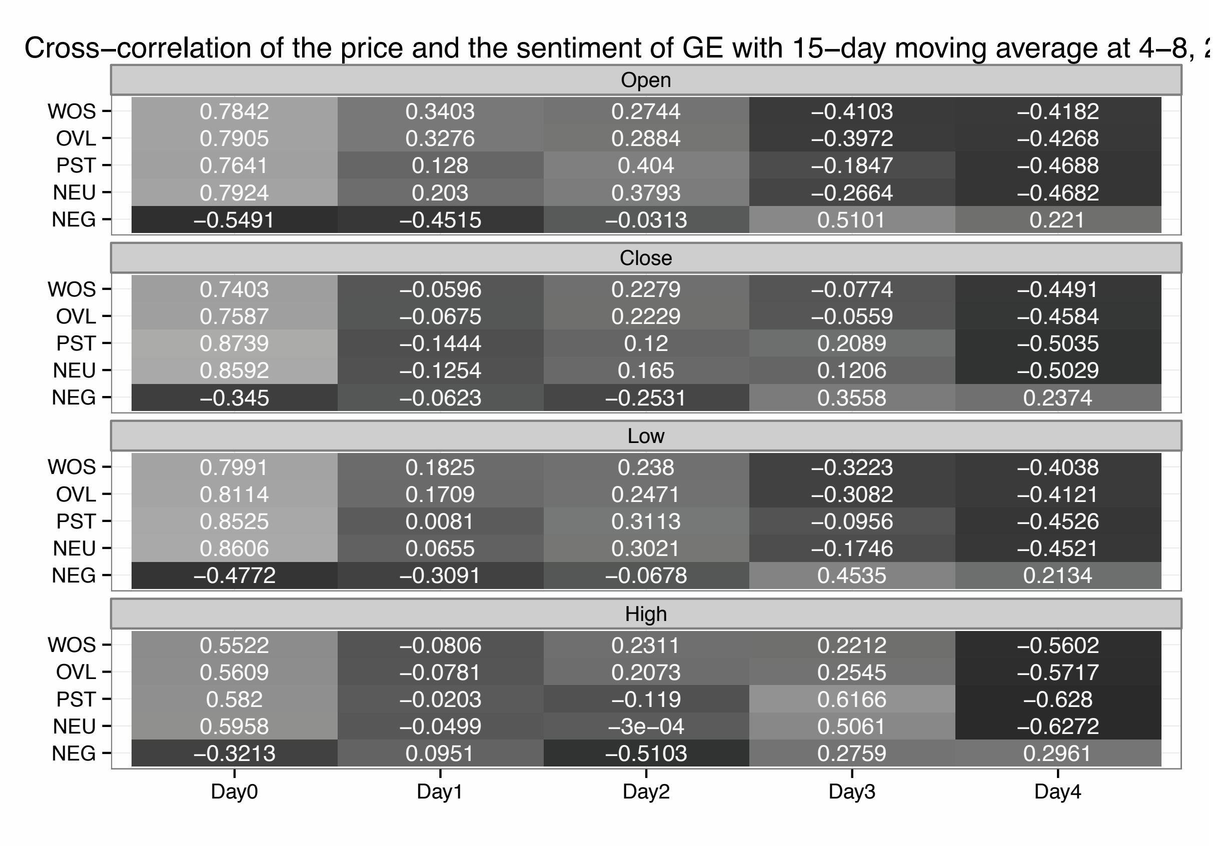
	Day0	Day1	Day2	Day3	Day4
WOS	0.2792	–0.3475	–0.1329	–0.0367	0.2122
OVL	0.316	–0.3821	–0.1467	0.0212	0.1797
PST	–0.0181	0.3861	0.0401	–0.0819	–0.1108
NEU	–0.0519	0.0057	0.1793	–0.693	0.4784
NEG	–0.262	0.3915	0.0873	0.0333	–0.2001

High

	Day0	Day1	Day2	Day3	Day4
WOS	0.3641	–0.1661	–0.2144	–0.065	0.1461
OVL	0.4068	–0.2267	–0.2149	–0.019	0.1237
PST	–0.1349	0.2949	0.1106	–0.0356	–0.0763
NEU	–0.1631	0.4522	–0.0393	–0.5667	0.3293
NEG	–0.3312	0.2085	0.1762	0.0604	–0.1378

Figure 5.4.7 Cross-correlation result of GE sentiment and price with a 7-day moving average

For the 15-day moving average (see Figure 5.4.8), there are many more strong correlations at the Day$_0$ lag: Except for the NEG sentiment, all the other four sentiment trends had strong correlations with the open, close and low price (with Pearson’s_r_ values of around 0.8). Specifically, the NEU sentiment had strong positive correlations with the low and close price (*r* = .8606 and *r* = .8592 respectively), and a positive correlation with the open price (*r* = .7924). Similarly, the PST sentiment had similar results: strong positive correlations with low (*r* = .8525) and close price (*r* = .8739), and a positive correlation with the open price (*r* = .7941). The OVL and WOS sentiments had very similar results: Both of them had correlations with the low (*r* = .8114 and *r* = .7991), close (*r* = .7587 and *r* = .7403) and open price (*r* = .7905 and *r* = .7842).



Open

	Day0	Day1	Day2	Day3	Day4
WOS	0.7842	0.3403	0.2744	−0.4103	−0.4182
OVL	0.7905	0.3276	0.2884	−0.3972	−0.4268
PST	0.7641	0.128	0.404	−0.1847	−0.4688
NEU	0.7924	0.203	0.3793	−0.2664	−0.4682
NEG	−0.5491	−0.4515	−0.0313	0.5101	0.221

Close

	Day0	Day1	Day2	Day3	Day4
WOS	0.7403	−0.0596	0.2279	−0.0774	−0.4491
OVL	0.7587	−0.0675	0.2229	−0.0559	−0.4584
PST	0.8739	−0.1444	0.12	0.2089	−0.5035
NEU	0.8592	−0.1254	0.165	0.1206	−0.5029
NEG	−0.345	−0.0623	−0.2531	0.3558	0.2374

Low

	Day0	Day1	Day2	Day3	Day4
WOS	0.7991	0.1825	0.238	−0.3223	−0.4038
OVL	0.8114	0.1709	0.2471	−0.3082	−0.4121
PST	0.8525	0.0081	0.3113	−0.0956	−0.4526
NEU	0.8606	0.0655	0.3021	−0.1746	−0.4521
NEG	−0.4772	−0.3091	−0.0678	0.4535	0.2134

High

	Day0	Day1	Day2	Day3	Day4
WOS	0.5522	−0.0806	0.2311	0.2212	−0.5602
OVL	0.5609	−0.0781	0.2073	0.2545	−0.5717
PST	0.582	−0.0203	−0.119	0.6166	−0.628
NEU	0.5958	−0.0499	−3e−04	0.5061	−0.6272
NEG	−0.3213	0.0951	−0.5103	0.2759	0.2961

Figure 5.4.8 Cross-correlation result of GE sentiment and price with a 15-day moving average

In addition, the NEG sentiment had a strong positive correlation with the Day$_3$ open price, and the Day$_3$ high price had strong positive correlations with the POS and NEU sentiments respectively (*r* = .6166 and *r* = .5061). Also, the Day$_4$ close price had strong negative correlations with the POS and NEG sentiments respectively (r = -.5035 and r = -.5029). Finally, the Day_4 high price had strong negative correlations with the WOS, OVL, POS, and NEU prices (r = -.5602, r = -.5717, r = -.628 and r = -.6272 respectively).

There were four extremely high correlations for the Day_0 tag in the CCF results for 30-day moving average (see Figure 5.4.9): the positive correlations between the NEG sentiment and the low price (r = .9907) and between the PST sentiment and the high price (r = .9876), as well as the negative correlations between the OVL sentiment, WOS sentiment and low price (r = -.9772 and r = -.9750 respectively). Apart from these, there were also a number of strong correlations for the Day_0 lag. Except for the correlations between the open price and the POS and NEU sentiments, all the correlations between price and sentiment for the Day0 lag were strong.

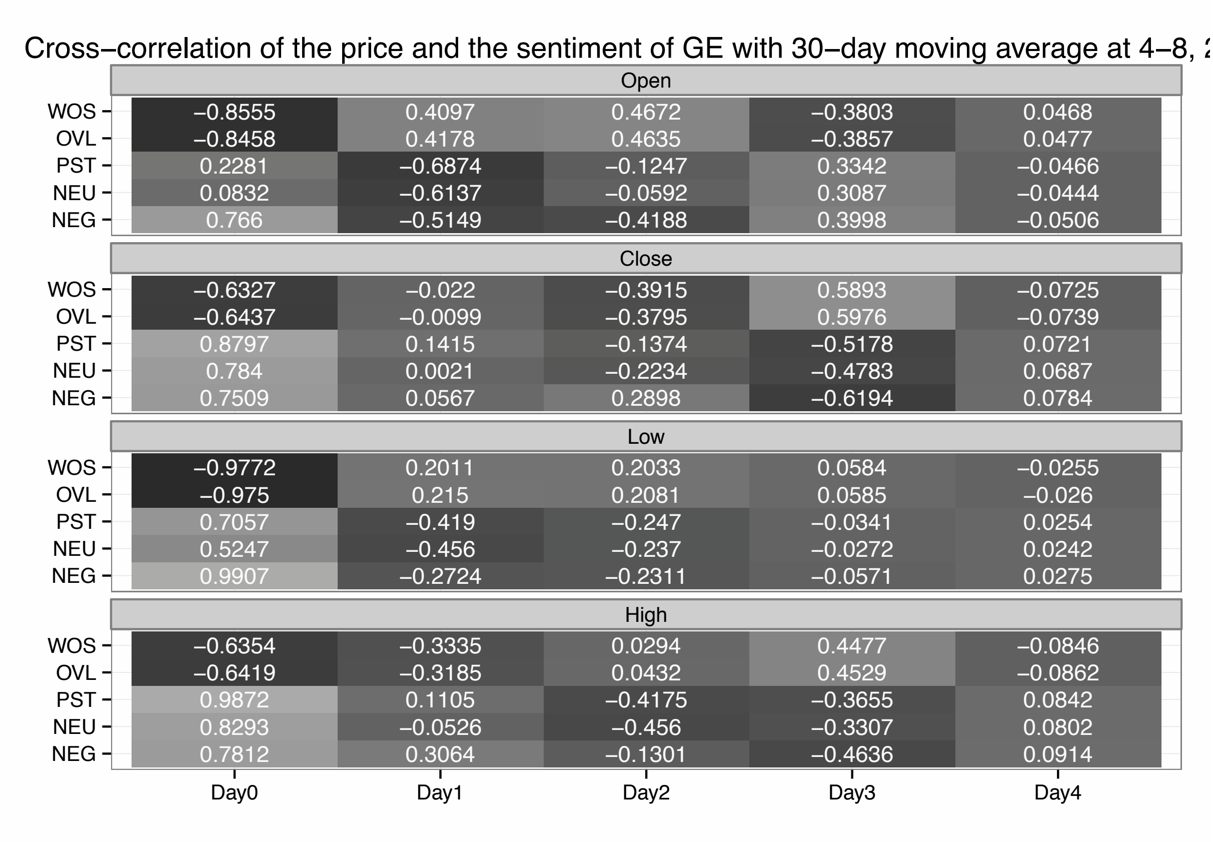
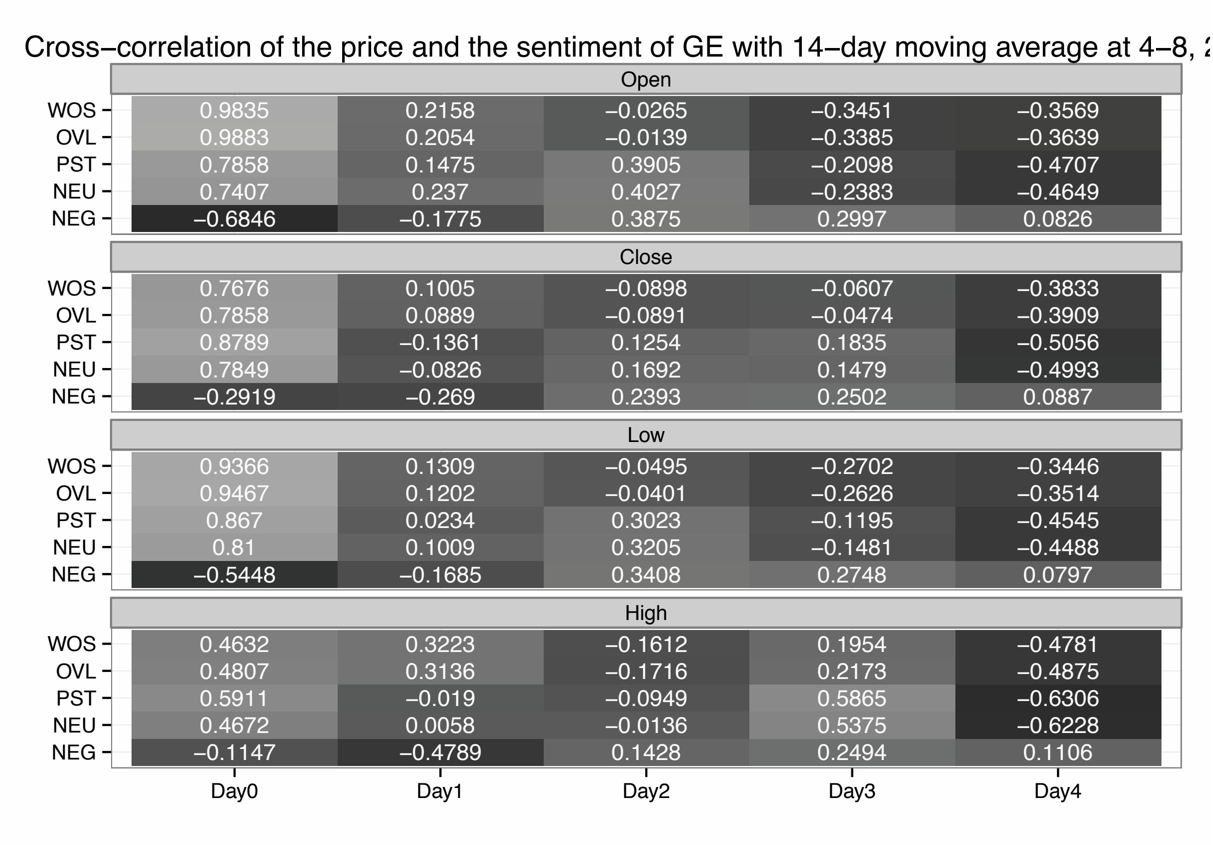


Figure 5.4.9 Cross-correlation result of GE sentiment and price with a 30-day moving average

Furthermore, the Day_1 open price had strong negative correlations with the POS, NEU, and NEG prices respectively (r = -.6874, r = -.6137 and r = -.5149). Also, the Day_3 close price had strong positive correlation with the WOS and OVL sentiments (r = .5893 and r = .5976 respectively), and strong negative correlations with the POS and NEG sentiments (r = -.5178 and r = -.6194 respectively).

The CCF test result showed that most correlations in the 15-day moving average manipulation become stronger than the baseline. In particular, for the Day_0 lag, all three individual sentiments turned much more positive than those in the baseline, but the OVL sentiment had an opposite trend.

Comparing these three results, it was clear that the 15-day moving average had the most strong positive correlations. Thus, it is worth going back to the data, and running the analysis again, this time testing out a 14-day moving average, to see if that performance was even better still (see Figure 5.4.10).



Cross–correlation of the price and the sentiment of GE with 14–day moving average at 4–8, 2

Open

	Day0	Day1	Day2	Day3	Day4
WOS	0.9835	0.2158	−0.0265	−0.3451	−0.3569
OVL	0.9883	0.2054	−0.0139	−0.3385	−0.3639
PST	0.7858	0.1475	0.3905	−0.2098	−0.4707
NEU	0.7407	0.237	0.4027	−0.2383	−0.4649
NEG	−0.6846	−0.1775	0.3875	0.2997	0.0826

Close

	Day0	Day1	Day2	Day3	Day4
WOS	0.7676	0.1005	−0.0898	−0.0607	−0.3833
OVL	0.7858	0.0889	−0.0891	−0.0474	−0.3909
PST	0.8789	−0.1361	0.1254	0.1835	−0.5056
NEU	0.7849	−0.0826	0.1692	0.1479	−0.4993
NEG	−0.2919	−0.269	0.2393	0.2502	0.0887

Low

	Day0	Day1	Day2	Day3	Day4
WOS	0.9366	0.1309	−0.0495	−0.2702	−0.3446
OVL	0.9467	0.1202	−0.0401	−0.2626	−0.3514
PST	0.867	0.0234	0.3023	−0.1195	−0.4545
NEU	0.81	0.1009	0.3205	−0.1481	−0.4488
NEG	−0.5448	−0.1685	0.3408	0.2748	0.0797

High

	Day0	Day1	Day2	Day3	Day4
WOS	0.4632	0.3223	−0.1612	0.1954	−0.4781
OVL	0.4807	0.3136	−0.1716	0.2173	−0.4875
PST	0.5911	−0.019	−0.0949	0.5865	−0.6306
NEU	0.4672	0.0058	−0.0136	0.5375	−0.6228
NEG	−0.1147	−0.4789	0.1428	0.2494	0.1106

Figure 5.4.10 Cross-correlation result of GE sentiment and price with a 14-day moving average

At the Day_0 lag, both the WOS and OVL sentiments had extremely strong positive correlations with the open price (*r* = .9835 and *r* = .9883 respectively); in addition, these two sentiments have very strong correlations with low price (*r* = .9366 and *r* = .9467 respectively). Besides, there are strong positive correlations between open, close and low price with the NEG, NEU, and PST sentiments, and the value of *r* ranges from 0.6 to 0.9.

In addition to these, two strong positive correlations could be found between the Day_3 high price and the POS and NEU sentiments (*r* = .5865 and *r* = .5375 respectively), and two negative correlations between the Day_4 high price with the POS and NEU sentiments (-.6306 and *r* = -.6228 respectively).

Compared with the baseline, the correlations for the Day_0 lag in the CCF test result of the 14-day moving average manipulation had a completely different trend: With the exception of the NEG sentiment, the other four sentiments turned to much more positive than the baseline. However, for the other time lags, most correlations became weaker than the baseline.

Comparing different moving average time lags, the following points can be concluded. To start, the strongest correlations often occur for the corresponding day lag (Day_0). Then, the correlation strength becomes weaker in the following time lags. Also, there are four extremely high correlations in the 14-day and 30-day moving average results for the corresponding day lag, but the 14-day moving average has more positive pairs. Next, using moving average to smoothen data can significantly improve the correlation between GE price trends and GE sentiment trends. Finally, at the Day_0 lag, the negative sentiment often showed an opposite trend to the other sentiment trends.

5.4.4 Discussion

It is clear from the analyses reported in this section that percentage normalisation weakened the correlation within the GE tweet data due to the changes in the dependence between individual sentiments, meaning they had weaker correlations with the price than the baseline, including the NEG, NEU, and PST sentiments. Also, this transformation showed no strong correlation between negative and positive sentiments. Adding the NEG, NEU, and PST sentiment trends together shows an overall positive trend, and the weighted overall sentiment showed improvement. The overall sentiment and the weighted overall sentiments showed stronger correlations with the price trends than any individual sentiment trends, and the strongest correlation occurred between the OVL sentiment and the Day_3 low price, r = .9311.

Using the window function significantly improved the correlation within the price and sentiment data respectively, but not between these two groups of data, compared to the baseline. Thus, it was not helpful to identify the temporal correlation between stock price and sentiment data.

By applying different moving average manipulations showed a clear tendency to increase, and the correlation test has different results. As addressed by O'Connor et al. (2010), using more aggregated window does not necessarily improves the overall performance of cross-correlation: In this case study, the 14-day moving average showed the best result at the same lag. Most strong correlations occur at the Day_0 lag: These indicate that there were certain correlations between GE price trends and sentiment trends, but they were not very helpful for predicting price developments.

Comparing the cross-correlation results of overall sentiment and weighted overall sentiment with percentage normalisation, the window function, and moving average, the overall sentiment was found to slightly outperform the weighted overall sentiment in most cases.

5.5 Discussion of Neutral Sentiment

As argued in Chapter 3, some previous sentiment analysis projects put unclassified or irrelevant content into a neutral category. However, this research used a hierarchy sentiment classification approach instead of classifying unidentified tweets as neutral tweets as shown in Chapter 4. In order to understand which classification method is more effective, this section reports on the comparison of the time series correlations of these two sentiment classifications with price changes.

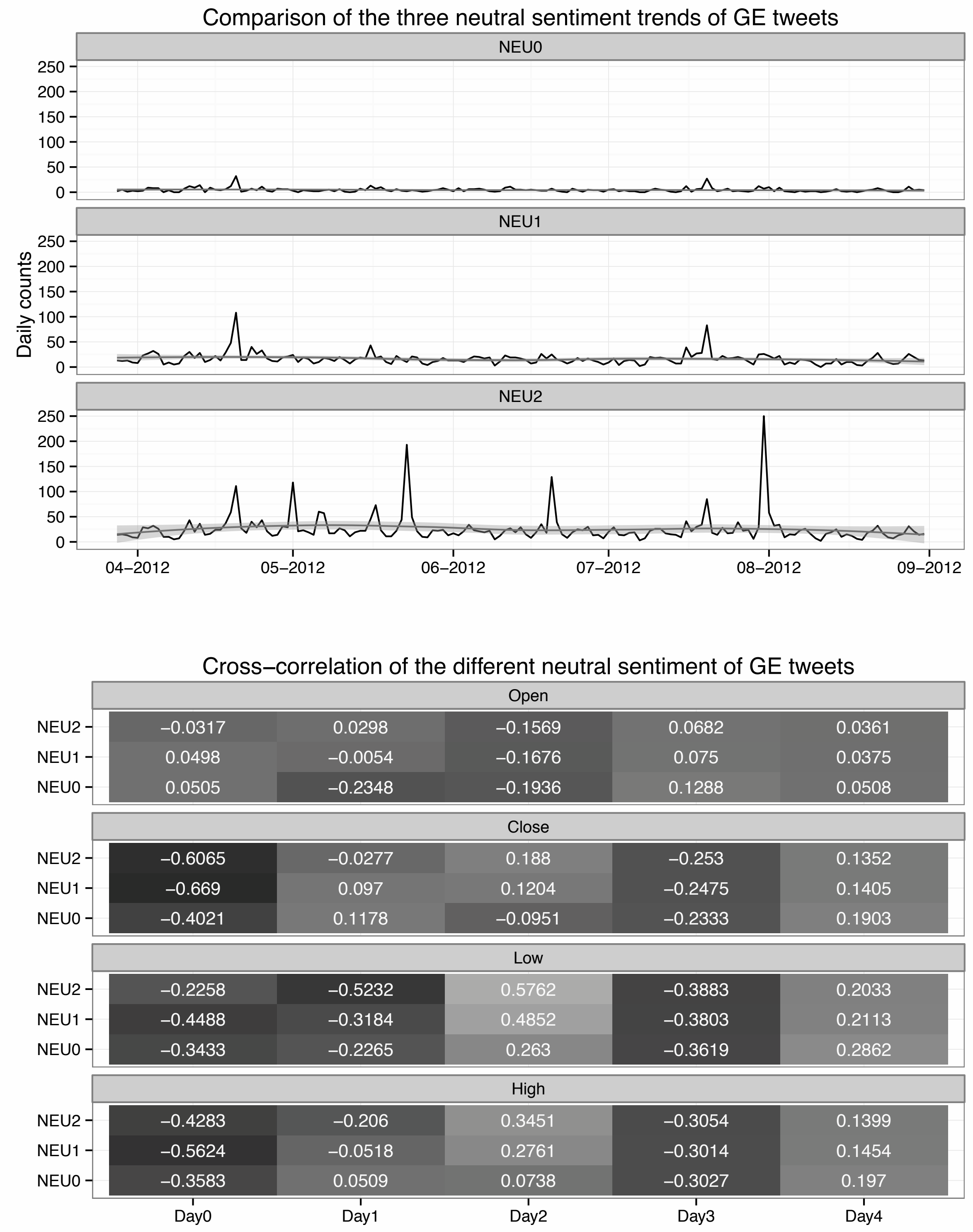
5.5.1 Three neutral sentiments

The comparison is among three groups of neutral sentiment:

1. NEU_0. The NEU group;
2. NEU_1. The NEU and NTR tweets, which are conventionally considered as neutral tweets in ticker-related tweets;
3. NEU_2. The NEU, NTR and NSR tweets, which are conventionally considered as neutral.

The overall trends of these three sentiments of the GE ticker are very different (see the upper part of Figure 5.5.1). The NEU_0 trend is relatively smooth, whereas the NEU_2 trend is the most dynamic. However, in the middle of April and July, there are two corresponding peaks among all three trends.

Applying the CCF test to the three different neutral trends with price trends, most correlations were weak (see the lower part of Figure 5.5.1). However, there were only a few strong correlations: for the Day_0 lag, the close price had strong negative correlations with the NEU_1 and NEU_2 trends, and the high price had strong negative correlations with the NEU_1 trend. Moreover, the NEU_2 trends strongly correlated with the Day_2 low price. This showed three neutral trends had



Open

	Day0	Day1	Day2	Day3	Day4
NEU2	–0.0317	0.0298	–0.1569	0.0682	0.0361
NEU1	0.0498	–0.0054	–0.1676	0.075	0.0375
NEU0	0.0505	–0.2348	–0.1936	0.1288	0.0508

Close

	Day0	Day1	Day2	Day3	Day4
NEU2	–0.6065	–0.0277	0.188	–0.253	0.1352
NEU1	–0.669	0.097	0.1204	–0.2475	0.1405
NEU0	–0.4021	0.1178	–0.0951	–0.2333	0.1903

Low

	Day0	Day1	Day2	Day3	Day4
NEU2	–0.2258	–0.5232	0.5762	–0.3883	0.2033
NEU1	–0.4488	–0.3184	0.4852	–0.3803	0.2113
NEU0	–0.3433	–0.2265	0.263	–0.3619	0.2862

High

	Day0	Day1	Day2	Day3	Day4
NEU2	–0.4283	–0.206	0.3451	–0.3054	0.1399
NEU1	–0.5624	–0.0518	0.2761	–0.3014	0.1454
NEU0	–0.3583	0.0509	0.0738	–0.3027	0.197

Figure 5.5.1 Comparison of different neutral sentiment trends of GE

different movements, meaning that they would have different influence on the later sentiment classification.

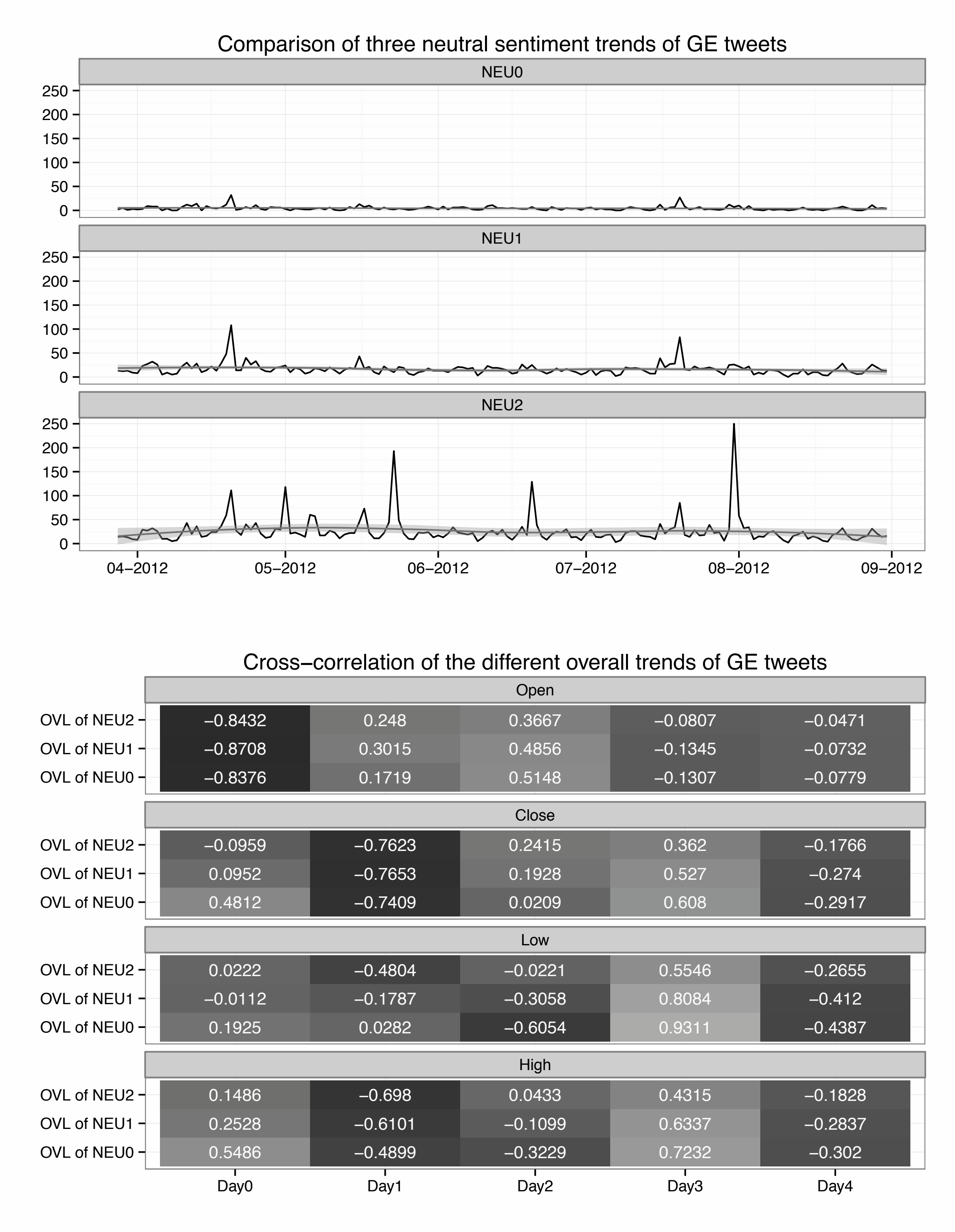
5.5.2 Overall trends of the different neutral sentiments

The following analysis computes the overall trends as below:

1. Overall trend based on NEU_0 = (PST - NEG)/(PST + NEU_0 + NEG), where NEU_0 = NEU
2. Overall trend based on NEU_1 = (PST - NEG)/(PST + NEU_1 + NEG), where NEU_1 = NTR + NEU
3. Overall trend based on NEU_2 = (PST - NEG)/(PST + NEU_2 + NEG), where NEU_2 = NSR + NTR + NEU

The upper part of Figure 5.5.2 presents the similarity of these three overall sentiment trends, and generally, they have similar movements. However, these three trends differ in terms of magnitude: the overall trend based on the NEU_0 is the most dynamic trend, while the overall trend based on the NEU_2 trend is the least dynamic. Considering the formulae above, the denominator of the overall trend based on NEU_0 is the smallest, and the overall trend based on NEU_2 is the biggest, so this condenses the changes of the overall trend based on NEU_1 and NEU_2.

Applying a CCF test to the different overall trends and the price trends, the overall trends based on NEU_1 and NEU_2 had similar performances with the overall trend based on NEU_0 in general (see the lower part of Figure 5.5.2). The most strong positive correlations cluster at the Day_3 lag, in particular, between overall trend based on the NEU_0 and the Day_3 low price, the *r* = .9311. The overall trend based on the NEU_0 slightly outperformed the other two trends. At the same time lag, the close and high prices also had strong positive correlations with these three overall sentiment trends, and the overall trend based on NEU_0 still outperformed the other two.



	Day0	Day1	Day2	Day3	Day4
Open					
OVL of NEU2	−0.8432	0.248	0.3667	−0.0807	−0.0471
OVL of NEU1	−0.8708	0.3015	0.4856	−0.1345	−0.0732
OVL of NEU0	−0.8376	0.1719	0.5148	−0.1307	−0.0779
Close					
OVL of NEU2	−0.0959	−0.7623	0.2415	0.362	−0.1766
OVL of NEU1	0.0952	−0.7653	0.1928	0.527	−0.274
OVL of NEU0	0.4812	−0.7409	0.0209	0.608	−0.2917
Low					
OVL of NEU2	0.0222	−0.4804	−0.0221	0.5546	−0.2655
OVL of NEU1	−0.0112	−0.1787	−0.3058	0.8084	−0.412
OVL of NEU0	0.1925	0.0282	−0.6054	0.9311	−0.4387
High					
OVL of NEU2	0.1486	−0.698	0.0433	0.4315	−0.1828
OVL of NEU1	0.2528	−0.6101	−0.1099	0.6337	−0.2837
OVL of NEU0	0.5486	−0.4899	−0.3229	0.7232	−0.302

Figure 5.5.2 Comparison of different neutral sentiment trends of GE in percentage scale

For Day_0, the strong correlations existed in the open, close, and high prices. The correlations between the open price and the three overall sentiments were negative, with the overall sentiment based on NEU_0 being the strongest (r = -.8376). Between the close and high price with the three overall sentiments, the overall sentiment based on NEU_0 had the strongest correlations (r = .4812 and r = .4812 and r = .5486). In addition, there were two groups of strong negative correlations between three overall sentiments and the Day_1 close price and the Day_1 low price. The correlation between the overall sentiment based on the NEU_0 and the Day_1 close price was the strongest negative correlation.

These suggest that, in the annotated dataset, the conventional sentiment classification outperformed in the Day_0 lag, but in the future time lags, the hierarchy sentiment classification developed in this research could improve the positive correlation strength by about 1% to 4%.

5.6 Summary

The chapter analysed the temporal correlation within the annotated tweet sentiment, within different types of stock prices, and between the tweet sentiment and corresponding market prices. The cross-correlation test on the tweet sentiment suggests that the NSR and NTR tweet sentiments are temporally different from the three ticker-correlated sentiments. Using three temporal data manipulation methods – percentage normalisation, window function transformation, and moving average smoothing – some strong temporal correlations could be found between the stock price and corresponding tweet sentiments. Most strong correlations occurred at Day_0, which suggests that the prediction window is short. Apart from the correlations for the Day_0, the strongest correlations are found in the CCF test based on the percentage normalisation, which had a positive correlation between the overall sentiment and Day_3 low price, where r = .9311. Moreover, using ticker-related tweets from the hierarchy classification category designed in the second case study to correlate the market price only achieved a slightly higher improvement than the conventional approach. Overall, there were several strong correlations between the stock price and tweet sentiment data. In short, this chapter demonstrated that tweet sentiment analysis is able to predict stock price changes.

Chapter 6 Methodology for Analysis and Classification

While analysis of the temporal correlations of the annotated tweets described in Chapter 5 revealed temporal distinctions between stock tweets and non-stock-related tweets, the linguistic differences between these two types of tweets are still unclear. Therefore, the main analyses reported in Chapters 7, 8 and 9 focus on the linguistic features of stock tweets. Although this research makes substantial use of various approaches in computational linguistics, it still focuses on the **linguistic** features of tweets, and as such can still be seen fundamentally as a piece of applied linguistic research. The analysis reported in the following three chapters required methods from three different areas to investigate the annotated tweet data, including corpus analysis, statistical analysis and machine learning, so this chapter provides a detailed introduction of those methods being used later.

6.1 A Linguistic Approach to Tweet Analysis

This section describes how corpus linguistics techniques were used to to analyse a collection of annotated stock tweets, including frequency analysis, concordance analysis and keyness analysis. Each of these methods are described in turn. First, it may be useful to provide a general introduction to corpus linguistics as a research methodology.

6.1.1 Introduction to corpus linguistics

According to McEnery and Hardie (2011), corpus linguistics can be considered as 'dealing with some set of machine-readable texts which is deemed an appropriate basis on which to study a specific set of research questions' (p. 1). Thus, corpus linguistics can be regarded as a methodology for exploring linguistic questions quantitatively and qualitatively based on textual data. Their definition also suggests that corpus linguistics has two fundamental requirements in terms of data:

that it be both sufficient in size, and machine-readable in format. The raw tweet data collected for the current research qualify on both counts: They are massive, and are presented in a well-structured format. With proper manipulations, a tweet corpus can thus be compiled.

As McEnery and Hardie (2011) pointed out, the corpus approach can cover 'a subset of all the research questions that a linguist might ask', and this subset 'overlap[s] with the subset of questions that a linguist can ask without a corpus, but it is almost certainly greater in size than that set' (p. 27). Tweet data are a relatively new domain to linguists, and undoubtedly there are many unanswered areas yet to be explored. In this sense, applying the corpus approach could extend this possibility much further.

One such avenue is suggested by Hunston (2002), who pointed out that 'corpora can be used to establish norms of frequency and usage against which individual texts can be measured' (p. 14). While Hunston's proposal was formulated in relation to stylistics and clinical or forensic linguistics, it also seems relevant to the study of tweets (It should perhaps be noted that Hunston proposed this idea at a time when social media did not exist at all). In any case, the present research aims to provide a benchmark in the form of a set of statements about the frequency and regularity of a range of linguistic features of tweets.

6.1.2 Frequency analysis

A frequency list is 'a list of all the types in corpus together with the number of occurrences of each type' (Hunston, 2002, p. 67). With a frequency list, a number of features of a corpus can be investigated. For example, Zappavigna (2012) compared the frequency list of the HERMES corpus (see Chapter 1) with that of the COCA corpus as a reference corpus (see Chapter 1), and she found out that the @ mention is the most frequent pattern in the tweet data, which does not exist in the top frequency list of the COCA corpus. Moreover, Page (2012) used the HERMES corpus as a reference corpus for the keyness analysis in her study of tweets, and she found that different groups of Twitter users have different tweeting behaviours in terms of hashtag usage. Thus, it is worth investigating whether the stock tweets present different tweeting behaviours. Frequency lists can comprise a list of single words (a unigram list), a list of adjacent word pairs (a bigram list), or a list of three adjacent words (a trigram list), etc. This research mainly used unigram and bigram frequency lists to explore whether it is possible to find frequently reoccurring structures that can provide a reliable indication of the stock market sentiment in tweets.

6.1.3 Concordance analysis

Concordances are 'a list of a word or phrase, with a few words of context either side' (Baker in Litosseliti, 2010, p. 106). Concordances are mainly used to qualitatively explore the corpus data because they make the identification of coocurring patterns easier. For example, Zappavigna (2011) used this approach to investigate tweets talking about President Obama's victory in the presidential election and found a 'general prosody of positive evaluation' (p. 11) around the word *won*.

Although the concordance is arguably the single most important tool in corpus analysis (McEnery & Hardie, 2011), it has one obvious shortcoming: A frequent pattern may bring too much data for manual analysis, particularly when the corpus under analysis is relatively big. One possible solution to this problem is to randomly sample the data. This is the solution applied by Zappavigna (2012), whose concordance analysis was based on a subset of 100 annotated tweets that she had drawn randomly from the HERMES corpus, which comprised 7 million tweets in total. On the other hand, another problem is that each tweet is short, and might contain a number of irregular spellings or instances of informal language usage. Therefore, if the corpus is relatively small, it is hard to generate sufficient concordance lines for analysis.

McEnery and Hardie (2011) suggested that ‘concordances and frequency data exemplify respectively the two forms of analysis, namely qualitative and quantitative’ (p. 2). These two methods are the most fundamental methods being used in corpus analysis, and in the analysis undertaken for this research, they were used to analyse the keywords in different polarities, in order to develop an automatic classifier for identifying the polarity of stock tweets.

6.1.4 Keyness analysis

Scott and Tribble (2006) defined keyness as ‘a quality words may have in a given text or set of texts, suggesting that they are important, they reflect what the text is really about, avoiding trivia and insignificant detail’ (pp. 55-56). Keyness analysis works by identifying words that have a marked association with the target corpus when compared to a reference corpus representing the language or language variety in general.

AntConc (Anthony, 2004) provides a convenient tool for identifying keywords in a specialised corpus. For the analyses outlined in the following chapters, *AntConc 3.3.5 for Mac* was used to calculate the keyness. Usually, there are two statistical methods to calculate keyness, namely the log-likelihood and chi-square test, and *AntConc* provides both options. The present analysis chooses the log-likelihood test because it is ‘better ’in general’ than the Chi-square test’ (Rayson, 2003, p. 152), and in particular because the corpus size does not affect the significance scores in the log-likelihood test, whereas the Chi-square test is notoriously vulnerable to this problem (Dunning, 1993).

6.1.5 Choice of using keyness analysis

As it will be shown later, due to the unbalanced size of the two polarity word lists and the low degree of overlapping of the two lists, the aforementioned frequency list analysis of unigrams and bigrams failed to identify the keywords in different polarity lists. Therefore, a more robust approach should be considered, and keyness analysis fulfills this need in a great deal.

In principle, unlike frequency analysis, keyness analysis ignores the different sizes of two words lists, because it uses the rank of the words, instead of the frequency, to calculate the degree of relevant importance of words in a list compared to a reference list. On the other hand, ngram analysis and concordance analysis are mainly conducted manually, and they are not able to indicate the importance of any specific words regarding its frequency or rank.

In comparison with the above two methods, keyness analysis can offer sufficient information of the importance of specific keywords even when they do not cooccur in two words at the same time. Therefore, using keyness analysis can improve the identification of the relevant importance of specific words regardless of the different size of two word lists.

Considering the above reasons, the later analysis decided to use keyness analysis, instead of other choices, to analyse the tweet corpus.

6.1.6 Tweet normalisation

Non-standard language usages are the main obstacle for any analysis of tweets (A Kumar & Sebastian, 2012). As will be shown later, this is particularly problematic for a small-size tweet corpus, because the non-standard language brings more word types, which in turn reduces the level of word density. Consequently, any analysis based on frequency will become more difficult.

To understand how frequent the non-standard language occurred in the sample corpus, VARD 2.5.4 by Baron (2014) was used to inspect these spelling variants. VARD has a preset of tweets, but still, it does not fit the sample data very well. In the first experiment, 27.86% of all tokens in the sample corpus are not standard. However, a manual inspection showed that some highly-frequent tokens were also mis-classified as non-standard tokens, including *ge*, *rt*, *url* and a number of ticker names. After removing these mis-classified non-standard tokens, only 12.13% of all tokens were regarded as non-standard tokens. Although this frequency of spelling variants seems low, it would still cause difficulties to a subsequent analysis, because the first method to be analysed, the 'bag of words' approach, is supposed to be primarily based on the analysis of word types. The incidence of spelling variants will increase the sparseness of the corpus, which, in turn, will inevitably affect the results of the 'bag of words' approach.

Therefore, to solve this problem, one possible solution is to normalise non-standard language usages in tweets to their standard forms. This method is derived from a similar approach being used in short message processing, such as Aw, Zhang, Xiao and Su (2006) and Han and Baldwin (2011). In recent years, several papers have discussed tweet normalisation (Porta & Sancho, 2013; Xue, Yin, Davison, & Davison, 2011), but few of them have provided a freely available solution. K. Bontcheva et al. (2013)'s solution seems the only available open-sourced one to this problem, but as a plug-in for GATE (General Architecture for Text Engineering) by Cunningham, Tablan, Roberts and Bontcheva (2013), it was neither available in the stable version of GATE, nor available as a stand-alone script by the time when the analysis was conducted. Given the limited time and limited resources available for this research, a decision was made against normalising the tweet data for this research, instead, the analysis was to use the stemming method as mentioned below.

6.1.7 Stemming

Stemming is a widely used technique in computational linguistics and information retrieval for removing suffixes from English words (Porter, 1980). According to Porter (1980), conflating various words with the same root to a single word can improve the performance of information retrieval, and this technique has been applied in sentiment analysis as well, where it helps to 'reduce the vocabulary size, thereby sharpening one's results' (Potts, 2011). The Porter Stemmer (Porter, 1980) and the Lancaster Stemmer (Paice, 1990) are two popular stemmers used in sentiment analysis.

According to Potts (2011), both stemmers are problematic in that they 'destroy too many sentiment distinctions' (1 Overview, para. 3), but the Lancaster Stemmer is 'even more problematic than the Porter Stemmer' (3 Lancaster stemmer, para. 1). Therefore, the analysis conducted in this research used the Porter Stemmer to conflate the various word types in the annotated stock tweet dataset.

As Porter (1980) explained, the Porter Stemmer is 'given an explicit list of suffixes, and, with each suffix, the criterion under which it may be removed from a word to leave a valid stem' (p. 131). In other words, it removes the suffix of a word according to a suffix list. Sample 6.1.2 is the stemmed form of the tweet in Sample 6.1.1. The word *apple* has been conflated to *app*, *heavy* to *heavi*, and *wednesday* to *wednesdi*.

> Sample 6.1.1 apple inc, dollar general, micron tech, ge attract heavy trading demand wednesday aapl dg mu ge mkt

> Sample 6.1.2 appl inc, dollar general, micron tech, ge attract heavi trade demand wednesdai aapl dg mu ge mkt

The main purpose of this stemmer is to improve the performance of information retrieval, so it aims to balance to simplicity, speed, and accuracy, instead of pursuing the maximum linguistic accuracy as a single goal. For example, the words *surprise* and *surprises* are the singular and plural forms of the same word root *surprise*, but in the following stemming results (see Section 7.2) based on the Porter Stemmer, they are considered as two words. In the experiment of a vocabulary of 10,000 words, this stemmer reduces the total number of word types by one third (Porter, 1980). Porter (2006) provides the stemmer in various languages on his website (`http://tartarus.org/martin/PorterStemmer/`). For this research, an implementation in `awk` language of the Porter Stemmer was used to conflate the stock tweets.

6.1.8 Part-of-speech tagging

Pak and Paroubek (2010) suggested that using tree tags, a binary POS tagging system, can distinguish between objective and subjective tweets, for example, using different pronouns to address the authors themselves. They also consider POS tags as an important indicator of the polarity of tweets. Accordingly, the present research examined the possibility of using POS tags to improve the accuracy of sentiment classification. Currently, there are two POS taggers designed exclusively for tweet data, including the Carnegie Mellon ARK tweet POS tagger (Gimpel et al., 2011; Owoputi et al., 2013), and the Sheffield GATE Twitter POS tagger (Derczynski, Ritter, Clark, & Bontcheva, 2013).

The ARK tweet POS tagger is a reliable tagger with a best overall accuracy of 93.2% (Owoputi et al., 2013), while the GATE Twitter POS tagger has a slightly lower accuracy of 88.7%. Given that the ARK tweet POS tagger has a longer development history and a better classification accuracy, the analyses reported in later chapters uses it to create the annotation. The ARK tweet POS tagger can be freely downloaded from the web page of the ARK laboratory at Carnegie Mellon University (`http://www.ark.cs.cmu.edu/TweetNLP/`). It contains two parts: a Twitter tokenizer for tokenising tweet data and a POS tagger for assigning POS tags for tokenised data (Owoputi et al., 2013). Both are written in Java, so they can be used on multi-platforms.

6.1.9 Summary

Given the size of the stock tweet corpus analysed, the research mainly used the frequency analysis to investigate the stock tweet data. Specifically, it analysed the most and the least frequent items in different sentiment categories of the annotated stock tweets and looked at the collocates around these frequent lexical items. It then used the Porter Stemmer and ARK tweet POS tagger to process the data. After that, it used keyness analysis to identify the keywords in different polarities, attempting to find distinct sentiment keywords in each polarity in order to establish whether the mainstream sentiment analysis approach — the bag-of-words approach — can successfully identify the polarity of the annotated stock tweets.

6.2 Statistical Tests

Statistical tests are widely used in corpus linguistics as a means of understanding data. This research uses normality tests to explore the distribution of the tweet data and significance tests to see if different groups of tweet data significantly differ from each other, in order to support the later automatic classifications. This section, therefore, provides a technical overview of different statistical tests that were applied in this research.

6.2.1 Normality tests

Statistically speaking, to understand whether two or more groups of data are significantly different from each other, significance tests should be applied; however, significance tests are based on specific assumptions of how the data were distributed. Therefore, data with different types of distributions need significance tests that correspond to those distributions. For example, some assumptions concern the normality, so a normality test, also known as goodness-of-fit test, is needed to validate these assumptions. Normality is critical in statistical tests as Baayen (2008) pointed out that

> Since many statistical procedures assume that vectors are normally distributed, it is often necessary to ascertain whether a vector of values is indeed approximately normally distributed. (p. 76)

According to Conover (1999), normality tests are used to examine 'a random sample from some unknown distribution in order to test the null hypothesis that the unknown distribution function is in fact a known, specified function' (p. 344). Thus, two commonly used normality tests were chosen to inspect whether the data are normally distributed.

The first of these, the Shapiro-Wilk test is a common approach to understand the normality of data. In R, the Shapiro-Wilk test can be performed with the `shapiro.test` function, which is based on the AS R94 algorithm by Royston (1995). This algorithm defines the range of the sample size from 3 to 5000 (Royston, 1995). The Shapiro-Wilk uses the p-value to indicate the significance levels, but unlike standard significance tests, if the p-value ≤ .05, the Shapiro-Wilk test rejects the null hypothesis that the data are normally distributed. In brief, if the p-value ≤ .05, then the data are not normally distributed.

Although the Shapiro-Wilk test is the most powerful normality test (Razali & Wah, 2011), it has a major drawback in that it only fits data of a small size. If the data size is larger than 5000, the Kolmogorov-Smirnov test, as an alternative, can be applied. The Kolmogorov-Smirnov test uses 'the largest distance between the two graphs $S(x)$, and $F(x)$, measured in a vertical direction' (Conover, 1999, p. 345). In R, the Kolmogorov-Smirnov test can be done with the `ks.test` function. Similar to the Shapiro-Wilk test, the Kolmogorov-Smirnov test also uses the p-value ≤ .05 to reject the normality of the data.

6.2.2 Significance tests

As shown in the analyses later, all data used in this thesis are not normally-distributed. Thus, given the nature of the non-normally-distributed data, two non-parametric tests, the Wilcoxon rank sum test, and the Kruskal-Wallis test were applied. Both the tests are designed for non-normally-distributed data because they are rank tests, which use the median to examine the difference of the data.

The Wilcoxon rank sum test is also called the *U*-test, and is similar to the *t*-test in that it can act as a means of examining the significance of the difference of two groups of data. The main difference between the two is that the *t*-test uses the arithmetic mean to test the data, and can only be applied to normally distributed data. As naturally-occurring linguistic data are rare, if ever normally distributed (cf. Gries, 2009; Kilgarriff, 2005), the *t*-test is less frequently applied to linguistic data, and the present research is no exception. Baayen (2008) provided clear instruction for using the *t*-test and *U*-test in R:

> The *t*-test is an excellent test for data that are more or less normally distributed. But it should not be used for variables with skewed distributions. For such variables, the ONE SAMPLE WILCOXON TEST, implemented in the function `wilcox.test`, should be used instead (p. 81).

According to this, the function `wilcox.test` in `R` will be used to test the significance of the difference of two non-normally distributed datasets in the tests reported later. In addition, it is worth nothing that the *U*-test uses the p-value ≤ .05 to reject the similarity of two groups of data.

The *T*-test and *U*-test are convenient methods for understanding the difference between two groups of data, but they 'cannot be applied to cases where you need to compare more than two means' (Gries, 2009, p. 274). In such cases, the Kruskal-Wallis test is applied if the data contain more than two groups of non-normally-distributed samples. The null hypothesis in the Kruskal-Wallis test is that 'all of the populations are identical against the alternative that some of the populations tend to applies furnish greater observed values than other populations' (Conover, 1999, p. 229). Though the Kruskal-Wallis test needs more computations, it can be done easily in `R` the function `kruskal.test` from the **core** package. The Kruskal-Wallis test also uses the p-value ≤ .05 to accept the difference.

However, the result of a Kruskal-Wallis test only reveals whether the difference of the entire data is significant or not, it cannot identify specific pair within the tested data. If the result of a Kruskal-Wallis test is significant, it only indicates that at least one pair data from the entire data is statistically different. To identify which pair of data is different, it is necessary to use a post-hoc test. The significance of individual pairs of differences in the Kruskal-Wallis test can be tested by multiple comparisons between treatments (Siegel & Castellan, 1988, p. 213), and the **stats** package in `R` provides the function `pairwise.wilcox.test` to address this need. The `pairwise.wilcox.test` function, as its name indicates, can 'calculate pairwise comparisons between group levels with corrections for multiple testing' by applying the Wilcoxon rank sum tests (R Team, 2013).

6.3 Machine Learning

As used in this research, machine learning is widely used in processing and analysing large-scale data. Machine learning techniques were used to undertake two tasks: identifying stock tweets and identifying the sentiment of stock tweets. This section first focuses on the theoretical background of machine learning techniques: It explains the motivation of using machine learning for the research and discusses suitable machine learning solutions for it.

6.3.1 Introduction to machine learning

Machine learning, often considered as a branch of artificial intelligence (AI), has developed rapidly in recent years (Segaran, 2008). The basic idea of machine learning is to enable computers to 'improve automatically with experience' (Mitchell, 1997, p. 1). Specifically, the computer learns to solve a problem by learning from given algorithms and prior training data, usually numerical data. Thus, it can significantly reduce human effort.

Mitchell (1997) defined the well-posed learning problems in machine learning as:

> A computer program is said to learn from experience *E* with respect to some class of Task *T* and performance measure *P*, if its performance at tasks *T*, as measured by *P*, improves with experience *E* (p. 2).

This explains the most noticeable characteristics of machine learning, that the algorithm 'learns' from the prior knowledge.

6.3.2 Motivation of applying machine learning

Although machine learning has been applied in data analysis and related areas for decades, it has not been widely used in linguistics or the other humanities. Applying machine learning methods usually requires a high level of knowledge of computing and mathematics, which poses a major obstacle to using machine learning methods in humanities research. On the other hand, the payoff of using machine learning on a small group of data might be limited, so it is often used on a large and complex set of data. However, processing large and complex data can cause another obstacle to applying machine learning.

The tweet data used in this research are relatively complex as discussed in Chapter 4. First, the sentiment information to be extracted is not explicit. For example, Potts (2011) pointed out the accuracy of sentiment analysis 'typically do[es] not exceed 90%'. Second, the tweet data are segmental. With the 140-character length limitation, each tweet contains limited information. Third, tweets contain a high proportion of non-standard language usages, so they can yield sparse word matrices as demonstrated later.

Thus, machine learning methods were applied to deal with the following two problems in the later analyses:

1. Identifying the ticker-related tweets from the pre-processing tweets and classifying them into different categories;
2. Identifying the sentiment polarity of the ticker-related tweets, and classifying them into different categories.

Chapter 7 focuses only on the second task, and Chapter 8 presents an investigation of both. Later, Chapter 9 tackles both problems with a hybrid approach.

6.3.3 Supervised and unsupervised learning

In general, there are two major branches of machine learning: supervised learning and unsupervised learning. Different learning methods are applied to specific tasks according to Bishop (2006):

1. Supervised learning, or machine induction, is suitable for dealing with 'applications in which the training data comprises examples of the input vectors along with their corresponding target vectors'.
2. Unsupervised learning is applied where 'the training data consists of a set of input vectors *x* without any corresponding target values' (p. 3).

In brief, a supervised learning task is a classification task, while an unsupervised learning task is a clustering task. As supervised machine learning methods often outperform unsupervised machine learning methods in the sentiment analysis field (Vohra & Teraiya, 2013), the chapter outlines experiments with the four main supervised methods currently being used in the sentiment analysis field: decision tree, random forests, Naïve Bayes and support vector machine methods in order to find the one with the best accuracy for this specific task.

According to Mitchell (1997), the performance of machine learning can be maximised when training examples follow a distribution similar to that of the future test examples. Thus, tweets of one ticker, General Electric, were chosen, and the sentiment of each tweet was manually annotated as shown in Section 4.4. The annotated tweets were used as prior experience for the further machine learning tasks: The sentiment classification uses different parts as the training and test sets. Therefore, the training data and the test data are from the same source but differ in size.

As introduced above, the classification task in the following analysis is two-fold: identifying relevant tweets and identifying tweet sentiment. The first task matches the definition of a classification task, which 'assign(s) each input vector to one of a finite number of discrete categories' (Bishop, 2006, p. 3). Accordingly, a suitable supervised machine learning method needed to be identified and tested for this research. The second task was more complicated. In the sense of complexity of the sentiment classification, there are two branches: the polarity problem and fine-grained problem (see Chapter 4). To match the trinary characteristics of the stock market (as discussed in Section 4.4.6), the sentiment analysis task in the following chapters simply uses a polarity approach, which classifies them into three categories: negative, neutral and positive.

The detailed introductions of specific machine learning algorithms are presented below.

6.3.4 Decision tree classification

'Approximating discrete-valued target functions' (Mitchell, 1997, p. 52), the decision tree is a robust method to classify data into discrete categories and present the result as a tree map. Simply put, each time, the decision tree generates a binary branch from the parent branch according to the most distinct features, so it is regarded as a single classification tree. Because Mitchell (1997) suggested that it is 'among the most popular of inductive inference algorithms and have been successfully applied to a broad range of tasks' (p. 52), it becomes the first choice in this section.

Moreover, Mitchell (1997) suggested that the decision tree method works particularly well when the following conditions are met:

1. Instances are represented by attribute-value pairs.
2. The target function has discrete output values.
3. Disjunctive description may be required.
4. The training data may contain errors.
5. The training data may contain missing attribute values (p. 54).

As introduced above, a word matrix was used as the training data. The word matrix was converted to numerical values, so it can well represent the instances by attribute-value pairs. Also, the matrix has a fixed number of categories, so they are discrete. Therefore, the data analysed in this research fit the above rules well.

In statistics, this method belongs to the recursive partitioning method, and `R` provides various packages to conduct decision tree classifications, such as the package **tree**, **rpart** and **party**. The following analysis used the package **rpart** to train and test the decision tree classifier.

6.3.5 Random forests classification

Generally, the random forests method can be considered as an improvement of the decision tree, which is 'intended to improve the predictive accuracy of tree-based learning algorithms' (Kononenko & Kukar, 2007, p. 95).

Defined by Breiman (2001),

> A random forest is a classifier consisting of a collection of tree-structured classifiers {*h*(**x**, Θ_k_), *k*=1, ...} where the {Θ_k_} are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input **x** (p. 2).

Put more simply, the random forests method generates a number of decision trees and lets them vote to select the best results.

Although the random forests method is generally considered more robust than the decision tree method, it is not easy to understand how the classifier works or reproduce the reasoning behind it (Kononenko & Kukar, 2007; Segaran, 2008). This is mainly caused by the complicated voting procedure, which is usually considered as a black box. In other words, the random forests method is less clearer than the decision tree method to present which feature works in each partition.

In R, several packages provide random forests training and testing, such as **e1071** and **randomForest**. Due to the convenience of conducting the 10-fold cross validation, this research used the **randomForest** package to carry out the random forest classification.

6.3.6 Naïve Bayes classification

Based on the Bayes theorem, in classification tasks, the Naïve Bayes method is 'among the most effective approaches algorithms known' (Mitchell, 1997, p. 155). According to the Bayes theorem, the posterior probability $P(h|D)$ is determined by the prior probability $P(h)$, $P(D)$ and $P(D|h)$. A typical machine learning problem often consists of several hypotheses, so the overall probability of the problem, namely *maximum a posterior hypothesis* (MAP), is decided by each candidate hypothesis (Mitchell, 1997). Due to the simplicity, it often yields a good result, so the following analysis also considered the Naïve Bayes method as a candidate.

Using the kernel function to smoothen the data is a common technique in Naïve Bayes algorithm. As a locally weighted regression method, the kernel function uses 'nearby or distance-weighted training samples to form the local approximation' (Mitchell, 1997, p. 236). By using the kernel function, the training can be more efficient (Mitchell, 1997). In the cross validation (see Section 6.3.8), the Naïve Bayes algorithm compares the results between using and not using the kernel function.

In R, packages such as **e1071** and **KlaR** provide the Naïve Bayes algorithm. Because the package **caret** includes the Naïve Bayes algorithm from the package **KlaR** to conduct the 10-fold cross validation, the classification training in this research followed this approach.

6.3.7 Support vector machine classification

The fourth supervised learning experiment is to use support vector machine method, because it often outperforms other supervised learning methods in sentiment classification (Mukherjee et al., 2012). Recently, the support vector machine method has become popular 'for solving problems in classification, regression and novelty detection' (Bishop, 2006, p. 325). Particularly in natural language processing, the support vector machine is among the most commonly used methods among many applications (Vohra & Teraiya, 2013). According to Cortes & Vapnik (1995),

> The support-vector network implements the following idea: it maps the input vectors into some high dimensional feature space Z through some non-linear mapping chosen a priori. In this space, a linear decision surface is constructed with special properties that ensure high generalisation ability of the network (p. 274).

In other words, the idea of the support vector machine is to convert data into a high-dimension map according to different features, and then find the best decision line, which leaves the largest margin to any data point in each feature category.

In R, there are several packages providing the support vector machine algorithm, such as package **e1071**, **klaR** and **kernlab**. The following analysis chooses the package **ksvm** to perform the support vector machine training because it contains the cross validation function in the package.

6.3.8 Cross validation

There are several approaches to evaluating the accuracy of a machine learning result, and cross validation is a common choice. According to Segaran (2008):

> Cross-validation is the name given to a set of techniques that divide up data into training sets and test sets. The training set is given to the algorithm, along with the correct answers (in this case, prices), and becomes the set used to make predictions. The algorithm is then asked to make predictions for each item in the test set. The answers it gives are compared to the correct answers, and an overall score for how well the algorithm did is calculated (p. 176).

Generally, data are divided into K-fold equal-sized parts, and K is often taken as 5 or 10 (Hastie, Tibshirani, & Friedman, 2001). For the analysis presented in the following chapters, K was chosen as 10, so the data were divided into 10 equally sized subgroups and the validation applied 10 times.

6.3.9 Test set of CAT tweets

In order to test whether the classification trained above worked on comparable data, they were tested on another dataset, tweets from the CAT (Caterpillar Inc.) ticker. Following the same procedure as introduced in Chapter 4, 702 CAT tweets were annotated. After annotation, 525 were found to be related tweets. These tweets were then submitted to the trained automatic classifiers in the following chapters. Then, the results of the automatic classifiers were compared with the manual annotation results to assess the accuracy of the automatic classifiers.

6.4 Summary

This chapter provided a detailed overview of the methodology used in the later main data analysis, covering linguistic analysis, statistical analysis and machine learning.

Chapter 7 Sentiment Classification Based on the Internal Linguistic Features

The bag-of-words model is among the most common approaches in current sentiment analysis. This chapter tests such a model with stock tweet data, and sets up a baseline accuracy of sentiment classification for later experiments. The key step for sentiment classification based on the bag-of-words model is to design a suitable sentiment word list, so this step becomes the focus of this chapter. As discussed in Chapter 3, to generate a word matrix for the bag-of-words approach, previous studies applied different pre-designed sentiment word lists based on the other types of textual data, but it remains unanswered how robust these word lists are when they are used to process stock tweet data. Therefore, this chapter first tests some popular sentiment word lists being used in previous studies.

Considering the poor performance of these word lists, mainly the poor coverage, it then focuses on the nature of tweet language itself to find proper features to build the bag-of-words model, particularly concentrating on frequent patterns and part-of-speech features. As all of these features relate to the meaning of tweets, they will be regarded as the **internal linguistic features** of tweets in this research. This chapter first analyses different internal linguistic features of stock tweets, including the most and least frequent unigrams, but neither of them works to classify polarity as will be shown. Thus, the analysis applies stemming and keyness analysis to generate a more robust sentiment word list, in order to convert stock tweets to a word matrix. A sentiment classification based on this word matrix is then performed by four supervised learning classifiers as introduce in Chapter 6. The four classifiers achieved a reasonably good result with an accuracy of around 0.6 at the training stage, and at the test stage, they achieved a better accuracy. The sentiment classification accuracy based on the bag-of-words model is thus considered as the baseline accuracy for the following chapters. Finally, it analyses the part-of-speech features of the annotated stock tweets, and trains automatic classifiers based on these features. However, these features offer limited help to identify the polarity.

With regard to this analysis of the internal feature of the annotated tweets, several points should be borne in mind. The comparison focuses on the ticker-related tweets only, and all tweets in the analysis were converted into the lowercase format. Only three groups of ticker-related tweets are then studied. NEG tweets (8200 tokens, 5613 word types), NEU tweets (9951 tokens, 6381 word types), and PST tweets (29,362 tokens, 12,411 word types). URLs were excluded from the analysis, and all hash marks and dollar marks were removed, such that, for example, the cashtag '$GE' was considered as a word 'ge'. Finally, PTW means 'per thousand words'. In corpus linguistics, the base for normalisation is usually PMW (per million words), but the corpus size is relatively small in this research, so PTW is applied.

7.1 Test of Word Lists in Previous Studies

As reviewed in Chapter 3, several word lists were used to build the word matrix for training a bag-of-words model for sentiment classification. Common choices of word lists included the Harvard-IV-4 classification dictionary (Ogilvie et al., 1982), Loughran and McDonald Financial Sentiment Dictionaries (Loughran & McDonald, 2011), Opinion Lexicon (Hu & Liu, 2004), Subjectivity Lexicon (Wilson et al., 2005b) and the Twitter sentiment analysis list (2011). This section evaluates the use of these word lists to classify the tweet data collected in this research. The primary concern is the coverage of these lists: if the word lists are unrepresentative of the terminology used in the collected tweet data, the majority of records will not be assigned sentiment. In the case that coverage of a list is inadequate for data taken from a given domain, that list may be viewed as an inappropriate choice of resource.

Among the above five word lists, only the first 100 entries of the Harvard-IV-4 classification dictionary were available on the web page *Descriptions of Inquirer Categories and Use of Inquirer Dictionaries* (`http://www.wjh.harvard.edu/~inquirer/homecat.htm`). The webpage *Maryland Webuse site* (`http://www.webuse.umd.edu:9090/tags/`), providing the full version, was not accessible at the time of analysing. Thus, this list could not be tested.

Table 7.1.1 shows the coverage of the other four popular sentiment word lists being applied to the dataset in this research. None of them covered all records in the dataset. The first three were designed based on textual data other than twitter data, so their coverages were rather poor. One possible approach to improve classification accuracy might be the normalisation of tweet data to reflect standard spelling forms. However, no tweet normaliser is available so far as discussed in Chapter 6, so this approach is not at present an option. Although the last word list outperformed the first three as it was designed based on Twitter data, it is biased between the negative and positive categories, at a ratio of 8:1. The consequence of such a biased list is that the generated word matrix will also be biased. Also, the word list contains 114 emoticons, which are rarely used in stock tweets. Therefore, it is of limited applicability to sentiment classification in this domain.

Table 7.1.1 Comparison of four popular word lists in previous studies

Word list	**Total words**	**Positive words**	**Negative words**	**Covered tweets**
Financial Sentiment Dictionaries	2683	354	2329	1209 (17.95%)
Opinion Lexicon	6789	2006	4783	2662 (39.52%)
Subjectivity Lexicon	8221	2718	4913	2894 (57.81%)
Twitter sentiment analysis list	7472	6654	818	6229 (92.47%)

According to this case study, two requirements for designing a better sentiment word list for stock tweet data are clearly identified: such a list needs to have a better coverage and needs to be less biased between the negative and positive polarities Therefore, the rest of this chapter attempts to develop such a better word list for generating a bag-of-words sentiment word matrix.

7.2 Frequency Analysis of Stock Tweets

Debbini et al. (2011) used a different approach, as they did not use a sentiment word list predesigned by others, but instead extracted the top 1000 frequent words to generate a sentiment keyword list. This section thus reports the analysis of this approach to see if it really works. As the first step, the research uses frequency lists to analyse the most and least frequent unigrams in different polarities, as both of them are possibly the most obvious features for generating a word matrix to build the bag-of-words model. To study these features, the current research adopts the methods taken from corpus linguistics analysis, which is often used to explore massive quantities of data. However, it will subsequently be seen that neither approach is fully able to reveal distinct characteristics of different polarities of stock tweets.

7.2.1 The most frequent unigrams in stock tweets

The first analysis focuses on the most frequent unigrams in the ticker-related tweets corpus. After removing all grammatical words from the top 50 frequent unigram list of the three groups of ticker-related tweets, all of them have 36 unigrams as listed in Table 7.2.1. In terms of the sequence of the unigrams, there are no explicit differences across these three groups.

Table 7.2.1 Meaningful unigrams in the top 50 frequent unigrams of ticker-related tweets

NEG			NEU			PST		
Unigram	**Freq.**	**PTW**	**Unigram**	**Freq.**	**PTW**	**Unigram**	**Freq.**	**PTW**
ge	699	85.24	ge	885	88.94	ge	2849	97.03
electric	118	14.39	electric	201	20.20	t	592	20.16
general	109	13.29	general	177	17.79	electric	476	16.21
stock	93	11.34	news	124	12.46	news	464	15.80
news	91	11.10	t	113	11.36	general	393	13.38
is	91	11.10	s	108	10.85	s	333	11.34
s	89	10.85	stock	105	10.55	rt	240	8.17
rt	79	9.63	is	97	9.75	stocks	234	7.97
stocks	64	7.80	rt	85	8.54	stock	196	6.68
analysis	58	7.07	stocks	81	8.14	is	175	5.96
bearish	45	5.49	i	74	7.44	earnings	168	5.72
0	42	5.12	company	70	7.03	i	126	4.29
trading	39	4.76	analysis	65	6.53	dividend	124	4.22
are	39	4.76	earnings	54	5.43	this	123	4.19
down	38	4.63	it	52	5.23	new	120	4.09
i	35	4.27	this	49	4.92	up	110	3.75
1	35	4.27	msft	49	4.92	capital	104	3.54
moody	34	4.15	trading	45	4.52	it	102	3.47
t	31	3.78	aapl	42	4.22	3	98	3.34
be	31	3.78	will	35	3.52	week	95	3.24
falling	30	3.66	jimcramer	35	3.52	msft	86	2.93
jpm	29	3.54	dow	34	3.42	energy	83	2.83
bac	29	3.54	more	33	3.32	buy	83	2.83
capital	28	3.41	buy	32	3.22	5	83	2.83
20	28	3.41	energy	31	3.12	bac	79	2.69
nbc	27	3.29	capital	31	3.12	are	78	2.66

NEG			NEU			PST		
taxes	26	3.17	4	31	3.12	today	77	2.62
2	26	3.17	20	31	3.12	2	77	2.62
today	25	3.05	you	30	3.01	trading	75	2.55
my	25	3.05	3	30	3.01	1	75	2.55
19	24	2.93	short	28	2.81	healthcare	70	2.38
will	23	2.80	be	28	2.81	20	70	2.38
it	23	2.80	jpm	26	2.61	analysis	69	2.35
3	22	2.68	but	26	2.61	aapl	68	2.32
spy	21	2.56	that	25	2.51	0	68	2.32
short	21	2.56	like	25	2.51	top	65	2.21

One interesting phenomenon is that a number of ticker names appear in these three groups of highly frequent unigrams. In total, 6 unigrams are ticker names: in addition to *ge*, *t* is for AT&T, *bac* for bank of America Corporation, *aapl* for Apple, *msft* for Microsoft, and *jpm* for JP Morgan. Also, *electric* and *general* also appear at the top of the list. The combination *general electric* occurs in 621 tweets (18.8% of all ticker-related tweets). These illustrate that the ticker-related tweets focus strongly on the market. Looking at these ticker names, it is hard to decide their relationship with General Electric: they spread across the domains of technology and finance, so it is unclear whether GE is in competition with them as a hi-tech enterprise, or (in the case of GE Capital) as a finance industry.

Apart from ticker names, other unigrams relating to the market are present. In the PST tweets group, *news* is the fourth most frequent unigram in the list, with a normalised frequency of 15.80 PTW. Compared with the other two groups, it is much higher. Observing the collocation containing *news*, the most frequent combination is *GE news*: 367 tweets in the PST tweets (17.87%), 98 tweets in the NEU tweets (13.96%), and 65 tweets in the NEG tweets (11.73%). Therefore, it is possible to conclude that most news about GE tend to be positive (69.24% in PST, 18.49% in NEU, and 12.26% in NEG). This roughly matches the price movement of the GE ticker as shown in Chapter 5. Next, unigrams such as *capital*, *energy*, and *healthcare*, all of which indicate areas in which GE is involved, appear more frequently in the PST group. This relates to the fact that the PST group contains more news tweets than the other two groups, so they use more news-related expressions.

Also, words indicating investment behaviours between these stock-related unigrams are divisible into three polarities: *earning*, *dividend*, *buy* and *up* occur in the PST group, and *bearish*, *down*, *moody*, *falling* and *short* occur in the NEG group. However, these words appear in a small proportion of the top list. Additionally, they appear across different categories of tweets. Thus, it is difficult to identify the polarity of tweets solely on the basis of these stock-related unigrams.

Finally, there are two terms indicating time in this top list: *today* and *week*. In total, in three polarities, there are 339 occurrences of *week*, and 177 occurrences of *today*, but most of them occur in the PST group (though the PTW of *time* in the NEG group is higher than in the PST group). Taking a closer look at the words around *week*, the most frequent bigrams are *this week* (117 occurrences) and *week high* (26 occurrences): the first indicates that most tweets with *week* focus on a short period, and *week high* is a stock term used to evaluate the overall trend of a ticker in a period, such as *52 week high*. The high frequency of the bigram *this week* and the unigram *today* shows that the ticker-related tweets tend to discuss the performance of the market over a relatively short period.

Due to the high degree of similarity among the highly frequent unigrams of ticker-related tweets, it is difficult to train a bag-of-words model based on these most frequent unigrams.Thus, the research proceeds with the analysis of an alternative approach – a reversal of the previously discussed method.

7.2.2 The least frequent unigrams in stock tweets

The reverse approach investigates the unigrams occurring in one group of tweets, but not in the other two groups. Often, these unigrams are the least frequent unigrams in the text sample. Therefore, with these unigrams in the sample data, defined as **unique unigrams**, the polarity of tweets might be identified. To understand how the reverse approach works, Figure 7.2.1 illustrates the relationship across three types of tweet unigrams.

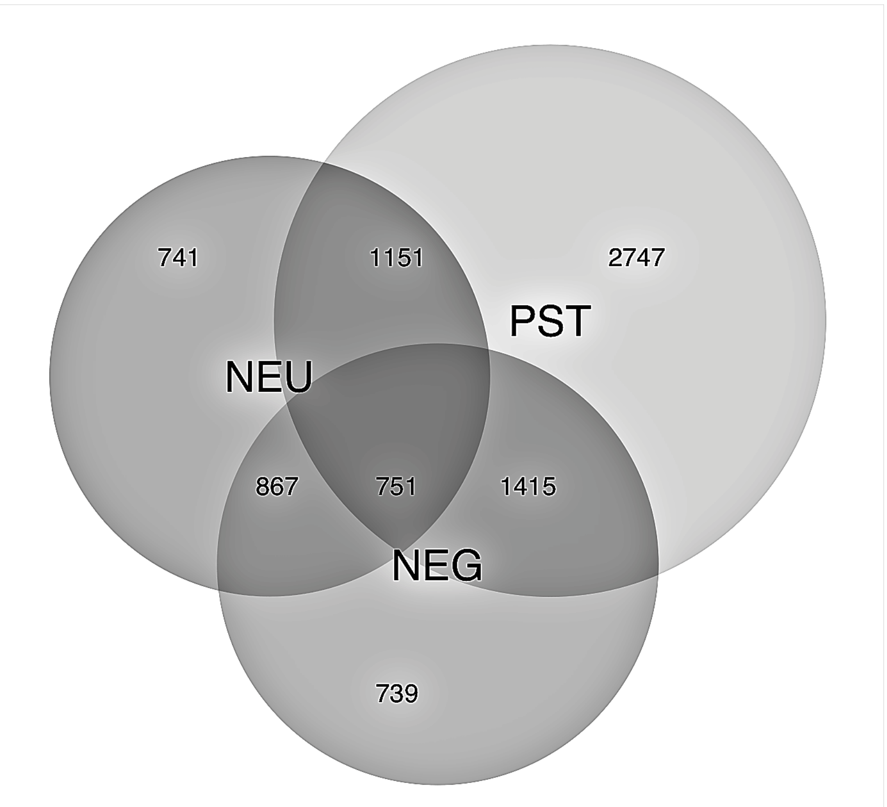


Figure 7.2.1 Statistics of unigrams in different polarities of tweets

Each circle represents one type of ticker-related tweets, and each zone with a different grey scale represents a different relationship among the unigrams. First, in the centre, all three circles overlap to form the darkest zone, indicating that these unigrams occur in all three types of ticker-related tweets. Then, in other shadowed areas around the centre, each unigram appear in two overlapped circles, suggesting these unigrams in any two types of tweets. Finally, as the remaining areas have no overlaps, these areas represent unigrams appearing in only one type of tweet, namely unique unigrams. In the reverse unigram analysis, the focus is on the last type of unigram – unique unigrams, which may significantly differentiate one type of tweets from other two groups.

In this graph, the PST category has the largest number of unique unigrams – 2747 unique unigrams, while the NEG and NEU group only shares 739 and 741 unique unigrams. However, there are a number of Arabic numbers and mixed spellings of numbers and characters. They are less helpful for recognising the polarity of a tweet, so they have been removed. As a result, the NEG, NEU and PST group has 680, 693 and 2531 unique unigrams respectively. Extracting all unique unigrams from three polarity groups, the NEG, NEU and PST group yield 999, 1152 and 5020 occurrences respectively (Table 7.2.2), which is 12.18%, 11.58% and 17.10% of the total occurrences of unigrams in each group. According to this, unique unigrams account for only a small proportion of the full set of tokens (15.09%).

Table 7.2.2 The relationship of unique unigrams and total tokens

Polarity	**Occurrence**	**Tokens**	**Percentage**
NEG	999	8200	12.18%
NEU	1152	9951	11.58%
PST	5020	29,362	17.10%
Total	7171	47,513	15.09%

Presumably, this method is only effective if the coverage of these unique unigrams is good. Table 7.2.3 shows the extracted results. The resulting unique unigram lists can identify the polarity of 75.32% of tweets on average: 74.01% and 80.14% of NEG and PST tweets respectively, but only 62.25% of NEU tweets. In addition, regarding the ratio of the count of unique unigrams with the count of covered tweets, each tweet has at least 2 unique unigrams on average, in particular, each tweet containing unique unigrams in the PST group contains more then 3 unique unigrams on average.

Table 7.2.3 The coverage of unique unigrams

Polarity	**Tweets**	**Tweets with unique unigrams**	**Percentage**	**Total unique unigrams**
NEG	554	410	74.01%	999
NEU	702	437	62.25%	1152
PST	2054	1646	80.14%	5020
Total	3310	2493	75.32%	7171

However, looking at these unique unigrams, one third of them only occur once in the corpus of ticker-related tweets (see Table 7.2.4), and a large proportion of them are non-standard spellings. In both NEG and NEU groups, about half of the unique unigrams have only one occurrence, although this figure is lower in the PST group. The high frequency of unique unigrams with only one occurrence indicates that unique unigrams are not stable, so they do not seem to be a reliable means of identifying polarities. In addition, considering the lowest percentage (not the frequency) of one-occurrence unique unigrams and the largest count of tokens of the PST group, it is reasonable to infer that, with a larger dataset, the long tail of a unigram list will become more stable, so, overall, the proportion of one-occurrence unigrams will reduce. This suggests that using a larger corpus to generate a more stable unique unigram list might be helpful for identifying the polarity. Nevertheless, for the current research, which is based on a relatively small annotated tweets corpus, this approach offers limited help.

Table 7.2.4 Relationship of one-occurrence unigrams and the whole unigrams

Polarity	**Total**	**one-occurrence**	**Percentage**
NEG	999	450	45.05%
NEU	1152	523	45.40%
PST	5020	1427	28.43%
Total	7171	2400	33.47%

7.2.3 Discussion

This section reported an analysis of the most and least frequent unigrams in the annotated stock tweets respectively, but neither of them can be used to build a bag-of-words model.

The first analysis, which focused on the most frequent unigrams in the annotated stock tweets showed that, in general, different polarities were similar in terms of the most frequent unigrams. All three polarities frequently contained ticker names. However, these frequent unigrams had subtle differences. Unigrams relating to news and time appeared much more often in the PST tweets than the other two groups. This showed the impossibility of using the most frequent unigrams in the sample data to generate a sentiment word list. In turn, a word matrix of the bag-of-words model could not be built for sentiment classification.

The second analysis with a focus on the least frequent unigrams in the annotated tweets indicated that there are distinct differences of unique unigrams in different polarities, and they could cover a large proportion of stock tweets. However, they were not stable as most of them only occurred once in the data. This may explain Yi (2009)'s finding that most of the vocabulary in the web domain make little contribution to his following training of machine learning classifiers. Also, in the one-occurrence unigrams, a large proportion of them are non-standard spellings, which supports Go and Bhayani (2010), Kaufmann (2010), Bar-Haim et al. (2011) and Zappavigna (2012)'s view that non-standard spellings are very common in tweets as discussed in Section 3.4.1. Therefore, this analysis showed that using the least frequent unigrams in the sample data could not generate a stable sentiment word list usable for the subsequent automatic classification.

These two unigram analyses suggested that two main problems still remain. First, a number of unigrams were different word types of the same token; second, a number of unigrams had overlaps across different categories. In such a small size corpus, these two problems caused further challenges of their own. Therefore, the following analyses try to reduce the impact of these two problems by using stemming and keyness analysis respectively.

7.3 Design of A Localised Bag-of-words Model

As shown above, neither the most and least frequent unigrams can be used to create a word matrix for sentiment classification due to their indistinct or unstable features. To tackle the first problem, this section uses keyness analysis to see if these three polarities contain any distinct keywords to differentiate them from each other. To solve the second problem, two possible approaches were discussed in Chapter 6: either using a suitable normaliser to reduce the number of non-standard spellings, or using a stemmer to conflate words with the same stems. However, due to the unavailability of a suitable tweet normaliser, this section focuses solely on the use of a stemmer to deal with this problem. Finally, this section creates a sentiment word list based on the keyness analysis of stemmed tweets, and then convert the stock tweets to a word matrix according to this word list.

7.3.1 Stemming ticker-related tweets

Applying the Porter stemmer (Porter, 1980) to the annotated tweets reduced the number of word types in each category to one third of their original size, as shown in Table 7.3.1. This conflation is much more significant than the example in Porter (1980), in which the stemmer reduced the word types to two thirds. Although it is not clear what type of text Porter used in his experiment, this comparison shows that the annotated tweets contain a higher percentage of variant spellings than the textual data in Porter (1980)'s experiment.

Table 7.3.1 Comparison of the word types in different polarities before and after stemming by the Porter stemmer

Polarity	Before stemming	After stemming	Percentage
NEG	5613	2004	35.70%
NEU	6381	2257	35.37%
PST	12,411	4532	36.52%

Removing grammatical words, Table 7.3.2 shows an extract of the highest frequent unigrams after stemming. Compared to Table 7.2.1, there are several clear differences. For instance, the words *stock* and *stocks* in Table 7.2.1 are conflated to *stock*, while *general* and *electric* are conflated to *gener* and *electr* respectively. Also, most items in this list have a slight increase in terms of the degree of density, which is also helpful for generating a less sparse word matrix. However, the order of these top frequent items does not change much, compared to the most frequent unigram list as shown in Table 7.2.1. Considering that the stemmer conflated a higher percentage of words than the proportion achieved in the original experiment by Porter (1980), the top stemmed unigram list suggests that the stemmer mainly conflates words with a lower frequency. This method significantly reduces the number of word types, so a word list based on this result will be more stable than a list generate without stemming. Therefore, the following analysis uses the stemmed data to conduct a keyness analysis.

Table 7.3.2 Sample of the top 20 frequent unigrams after stemming

	NEG			NEU			PST	
Unigram	**Freq.**	**PTW**	**Unigram**	**Freq.**	**PTW**	**Unigram**	**Freq.**	**PTW**
ge	699	85.24	ge	885	88.94	ge	2849	97.03
stock	149	18.17	gener	179	17.99	t	593	20.20
gener	109	13.29	electr	179	17.99	electr	425	14.47
electr	107	13.05	stock	175	17.59	stock	409	13.93
is	92	11.22	a	121	12.16	gener	407	13.86
a	87	10.61	news	107	10.75	news	399	13.59
rt	79	9.63	is	97	9.75	a	277	9.43
news	74	9.02	rt	85	8.54	rt	240	8.17
analysi	58	7.07	t	77	7.74	compani	198	6.74
bearish	45	5.49	compani	77	7.74	new	185	6.30
0	42	5.12	i	74	7.44	is	175	5.96
downgrad	38	4.63	analysi	65	6.53	dividend	139	4.73
down	38	4.63	it	61	6.13	it	138	4.70
ar	38	4.63	trade	60	6.03	earn	127	4.33
trade	37	4.51	msft	49	4.92	i	126	4.29
tax	36	4.39	thi	48	4.82	thi	123	4.19
i	35	4.27	earn	44	4.42	up	116	3.95
1	35	4.27	aapl	42	4.22	bui	107	3.64
moody	34	4.15	new	38	3.82	week	103	3.51
be	33	4.02	will	34	3.42	look	102	3.47

Using VARD to reexamine the stemmed result, the percentage of spelling variants appeared higher than the unstemmed result as shown in Chapter 6: 28.37% of all tokens were regarded as non-standard spellings. This was a seemly worse result. However, considering that the Porter stemmer only leaves a valid stem according to a list of suffixes (Porter, 1980), it is reasonable to see that stemmer destroyed standard spellings. For example, highly frequent tokens *company* and *energy* were stemmed to *compani* and *energi* respectively in the result. In other words, a number of originally standard words were conflated to designed forms, which could not be regarded as standard spellings by VARD.

In short, although the stemmer produced a stemmed result with more spelling variants, it did reduce the count of the word types considerably. This is particularly important given that the tweet corpus being analysed in this thesis is relatively small: the greater the number of word types, the sparser the corpus becomes, which will affect the accuracy of further analyses based on the 'bag of words' model. Conflating words with the same root to similar forms decreases the counts of word types in such a small corpus, which in turn decreases the sparseness of the corpus. In summary, then, this thesis regards stemming as an efficient way to improve word density in the corpus to be analysed.

7.3.2 Keyness analysis of the unigram lists

The above analyses based on the unigram features in Section 7.2 have not revealed sufficiently distinct features across different polarities; on the contrary, it seems – somewhat disappointingly – to be better at identifying overlaps. Thus, this section uses keyness analysis to test whether a keyword across polarities can play a better role of identifying any differences that might lie in the data. It uses the neutral category as the reference list, and compares the reference list with the positive and negative categories The assumption here is **not** that the keywords in the positive category are positive, or in the negative categories are negative; instead, the analysis assumes that the keywords in these polarities are **more** positive or negative than those in the other categories. By "more", it means that the word has a higher rank of keyness score in that category, and this will be explained by the following analysis.

Using keyness analysis to identify the keywords from the stemmed tweets, 1382 and 3048 keywords were identified from the negative and positive category respectively, and 520 keywords, namely **mutual keywords**, occur in both categories. Table 7.3.3 shows the top 20 stemmed unigrams in these two categories, and the polarity difference is clearer than the unigram frequency list in Table 7.2.1. In the negative category, words such as *downgrad* (the stemmed form of *downgrade*), *bearish*, *fall*, *down*, *cut*, *lower*, *alert*, and *suspends* have a direct association with negative sentiment, or link to the selling sentiments as discussed in Chapter 4. Conversely, in the positive category, words such as *dividend*, *expect*, *rise*, *beat*, *high*, *expand*, *surprises*, *profit*, *growth*, *top*, *bull*, *high*, *surprise* and *bull* have a closer relationship with the positive or buying sentiment. In addition, words relating to the ticker name, such as *ge*, *gener*, or *electr*, occur neither in this top keyness list, nor in the entire keyness list. This shows that this keyness list approach can remove a certain proportion of topic-related words from the frequency list, so the result is less sensitive to topic. Therefore, the generated sentiment unigram list might also be useful for classifying tweets with other ticker names.

Table 7.3.3 Comparison of words with top 20 keyness scores in negative and positive category

Negative			Positive		
Unigram	**Keyness**	**Keyness**	**Unigram**	**Keyness**	**Keyness**
38	61.441	downgrad	139	50.666	dividend
45	53.854	bearish	593	32.305	t
32	43.957	fall	67	31.030	expect
34	41.885	moody	50	28.871	whose
36	36.919	tax	62	28.296	wind
17	27.487	rsi	48	27.716	rise
25	24.893	pai	73	25.178	beat
38	22.228	down	98	24.517	ttfb
27	21.816	nbc	63	20.269	healthcar

	Negative			Positive	
13	21.019	tv	69	19.898	high
23	19.254	cut	646	19.894	to
11	17.786	credit	31	17.900	expand
11	17.786	dishwash	29	16.745	calendar
11	17.786	mfi	29	16.745	surprises
17	17.059	lower	62	16.687	profit
10	16.169	alert	27	15.590	trv
9	14.552	fx	53	15.506	growth
9	14.552	recal	80	14.709	top
9	14.552	suspend	25	14.436	surprise
21	14.396	rate	24	13.858	bull

To reduce the ambiguous effect of the mutual keywords, they were first classified according to their relative frequency. For example, *dividend* has 139 occurrences in the positive category, and 9 occurrences in the negative category. Therefore, the relative frequency for *dividend* in the PST category is 139/29,362 = 0.0047, and in the NEG category is 9/8,200 = 0.0011. The subtract result is 0.0036. This term is therefore regarded as a positive unigram. Based on this, 328 mutual unigrams were classified as positive unigrams, and 88 as negative unigrams, and 104 further mutual unigrams with a zero value are removed. This seems a useful way to classify these mutual keywords, but it remains rather biased.

Then, it compared the rank of each keyword in the two categories of the keyness list. For example, the above term, *dividend*, has the highest keyness score in the positive keyness list, so it ranks at the first place. Additionally, it also appears as the 462th item in the negative keyness list. Therefore, the rank of *dividend* in the positive list is much higher than its rank in the negative list. To improve accuracy, the relative rank was then used to calculate the score as

Score = Rank in NEG list / Size of NEG list - Rank in PST list / Size in PST list

For instance, the negative list contains 1382 unigrams, and the positive list has 3048 unigrams, so the score of *dividend* is calculated as 1/1382 - 462/3048 = -0.2248623. In a nutshell, if a mutual keyword's relative rank score is negative, it is then regarded as a positive unigram, and vice versa. Based on this classification, 277 mutual unigrams were classified as the positive unigrams, and 243 mutual keywords as the negative unigrams. Although few words are misclassified, this approach provides a less biased classification result than the previous attempts.

To summarise, using keyness analysis can generate a less biased, and less topic-sensitive localised sentiment word list. In this case, after removing those unigrams with only one occurrence, the list retains 1610 positive keywords, and 733 negative unigrams (see Table 7.3.4). The localised word list contains only one third of the words appearing in Davies and Ghahramani (2011)'s list. In addition, the localised word list is much more balanced Davies and Ghahramani (2011)'s list in terms of the distribution of positive and negative unigrams. However, ascertaining whether this word list is effective in the classification of sentiments in stock tweets, and whether it is less biased in the classification than the other lists may be assessed in a later automatic classifier training and testing.

Table 7.3.4 Size comparison of the localised sentiment word list and Davies and Ghahramani (2011)'s list

Word list	**Total words**	**Positive words**	**Negative words**
Twitter sentiment analysis list	7472	6654	818
Localised sentiment word list	2343	1610	733

7.3.3 Generating a word matrix based on a localised word list

In order to understand how the bag-of-words model works in sentiment analysis, and to set up a baseline for the following classifications, this section converts the annotated GE tweets to a word matrix – bag-of-words matrix – based on the aforementioned keyness analysis. According to the occurrence, the unigrams in the positive list are assigned with positive scores in the matrix, and the unigrams in the negative list are assigned with negative scores. Table 7.3.5 shows the converted word matrix based on the localised word list. For example, a tweet *GE screwed up everything* contains a word *screw* from the negative key word list, so it is assigned a negative score (see the second line in Table 7.3.5). After the conversion, the sum of the negative score, positive score and total score of each tweet will be calculated respectively.

Table 7.3.5 An extract of the word matrix based on the annotated GE tweets

Polarity	**effici**	**imsc**	**puls**	…	**screw**	**reaffirm**	**spy**	…	**Sum of PST**	**Sum of NEG**	**Sum of total**
NEG	0	0	0	…	-1	0	0	…	0	-1	-1
NSR	1	0	0	…	0	-1	0	…	1	-1	0
PST	1	0	1	…	0	0	-1	…	2	-1	1
NEU	0	0	0	…	0	0	0	…	1	-2	-1

In Table 7.3.5, each row represents the record of a converted tweet. Each cell contains the number of occurrences of a unigram appearing in a given tweet, and the occurrence is then regarded as a score. The first column indicates the polarity of a tweet, the second to 1612th columns indicate the positive scores, and the 1613th column to 2344th columns indicate the negative scores. The last three columns indicate the sum of the positive, negative and total scores. As can be observed from this extract, the generated matrix is extremely sparse.

Table 7.3.6 presents the average score of each polarity group of the GE stock tweets. In this case, the total score and the absolute total score are calculated as:

1. Total score = positive score + negative score
2. Absolute total score = positive score - negative score

The absolute total score indicates how many unigrams in a tweet can be identified by the word list, while the total score indicates the sentiment polarity of a tweet. From the average total score, the three polarities have clear differences: the NEG and PST groups have a higher score in different directions, and the NEG group is in the middle. The score of the NEU group is close to 0, which represents the ideal score of a neutral set. This shows that the localised word list can clearly identify the polarity of each group. In addition, both the NEG and PST group have a similar absolute total score, and the NEG has a slightly lower absolute total score. As will be shown later in Chapter 8, the average word count of the GE stock tweets is about 15, and the localised word list can identify more than 50% of the words used in each tweet. This shows that the localised word list can cover the words in each individual tweet well.

Table 7.3.6 Statistics of average score of each polarity group in the GE tweets

Dataset	**Positive score**	**Negative score**	**Total score**	**Absolute total score**
NEG	2.77	-6.40	-3.63	9.17
NEU	4.46	-3.40	1.06	7.86
PST	6.40	-2.53	3.87	8.93

After the conversion, 6631 out of 6736 tweets were assigned with sentiment scores (see Table 7.3.7). This coverage is better than that resulting from the use of Davies and Ghahramani (2011)'s Twitter sentiment analysis list, though the localised word list contains only one third of the words of their list. Most importantly, the average absolute total score based on the localised word list is more than double of the result based on the aforementioned result. This clearly demonstrates that the localised word list is much more robustly capable of identifying words in each record of data, which in turn can reduce the sparseness of the converted word matrix afterwards.

Table 7.3.7 Coverage comparison of the localised sentiment word list and the list by Davies and Ghahramani (2011)

Word list	**Coverage**	**Average absolute total score**
Twitter sentiment analysis list	6229 (92.47%)	3.25
Localised sentiment word list	6631 (98.44%)	8.75

Figure 7.3.1 presents the score distribution of each tweet based on two word lists. It is clear that the score distribution based on the localised word list is less skewed than the one based on the Twitter sentiment analysis word list. Both suggest that the localised word list has a better coverage than Davies and Ghahramani (2011)'s list.

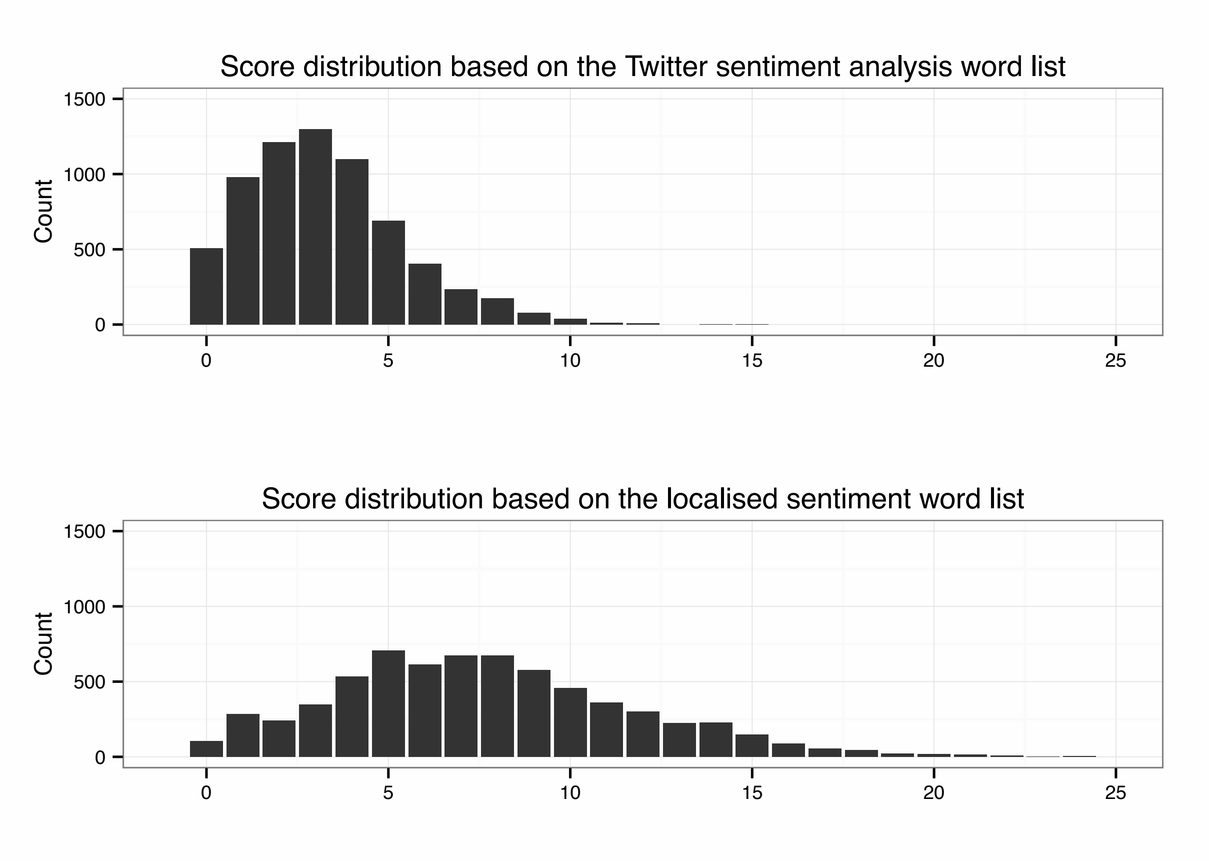


Figure 7.3.1 Score distribution comparison of the localised sentiment word list and the list by Davies and Ghahramani (2011)

7.3.4 Test of the bag-of-words model based on a localised word list

To understand if the converted word matrix can distinguish each polarity, it is better to apply statistical tests to see their difference. However, it is not realistic to test all unigrams in the localised sentiment word list in turn, so the statistical analysis is only applied to the three scores: the overall score, positive score and negative score.

The first group of tests were applied to the overall scores. Carrying out the Shapiro tests to the three subsets, the NEG group, NEU group and PST group, none of them was normally distributed (see Table C.1.1 in Appendix C.1). Therefore, the U-test was applied, and the polarity level had a significant effect ($\chi^2(2) = 1166.53$, $p < 2.2e-16$). A post-hoc test using Mann-Whitney tests with Bonferroni correction showed the significant differences between PST and NEG ($p < 2.2e-16$, $r = .612$), between PST and NEU ($p < 2.2e-16$, $r = .332$), and between NEG and NEU ($p < 2.2e-16$, $r = .571$). In other words, the overall scores across all three polarities have a significant effect of difference respectively.

Then, the statistical tests focused on the positive scores. Similarly, the Shapiro tests showed that all three subsets were not normally distributed (see Table C.1.2 in Appendix C.1). Therefore, the U-test was applied, and the polarity level had a significant effect ($\chi^2(2) = 701.391$, $p < 2.2e-16$). A post-hoc test using Mann-Whitney tests with Bonferroni correction showed the significant differences between PST and NEG ($p < 2.2e-16$, $r = .483$), between PST and NEU ($p < 2.2e-16$, $r = .279$), and between NEG and NEU ($p < 2.2e-16$, $r = .302$). These results suggest that the positive scores across all three polarities have a significant effect of difference respectively.

Finally, the statistical tests were carried out to the negative scores. Again, the Shapiro tests showed no normal discussion among the three subset (see Table C.1.3 in Appendix C.1). Therefore, the U-test was applied, and the polarity level had a significant effect ($\chi^2(2) = 632.414$, $p < 2.2e-16$). A post-hoc test using Mann-Whitney tests with Bonferroni correction showed the significant differences between PST and NEG ($p < 2.2e\text{-}16$, $r = .481$), between PST and NEU ($p = 5.93e\text{-}16$, $r = .154$), and between NEG and NEU ($p < 2.2e\text{-}16$, $r = .464$). These results suggest that the negative scores across all three polarities have a significant effect of difference respectively.

To summarise, all three sentiment scores generated from the converted word matrix present significant effects among each groups of tweets (see Table 7.3.8), so they may help to identify the polarity in an automatic way in later experiments.

Table 7.3.8 Summary of significance tests of three groups of sentiment scores

Features	**Polarity level**	**PST & NEG**	**PST & NEU**	**NEG & NEU**
Overall score	$\chi^2(2) = 1166.531$ ***	$r = .612$ ***	$r = .332$ ***	$r = .571$ ***
Positive score	$\chi^2(2) = 701.391$ ***	$r = .483$ ***	$r = .279$ ***	$r = .302$ ***
Negative score	$\chi^2(2) = 632.414$ ***	$r = .481$ ***	$r = .154$ ***	$r = .464$ ***

* indicates that $p < 0.05$, ** indicates that $p < 0.01$, and *** indicates that $p < 0.001$

7.3.5 Discussion

This section described the use of the Porter stemmer to conflate words according to their stems, which resulted in a reduction of the number of word types within the dataset to only one third of the original count. This process mainly conflated words with a lower frequency and therefore resulted in an increase in the density of the resulting word matrix. This also showed that tweets often contain variant spellings, which supports previous studies (Bar-Haim et al., 2011; Go & Bhayani, 2010; Zappavigna, 2012). Subsequently it described an application of keyness analysis in order to identify characteristic keywords within positive or negative polarities. The results presented a more distinct word list than the unigram frequency list in the last section. Additionally, the results were found to exhibit less sensitivity to the topic of the tweets, which suggested that this approach might usefully be applied to tweet data with other ticker names. However, some keywords occurred in both polarities, as is the case with mutual keywords. These keywords might confuse the automatic classifier to assign the right score. To reduce the impact of mutual keywords, a comparison of the keyness rank of mutual keywords was used in order to classify term polarities.

Finally, a smaller but more robust localised sentiment word list was generated. The improved robustness of this resource can be seen from three aspects: (a). it covered 98.44% of the annotated stock tweets. This coverage is much better than pre-designed sentiment word lists based on datasets other than tweet data; it also outperformed the Twitter sentiment analysis list by Davies and Ghahramani (2011). (b). it identified more unigrams in each individual tweet. With a higher score of each tweet, it can then generate a less sparse word matrix. (c). it displayed distinct differences across different polarities. As the statistical tests suggested, the sum of negative unigrams, of positive unigrams, and of the combined set showed a significant effect respectively, so they may help to develop an automatic classifier in later experiments.

In summary, this section presented a new approach to generate a sentiment word list, which provides better coverage than other lists. This approach consisted of three steps: 1. using a Porter stemmer to conflate the corpus; 2. using log-likelihood to calculate the keyness of the stemmed unigrams; 3. classifying mutual unigrams according to the keyness rank.

7.4 Sentiment Classification Based on the Bag-of-words Model

Previous sections of this chapter examined different frequency lists of the annotated tweets, ascertaining that unigrams in a keyness list generated using stemmed words are indicative of sentiment. Building on this work, the following section develops an automatic approach to the identification of sentiment in the ticker-related tweets based on the identified feature set. Four different machine learning methods introduced in Chapter 6 were used to classify the annotated GE tweets based on the above word matrix. The effectiveness of these methods at accurately classifying a sample tweet corpus is then tested using cross-validation (see Section 6.3). In addition, the four classifiers were explored further by testing them on tweets about a different company (the Caterpillar company – CAT). This process provided an indication of the validity of the classifiers and offered possible insights into how they might perform on a larger scale.

Each automatic classification process used the converted word matrix described in Section 7.3.3 to train the classifiers. The work described in the previous section yielded a word matrix with 6736 rows and 2344 columns. 3310 of the rows represent ticker-related tweets. 500 rows of each polarity were randomly selected to compile a training set, hereafter referred to as the GE training set. The remaining 1810 tweets were used to compile a test set, namely the GE test set. Taking a closer look at this training set and test set, the average score in the training set is very balanced (see table 7.4.1). However, in the test set, the score is skewed. This is because in the training set, all three polarities have the same number of tweets, while in the test set, the PST group has 1554 tweets, and the NEG group only has 54 tweets. The balanced score in the test set shows that the localised sentiment word list can provide a far less biased bag-of-words matrix than the other lists tested.

Table 7.4.1 Statistics of average score of GE training and test set

Dataset	Positive score	Negative score	Total score	Absolute total score
Training set	4.34	-4.31	0.03	8.65
Test set	6.25	-2.59	4.66	8.84

The second column of Table 7.4.2 shows the 10-fold cross validation results of four supervised learning classifiers. Four classifiers had very similar performance that achieved reasonably good results. In each case, an accuracy of around 0.6 was achieved. The support vector machine slightly outperformed the other three classifiers as its accuracy surpassed 0.6. As shown in the third column, applying the trained classifiers to the GE test dataset, all four classifiers performed better than they had done on the training dataset, especially the decision tree, Naïve Bayes and support vector machine classifiers, all of which showed a significant improvement. However, as shown in the fourth column, applying four classifiers to the CAT tweet dataset, all of them had a poorer performance. This indicates that classification results based on the localised word list is still relatively topic-related, though less dependent, contrary to the expectations raised by the analysis described in section 7.3.3. Thus, further development is recommended before this word list is used on tweets with other topics.

Table 7.4.2 Summary of the best accuracy of four classifiers based on the bag-of-words model

Method	**Best training accuracy**	**Test accuracy of GE tweets**	**Test accuracy of CAT tweets**
Decision tree	0.583	0.747	0.473
Random forests	0.592	0.611	0.381
Naïve Bayes	0.591	0.710	0.417
Support vector machine	0.610	0.671	0.397

According to the performance of four classifiers in the training stage, the research therefore regards the sentiment classification accuracy achieved at the training stage as the baseline accuracy, which provides a benchmark for evaluating the performance of classifiers in the following analysis.

7.5 Part-of-speech Analysis

The above classification showed that a localised sentiment word list based on the keyness analysis of stemmed tweets still bounds to the topic on which it is trained. This being the case, it was decided to annotate and analyse these ticker-related tweets with part-of-speech tags. Part-of-speech closely associates to the meaning of a word, so it is therefore regarded as an internal linguistic feature. The following annotation was based on the unstemmed data, because POS taggers are unable to identify the part-of-speech of stemmed words without a complete suffix.

7.5.1 Part-of-speech annotation

According to Gimpel et al. (2011), the ARK tweet POS tagger can classify words within tweets into 25 part-of-speech categories. After annotating the unstemmed tweets, the categories *existential 'there'* and *predeterminers + verbal (Y)* do not appear in the result, and the category *URL or email address (U)* becomes meaningless as all the URL links have been replaced by a special mark *_THIS_IS_A_URL_LINK_* as discussed earlier in Section 4.3.4. In addition, the ARK tweet POS tagger took punctuation marks into consideration. In order to keep the current analysis coherent with the previous unigram analysis, this category was also removed. Hence, there were only 22 categories in the annotation results (see Table 7.5.1). Finally, the ARK tweet PST Tagger treated all cashtags as proper nouns by default, because it does not have a particular category for the cashtag format. This is a reasonable strategy, because ticker names are used as proper nouns in the stock-market context.

Table 7.5.1 Statistics of the POS tagging results of five groups of tweets by the ARK tweet POS tagger (The original tags are in brackets.)

POS tag	**NEG**	**NEU**	**PST**
Verb participle (T)	29	35	156
Verb including copula, auxiliaries (V)	1125	1139	3532
Proper noun & possessive (Z)	47	45	179
Proper noun (ˆ)	2104	2560	7610
Pronoun (O)	149	215	368
Pre- or postposition, conjunction (P)	701	904	2939
Other abbreviations (G)	106	125	292
Numeral ($)	363	380	1240
Nominal & verbal, verbal & nominal (L)	25	60	170
Nominal & possessive (S)	4	9	19
Interjection (!)	23	43	84
Hashtag (#)	372	420	1339
Existential *there*, predetermines (X)	3	1	10
Emoticon (E)	8	6	11
Discourse marker (~)	175	189	528
Determiner (D)	319	438	1157
Coordinating conjunction (&)	159	206	537
Common noun (N)	1529	1923	6045
At-mention (@)	206	241	524
Adverb (R)	224	285	643
Adjective (A)	463	646	2068

After normalising the POS counts in each polarity of tweets, several similar features can be summarised (see Figure 7.5.1). Overall, the POS tags in three groups follow Zipf's law. Also, all three categories tend to be more similar in terms of the distribution of the POS tags. The only obvious differences are in the *pronoun*, *adverb* and *verb including copula, auxiliaries* categories. Hence, it is useful to take a closer look at these similar groups.

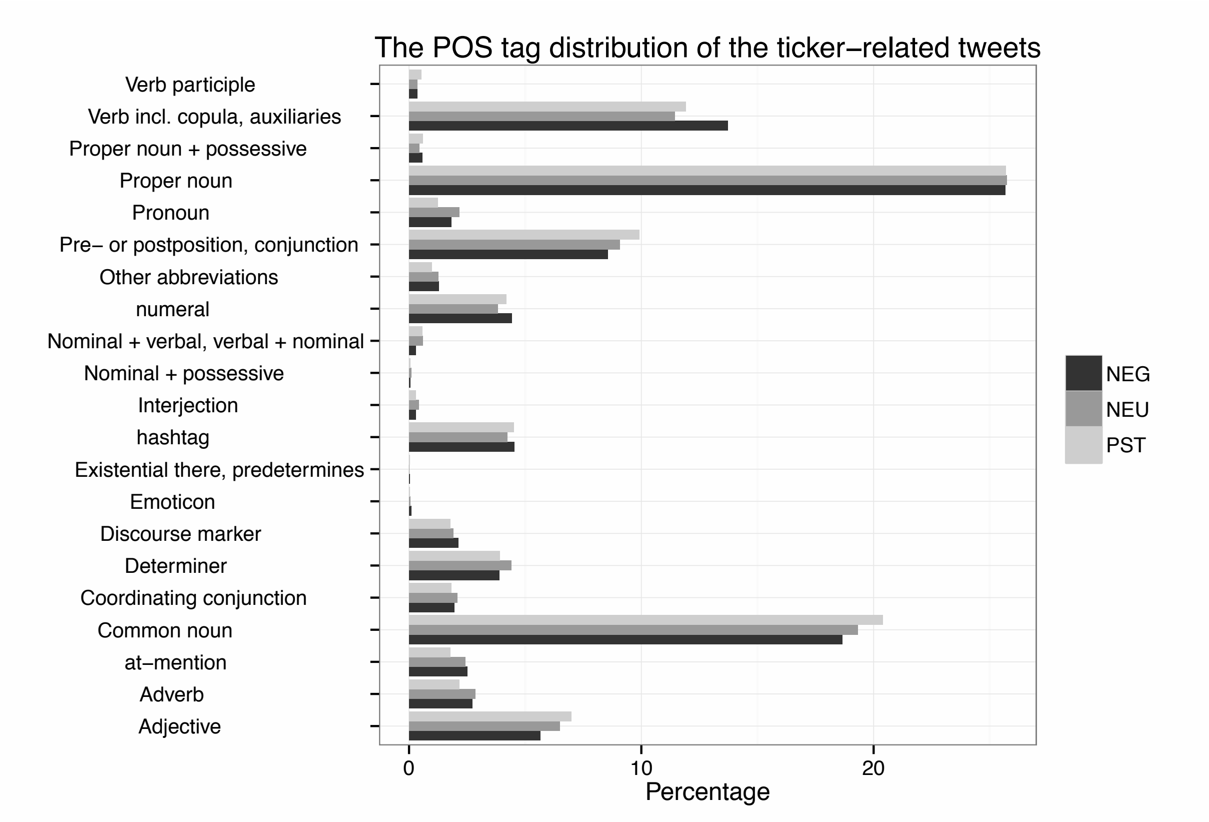


Figure 7.5.1 Distribution of POS tags of ticker-related tweets

The *proper noun* category is the largest category in all three polarities of ticker-related tweets. The next largest group is the *common noun* category, followed by these categories: *adjective*, *verb including copula, auxiliaries*, *preposition or postposition*, and *conjunction*. These five groups account for more than 70% of all tokens in the ticker-related tweets. The POS annotation also makes the differences across different groups clearer. For example, the *pronoun* category and the *proper noun* category differ across these three polarities of tweets. In addition, as superlative adverbs and possessive endings occur more often in the positive tweets (Pak & Paroubek, 2010), the following analysis focuses on the adjectives and verbs in particular. In the next section, these differences are discussed in further detail.

7.5.2 Analysis of POS tagged unigrams

The adjective *general* is the most frequent adjective in both NEU and PST group, and the second most frequent one in the NEG group, but Section 7.2 suggested that most instances of the word *general* collocate with the word *electric*. In this case, therefore, the word *general* should be considered as a part of a proper noun, instead of an individual adjective.

In general, as shown in Table 7.5.2, the NEU and PST groups share a similar frequency of these top frequently-occurring adjectives and are more condensed than the NEG group. Although the NEG group has a number of negative adjectives, such as *bearish*, *short* and *pointless*, it also contains positive adjectives at the top, for example, *fine*, *new*. In the PST group, in contrast, the most frequently-used adjectives tend to be positive.

Table 7.5.2 Top 20 frequent adjectives in the ticker-related tweets

NEG			NEU			PST		
Unigram	**Count**	**Freq.**	**Unigram**	**Count**	**Freq.**	**Unigram**	**Count**	**Freq.**
general	38	0.40	more	29	0.25	new	113	0.33
short	21	0.22	short	26	0.22	good	52	0.15
big	9	0.10	new	20	0.17	more	51	0.15
small	8	0.08	good	20	0.17	top	48	0.14
more	8	0.08	long	18	0.16	long	48	0.14
fine	7	0.07	next	13	0.11	bullish	47	0.14
close	7	0.07	technical	12	0.10	first	40	0.12
pointless	6	0.06	less	12	0.10	big	35	0.10
own	6	0.06	diversified	12	0.10	next	34	0.10
only	6	0.06	social	10	0.09	higher	32	0.09
long	6	0.06	double	10	0.09	strong	29	0.09
critical	6	0.06	big	9	0.08	short	25	0.07
biggest	6	0.06	major	8	0.07	best	25	0.07
annual	6	0.06	active	8	0.07	most	24	0.07
real	5	0.05	own	7	0.06	active	22	0.06
overweight	5	0.05	interesting	7	0.06	upside	21	0.06
last	5	0.05	second	6	0.05	industrial	21	0.06
broad	5	0.05	other	6	0.05	great	21	0.06
new	4	0.04	last	6	0.05	last	19	0.06

In terms of verbs, the NEG group has the highest density, and the PST group has the lowest (see Table 7.5.3). This suggests that tweets in the NEG group use verbs more frequently than those in the PST group. To be specific, the NEG group contains frequent usages of verbs such as *falling*, *sold*, *pay*, *sell*, and *downgraded*, while the PST group contains verbs such as *expected*, *beat* and *rise*. Though *buy* also occurs in the NEG group of the top 20 frequent list, the PST group has a much higher frequency, because it contains three variants of *buy* – *buy*, *buying* and *bought*.

Table 7.5.3 Top 20 frequent verbs in the ticker-related tweets

NEG			NEU			PST		
Unigram	**Count**	**Freq.**	**Unigram**	**Count**	**Freq.**	**Unigram**	**Count**	**Freq.**
are	39	0.41	will	34	0.29	ge	98	0.29
be	31	0.33	be	28	0.24	are	78	0.23
falling	30	0.32	are	25	0.22	buy	68	0.20
will	23	0.24	buy	21	0.18	looking	62	0.18
sold	16	0.17	have	17	0.15	be	60	0.18
have	16	0.17	do	17	0.15	expected	56	0.17
pay	14	0.15	check	17	0.15	see	54	0.16
has	14	0.15	see	16	0.14	will	51	0.15
get	14	0.15	am	16	0.14	has	42	0.12
been	13	0.14	was	14	0.12	going	41	0.12
sell	10	0.11	makes	14	0.12	get	36	0.11
moving	10	0.11	sell	12	0.10	check	36	0.11
downgraded	10	0.11	closing	12	0.10	beat	33	0.10
should	9	0.10	based	12	0.10	rise	28	0.08
did	9	0.10	valued	11	0.09	could	28	0.08
coming	9	0.10	looking	11	0.09	report	27	0.08
buy	9	0.10	trending	10	0.09	buying	24	0.07
would	7	0.07	think	10	0.09	have	23	0.07
was	7	0.07	get	10	0.09	bought	21	0.06

Although the overall distributions of POS tags are similar across the different polarities of ticker-related tweets, subtle differences may be found in individual tweets. To understand these differences, the five most frequent POS groups were analysed: adjectives, common nouns, prepositions, proper nouns and verbs. The Shapiro test suggested that none of these five groups of data is normally distributed (see tables in Appendix C.1), so the Kruskal Wallis test was used to establish whether there is a significant effect across three polarities among these five groups of data (see Figure 7.5.2).

First, a Kruskal Wallis test revealed a significant effect of the difference among adjectives ($\chi^2(2)$ = 14.433, *p*-value = 0.0007). A post-hoc test using Mann-Whitney tests with Bonferroni correction showed significant differences between NEG and PST (*p*-value = 0.0006, *r* = 1.263e-05) and between PST and NEU (*p* = 0.0151, *r* = .0003).

Then, a Kruskal Wallis test revealed a significant effect of the difference among common nouns ($\chi^2(2)$ = 7.276, *p* = 0.0263). A post-hoc test using Mann-Whitney tests with Bonferroni correction showed significant differences between PST and NEU (*p* = 0.0178, *r* = .0003).

Next, a Kruskal Wallis test revealed a significant effect of the difference among prepositions ($\chi^2(2)$ = 24.700, *p* = 4.329e-06). A post-hoc test using Mann-Whitney tests with Bonferroni correction showed the significant differences between NEG and PST (*p* = 0.0001263, *r* = 2.473e-06) and between PST and NEU (*p* = 7.104e-05, *r* = 1.353e-06).

However, a Kruskal Wallis test revealed an insignificant effect of the difference among proper nouns ($\chi^2(2)$ = 2.867, *p* = 0.2385).

Finally, a Kruskal Wallis test revealed a significant effect of the difference among verbs ($\chi^2(2)$ = 25.549, *p* = 2.832e-06). A post-hoc test using Mann-Whitney tests with Bonferroni correction showed significant differences between NEG and PST (*p* = 0.0001168, *r* = 2.287e-06), between PST and NEU (*p* = 0.01099, *r* = .0002), and between NEG and NEU (*p* = 2.225e-06, *r* = 6.278e-08).

In a nutshell, as shown in Table 7.5.4, only *verbs* showed a significant effect of variation across all subsets of tweets, and *adjectives*, *common nouns* and *proper nouns* only differentiated negative annotated tweets from their positive counterparts. These differences were much weaker than

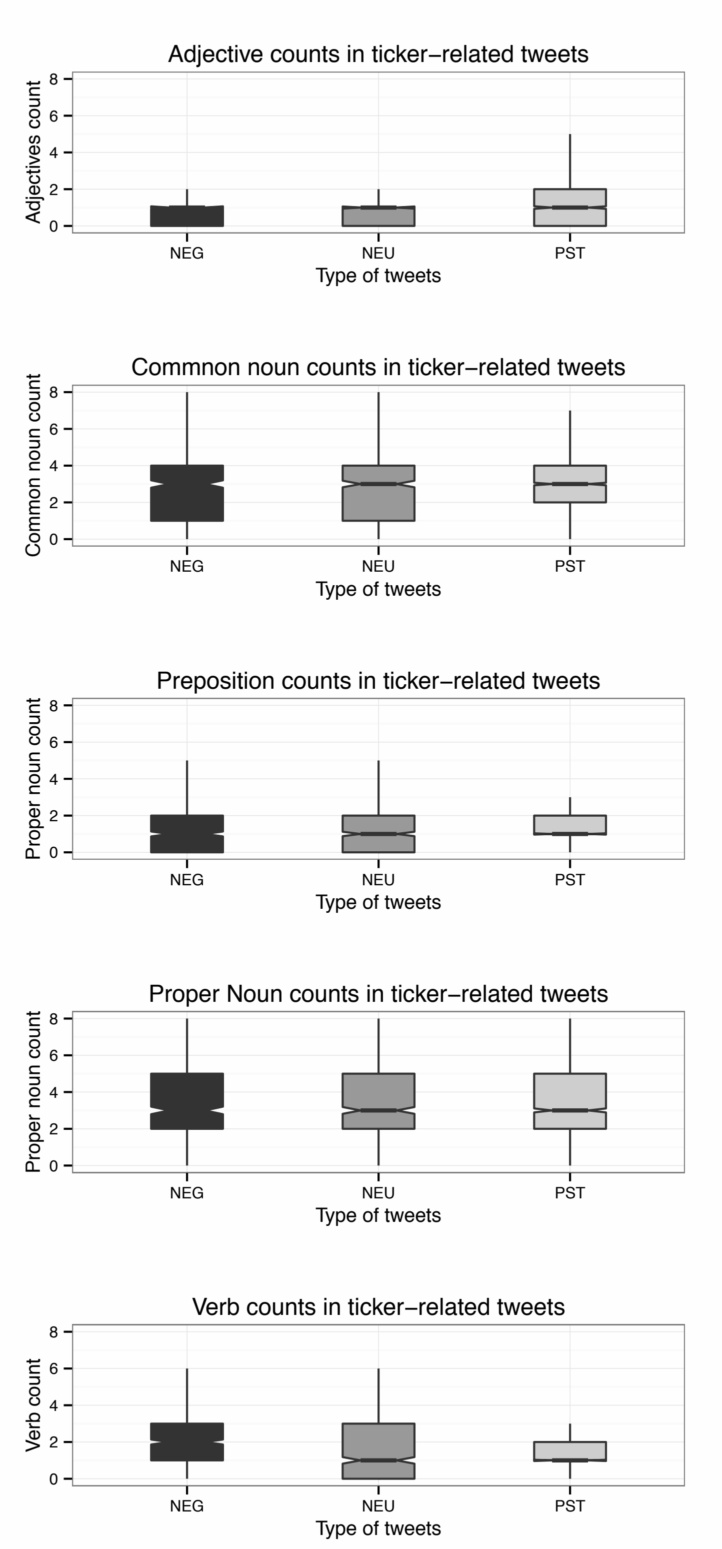


Figure 7.5.2 Five groups of the most frequent part-of-speech features in ticker-related tweets

those tested in the bag-of-words model above. In addition, *prepositions* had no significant effect of differences at all. Thus, it seemed that it would be possible, although less encouraging, to develop machine learning methods to automate the identification of the polarity of stock tweets based on the feature-set discussed here.

Table 7.5.4 Summary of significance tests of five most frequent POS tag categories

Features	**Polarity level**	**PST & NEG**	**PST & NEU**	**NEG & NEU**
Adjectives	$\chi^2(2) = 14.433$ ***	r = 1.263e-05 ***	r = .0003 ***	N.S.
Common Nouns	$\chi^2(2) = 7.276$ *	N.S.	r = .0003 *	N.S.
Prepositions	$\chi^2(2) = 24.700$ ***	r = 2.473e-06 ***	r = 1.353e-06 ***	N.S.
Proper Nouns	N.S.	N.S.	N.S.	N.S.
Verbs	$\chi^2(2) = 25.549$ ***	r = 2.287e06 ***	r = .0002 ***	r = 6.278e-08 ***

* indicates that $p < 0.05$, ** indicates that $p < 0.01$, and *** indicates that $p < 0.001$

7.6 Sentiment Classification Based on Part-of-speech Features

The above classification based on the bag-of-words model presented a reasonably good result at both the training and test stages. However, applying four trained classifiers to the CAT tweet dataset, they did not achieve as good a performance as had been achieved on the GE tweet dataset. Attempting to solve this topic-related problem, it proposed the use of part-of-speech tags as features on which to train classifiers. As shown in Section 7.4, four out of five of the most frequently used part-of-speech features – *adjective*, *common noun*, *proper noun*, and *verb* – exhibited significant variation across different polarities. Based on this, in the following section, it described to use these feature to train the classifiers.

The four most frequent features within the part-of-speech terms assigned – *adjective*, *proper* *noun*, *common noun* and *verb* – were extracted and converted to a matrix similar to that used in the previous analysis (see table 7.6.1). Each cell represents the occurrence count of a feature, and each line comprises a record for an individual tweet. The first column indicates the type of tweets, and the remaining columns indicate different part-of-speech features.

Table 7.6.1 A sample set of POS features generated from annotated GE ticker-related tweets

Type	**Adjective**	**Common Noun**	**Proper Noun**	**Verb**
NEG	1	1	1	3
NEG	2	1	1	0
NEG	1	0	6	2
NEG	1	3	5	0
NEG	2	1	1	1
NEG	1	2	1	0

Applying the four supervised learning methods to train the classifiers, the best accuracy of the 10-fold cross validation results are presented in Table 7.6.2. Among them, at the training, the random forests classifier outperformed the other three classifiers with an accuracy of 0.468, and the other three exhibited similar, poorer performances. Surprisingly, the decision tree and Naïve Bayes classifiers achieved better performances at the test stage with GE stock tweets. The decision tree classifier in particular exhibited the greatest improvement and the best accuracy. However, applying the trained classifiers to the CAT dataset, all of them failed to classify the polarity.

Table 7.6.2 Summary of the accuracy of four classifiers based on the most frequent part-of-speech features

Method	Best training accuracy	Test accuracy of GE tweets	Test accuracy of CAT tweets
Decision tree	0.397	0.535	0.112
Random forests	0.468	0.378	0.200
Naïve Bayes	0.397	0.415	0.151
Support vector machine	0.425	0.414	0.174

Adding all POS features seems to be a poor choice, as they did not improve the overall accuracy at the training stage (see table 7.6.2). In machine learning literature, this is known as the overfitting problem. What is worse is that, at the test stage, all four classifiers exhibited a clear drop in accuracy. Therefore, in the later classification, it is proposed that only five most frequent POS features are used to train the classifier. Again, none of the trained classifiers can classify the CAT tweets efficiently.

Table 7.6.3 Summary of the accuracy of four classifiers based on the use of all frequent part-of-speech features

Method	Best training accuracy	Test accuracy of GE tweets	Test accuracy of CAT tweets
Decision tree	0.391	0.359	0.230
Random forests	0.553	0.438	0.197
Naïve Bayes	0.397	0.160	0.192
Support vector machine	0.505	0.438	0.174

This section showed that, regardless of the significant effects exhibited in some part-of-speech features of stock tweets, using solely POS features to train an automatic classifier cannot beat baseline accuracy, especially when they are applied to stock tweets with another ticker name.

7.7 Discussion

This chapter focused on Research Question 4: *How to design a more robust sentiment list for tweet sentiment classification?* It first tested a number of predesigned sentiment word lists that have been used extensively in previous studies, but none of which were found to satisfy the needs of this research. The main problem is that these lists cannot provide an adequately extensive coverage of tweets, so they would not be sufficient to identify sentiment keywords in tweets. Similarly, a bag-of-words model based on such lists would exhibit a poor performance in the later training. This is a serious problem, but no previous research has addressed this. Previous research in this area focused on the selection of a widely used word list, and the conversion of tweet data to a word matrix according to such a list. The bag-of-words approach is often regarded as one of the most fundamental approaches in computer science, but, at least in the stock tweet sentiment analysis context, previous studies have paid little attention to this approach.

The only exception can be found in Debbini et al. (2011). They used 1000 most frequent words taken from their data to train the bag-of-words model. Following this approach, this chapter analysed the most frequent unigrams in the tweet data. The main problem identified was that these most frequent words appear in all three polarities, so they could, to a limited extent, indicate any distinct features of polarity. This shows that Debbini et al. (2011)'s approach is also not a good choice. Specifically, the top frequent unigrams show a high degree of similarity between the three polarities of tweets: there are a number of ticker names and stock-related unigrams among them, and the time-relevant unigrams indicate that the ticker-related tweets focus on short-term discussions. Although the stock-related unigrams are different in the three subgroups, the overlaps between them and their generally low frequency limit their usefulness for identifying the polarity of tweets. Furthermore, the most frequent unigrams in each polarity of tweets are similar, and therefore also offer little help in identifying tweet polarity.

In an attempt to address these shortcomings, this chapter then used a reverse approach to investigate the least frequent unigrams in each polarity of tweets. The analysis of unique unigrams, in contrast, does seem to show more potential for identifying the polarity of tweets, as unique unigrams cover nearly 75% of ticker-related tweets. However, a large percentage of unique unigrams only occur once, which suggests that they might not be stable. Thus, while using unique unigrams can be regarded as a possible solution for identifying the sentiment of tweets, a larger corpus than the one used in the current research would be needed in order to provide the best chance of obtaining a stable unique unigram list.

The next attempt to improve the result involved using a Porter stemmer to reduce the number of word types in the annotated tweet corpus, and using keyness analysis to identify the keywords in the positive and negative category respectively. Through stemming, two thirds of word types were reduced, which resulted in a sentiment word list of a much smaller size, and in turn, the generated word matrix would be less sparse. Keyness analysis provided two distinct word lists for the negative and positive category respectively. One problem with the use of keyness analysis is that it contains a number of mutual keywords in both categories. To reduce the ambiguousness of mutual keywords in later automatic classification, it then used the relative rank of items within the keyness list to classify these mutual keywords. The result provided a less biased sentiment word list. More importantly, applying statistical tests to the positive, negative and overall scores of this converted word matrix, all of them showed a significant effect of difference across all subsets, suggesting that these scores can distinguish different polarities from each other. This might offer great help to the later automatic sentiment classification. Compared with Davies and Ghahramani (2010)'s list, the localised sentiment word list provides a better coverage of tweets, and more importantly, an improved coverage of keywords in each individual tweet. Obviously, this approach is also easy to apply. Thus, this attempt provided a much better approach to generate a sentiment word list than others.

This chapter then described the application of four supervised learning methods to ticker-related tweets in order to classify the sentiment. Focusing on the keywords in the positive and negative polarity, it trained the classifiers accordingly. The sentiment classification based on the bag of words model achieved fairly good results with four supervised learning methods at the training stage, which is an accuracy of around 0.6, and these results were then regarded as the baseline accuracy for the following classification. Even better, all four classifiers had a better performance at the test stage. However, applying these four classifiers to another dataset – the CAT tweet set, all of them exhibited a poorer performance. This suggests that, although the localised sentiment word list is less biased than other word lists, it is still sensitive to the topic.

It then applied four supervised learning methods to the ticker-related tweets in order to classify the sentiment. Focusing on the keywords in the positive and negative polarity, it trained the classifiers accordingly. The sentiment classification based on the bag-of-words model achieved fairly good results with four supervised learning methods at the training stage, which is an accuracy of around 0.6, and these results were then regarded as the baseline accuracy for the following classification. Even better, all four classifiers had a better performance at the test stage. However, applying these four classifiers to another dataset – the CAT tweet set, all of them had a poorer performance. This suggests that, although the localised sentiment word list is less biased than other word lists, it is still sensitive to the topic.

To tackle the topic-dependent problem, the last segment in this case study dealt with the use of part-of-speech taggers of speech features. Using the ARK tweet POS tagger to annotate ticker-related tweets, the five most frequent POS categories were found to be *adjectives*, *common nouns*, *prepositions*, *proper nouns* and *verbs*. Applying significance tests on these five features, only *verbs* presented a significant effect of differences among all subsets of tweets, *adjectives* and *prepositions* were found to be significantly different between NEG and PST groups, and *adjectives*, *prepositions* and *common nouns* were found to be significantly different between NEU and

PST groups. Thus, these features can be used to develop an automatic classifier to identify the sentiment of ticker-related tweets. This finding suggests that these features could be used to develop an automatic classifier to identify the sentiment of ticker-related tweets. Thereafter, the chapter switched the focus to the discussion of part of speech features. However, the sentiment classification based on part of speech features did not outperform the baseline. According to this, classifying sentiment of stock tweets solely based on the part of speech features. It is likely to result in low reliability of classification.

7.8 Summary

This chapter answered Research Question 4: How can a robust sentiment word list for tweet sentiment classification be designed?. By a keyness analysis based on the stemmed tweets, it successfully developed a localised sentiment word list with a better coverage, and less biased polarity. This approach is also shown to be easy to apply.

Based on the above sentiment word list, four supervised learning classifiers were then trained, and all of them achieved a reasonably good result at the training stage, and a slightly better test stage. The accuracy achieved at the training stage was then regarded as the baseline accuracy for the following sentiment classification.

However, one problem still existed that the sentiment classification based on a localised sentiment word list is relatively topic dependent. To solve this problem, the chapter then tried to classify tweets according to part-of-speech features. Nevertheless, this approach did not work as had been expected.

Apart from the above classification, this chapter also analysed internal linguistic features of stock tweets extensively, including the most and least frequent unigrams, and five most frequent part-of-speech features. Except the most frequent unigrams, the other internal linguistic features were all indicative to polarity. However, as shown, the least frequent unigrams difficult to apply as they were not stable, while the most frequent part-of-speech features did not perform as well as the bag-of-words model in the classification. The analysis presented in the following chapter tests external linguistic features, in order to improve sentiment classification over the baseline accuracy.

Chapter 8 Sentiment Classification Based on the External Linguistic Features

As shown in the last chapter, the bag of words model based on keyness analysis of the stemmed tweets provided a reasonably good performance for classifying the polarity of stock tweets, and the accuracy of those classifications was therefore regarded as the baseline accuracy. However, experiments also demonstrated that these classifiers did not perform well when they were applied to tweets with another ticker name. In addition, the baseline accuracy may still have room to improve. Furthermore, as discussed in Chapter 3 and 4, the collected raw tweets contained a large proportion of irrelevant tweets, but how to identify them has not been investigated. Thus, this chapter focuses on the low-level linguistic features of stock tweets, considered as the **external features** in this study, in order to answer Research Question 2. Are stock tweets a linguistically distinct type of tweet? And what specific linguistic features do they have? Research Question 3 is also relevant: If stock tweets display explicit linguistic features, is it possible to automatically identify stock tweets based on their linguistic features? In other words, this chapter explores whether these features can help to identify the type of a given tweet, hence resulting in an improved ability to automatically establish the polarity of stock tweets.

The first part of this chapter reports on analysis of the external linguistic features of the annotated tweets – word count, character count, retweet count, hashtag and cashtag count. The statistical analyses show that these features are able to indicate whether these tweets are (a). stock tweets or not, (b). ticker tweets or not. Next, variation in external linguistic features is evaluated at the polarity level, with the finding that tweets in different polarities can be differentiated solely by word count, excluding URLs and retweet count.

Based on these statistical analyses, the second part of this chapter returns to the study described in the previous chapter, using different supervised learning methods to develop automatic classifiers to identify stock-related and ticker-related tweets. The classification results show that the external features can be used to identify stock-related tweets very well, and can identify ticker-related tweets with a fair accuracy, but that classifiers trained using these features fail to identify the polarity of the ticker-related tweets.

8.1 External Linguistic Features of Stock Tweets

This section hypothesises that different groups of the annotated tweets present different tweeting behaviours. Such differences might be useful for the automatic classification of tweet type and polarity. To test this hypothesis, different types of annotated GE tweets are compared based on external linguistic features (word count, character count, retweet count, hashtag count, cashtag count and URL count). The analyses are carried out at three levels respectively, namely the stock level, ticker level and polarity level. These are shown in Table 8.1.1. The stock-related group contains those tweets classified as non-ticker-related (NTR), negative (NEG), neutral (NEU) and positive (PST) groups. The ticker-related group contains the NEG, NEU and PST groups.

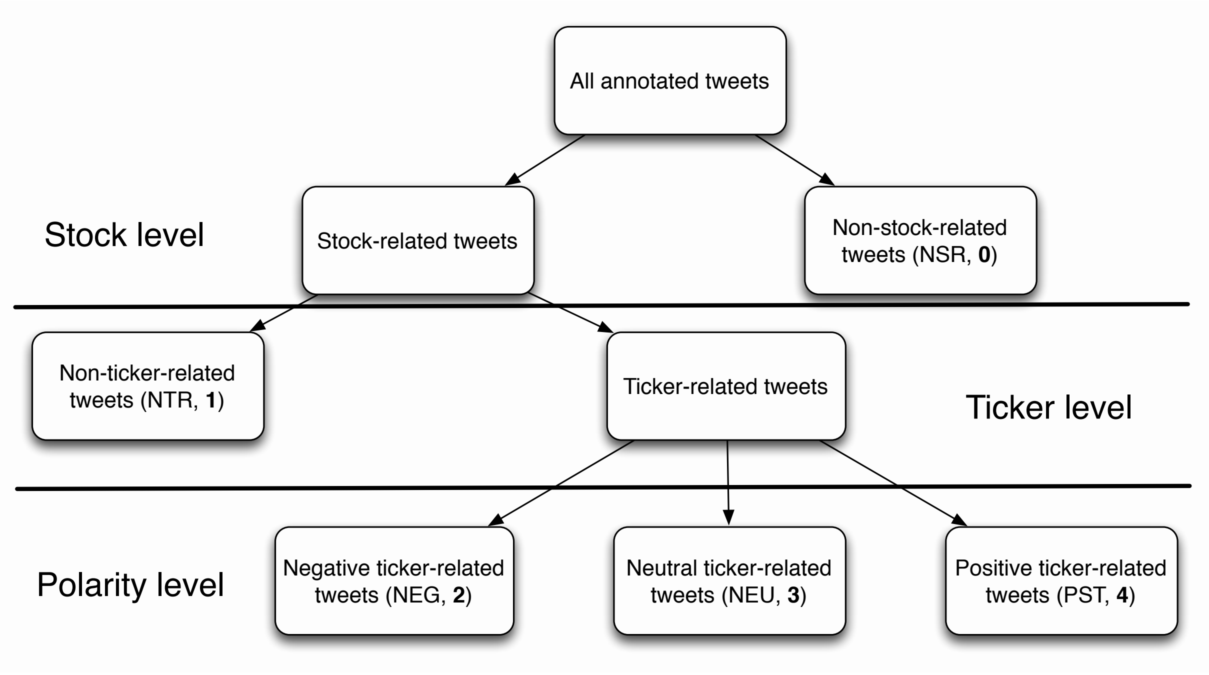

All annotated tweets
Stock level
Stock-related tweets
Non-stock-related tweets (NSR, **0**)
Non-ticker-related tweets (NTR, **1**)
Ticker-related tweets
Ticker level
Polarity level
Negative ticker-related tweets (NEG, **2**)
Neutral ticker-related tweets (NEU, **3**)
Positive ticker-related tweets (PST, **4**)

Figure 8.1.1 The relationship of annotated tweets as introduced in Chapter 4

8.1.1 Word count including URLs

The first test evaluates differences in word count including URLs. As shown in Figure 8.1.2, the average word count including URLs of stock-related tweets is 15.56 words, while the NSR group's is 14.07 words. The differences at the ticker level and polarity level are much smaller. Tests for statistical significance were applied to assess whether these differences are greater than would be expected by chance.

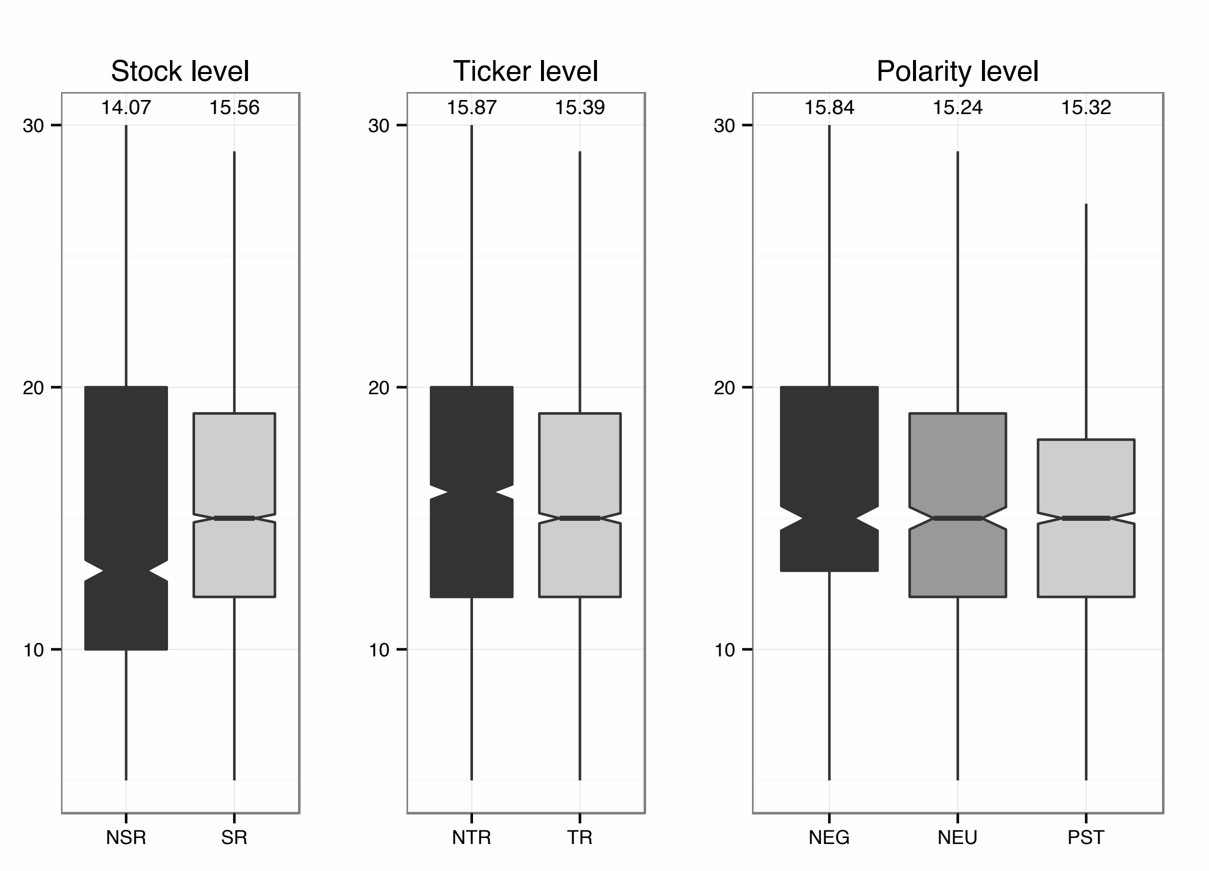


Figure 8.1.2 Boxplot of word count including URLs

The Shapiro and Kolmogorov-Smirnov tests showed that these seven groups of data were not normally distributed as the *p*-values ≤ .05 (see Tables in Appendix C.2). Therefore, the *U*-test was first applied to understand whether the differences are statistically significant. At the stock level, the NSR group and stock-related tweets were statistically significant in terms of word count including URLs (Z = -9.7837, p < 2.2e-16). Similarly, at the ticker level, the NTR and ticker-related tweets were statistically significant (Z = 4.7407, p = 2.13e-06). At the polarity level, a Kruskal Wallis test revealed an insignificant effect of the difference among word count including URLs ($\chi^2(2)$ = 5.9557 , p = 0.0509).

Thus, word count including URLs has a significant effect of differences at both the stock and ticker levels. This suggests that word count including URLs might be useful in classification of the tweet types at both the stock and ticker level.

8.1.2 Character count including URLs

As for the average character count including URLs, the stock-related tweets are much longer than the NSR tweets (see Figure 8.1.3). Also, the NTR tweets are 4 characters longer than the ticker-related tweets on average. However, at the polarity level, the three polarities of ticker-related tweets do not exhibit much difference.

None of the data was normally distributed as the Shapiro tests and Kolmogorov-Smirnov tests suggested (see Tables in Appendix C.2). The *U*-tests indicated that the differences were statistically significant at the stock level (Z = -15.104, p < 2.2e-16), and at the ticker level (Z = 6.0474, p = 1.472e-09). In brief, on average, the annotated stock-related tweets were significantly longer than the non-stock-related tweets, and the non-ticker-related tweets were significantly longer than the ticker-related tweets. The Kruskal-Wallis rank sum test was then applied to the polarity level, and it indicated that the difference between the ticker-related tweets was not statistically significant

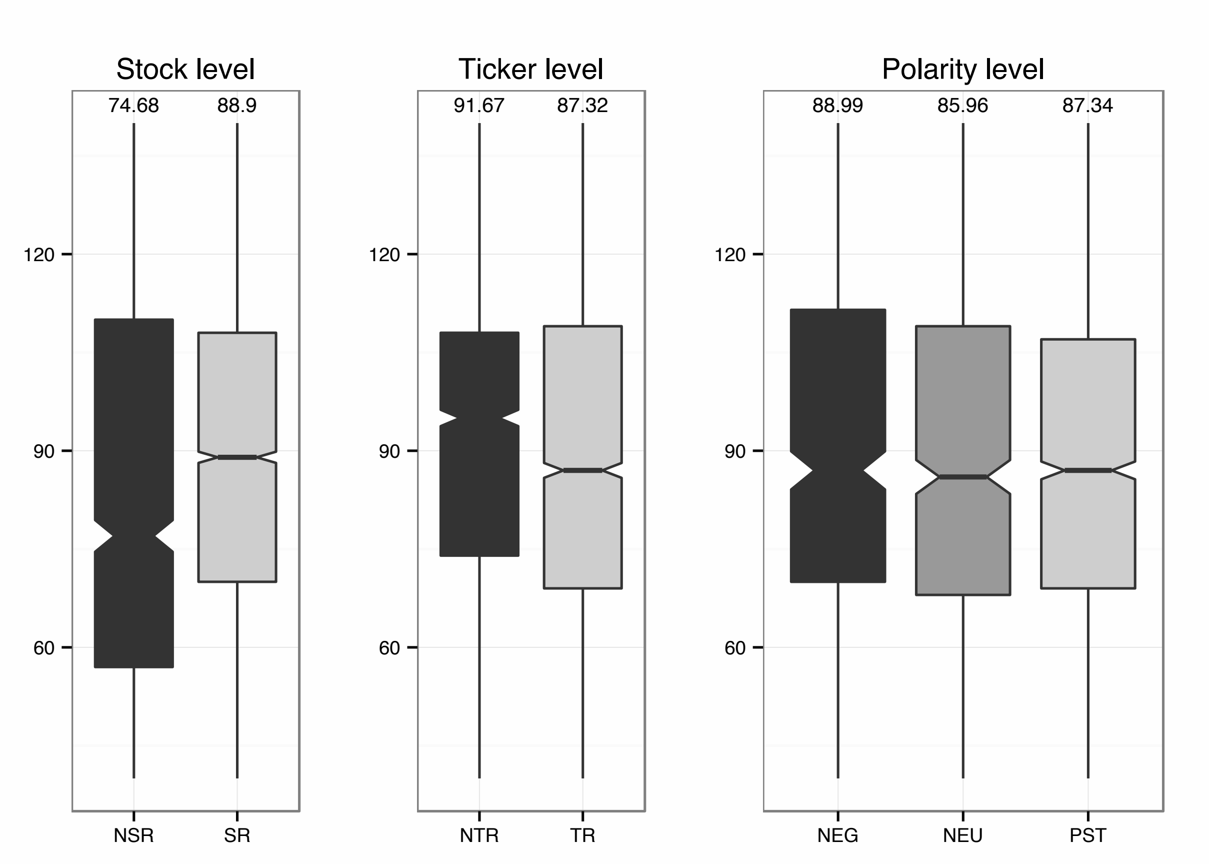


Figure 8.1.3 Boxplot of character count including URLs

($\chi^2(2)$ = 2.8142, p = 0.2449).

Therefore, character count including URLs has a significant effect at both stock and ticker levels. This suggests that character count including URLs might be a useful feature for classifying the tweet types at both stock and ticker levels.

8.1.3 Word count excluding URL

Although the initial processing in Chapter 4 used a special URL tag *_THIS_IS_A_URL_LINK_* to replace the shortened links, the high frequency of this special tag can affect the accuracy of the word type count method applied. Removing the URL tags from the annotated tweets (see Figure 8.1.4) produces results that differ from those generated by processing tweets including URL tags in the above analysis. Tweets in the NSR group still have the shortest average word count, but the difference between the NSR and the stock-related group is smaller than the difference of the above average word count including URLs.

Due to the non-normally distributed data (see Tables in Appendix C.2), the U-test was applied. The U-test results suggested that the difference at the stock level was significant (Z = -6.2215, p = 4.925e-10), and as well as at the ticker level (Z = 3.5144, p = 0.0004408). Thus, excluding URL tags, the average word count of the stock-related tweets was still significantly longer than the non-stock-related tweets, and as well as the non-ticker-related tweets than the ticker-related tweets. Carrying out the Kruskal-Wallis rank sum test at the polarity level, the difference was significant ($\chi^2(2)$ = 6.5805, p = 0.03724). Therefore, a post-hoc test of Kruskal-Wallis rank sum test – the Mann-Whitney test with Bonferroni correction was applied, showing a significant difference between the PST and NEG groups (p = 0.037, r = 0.04901176).

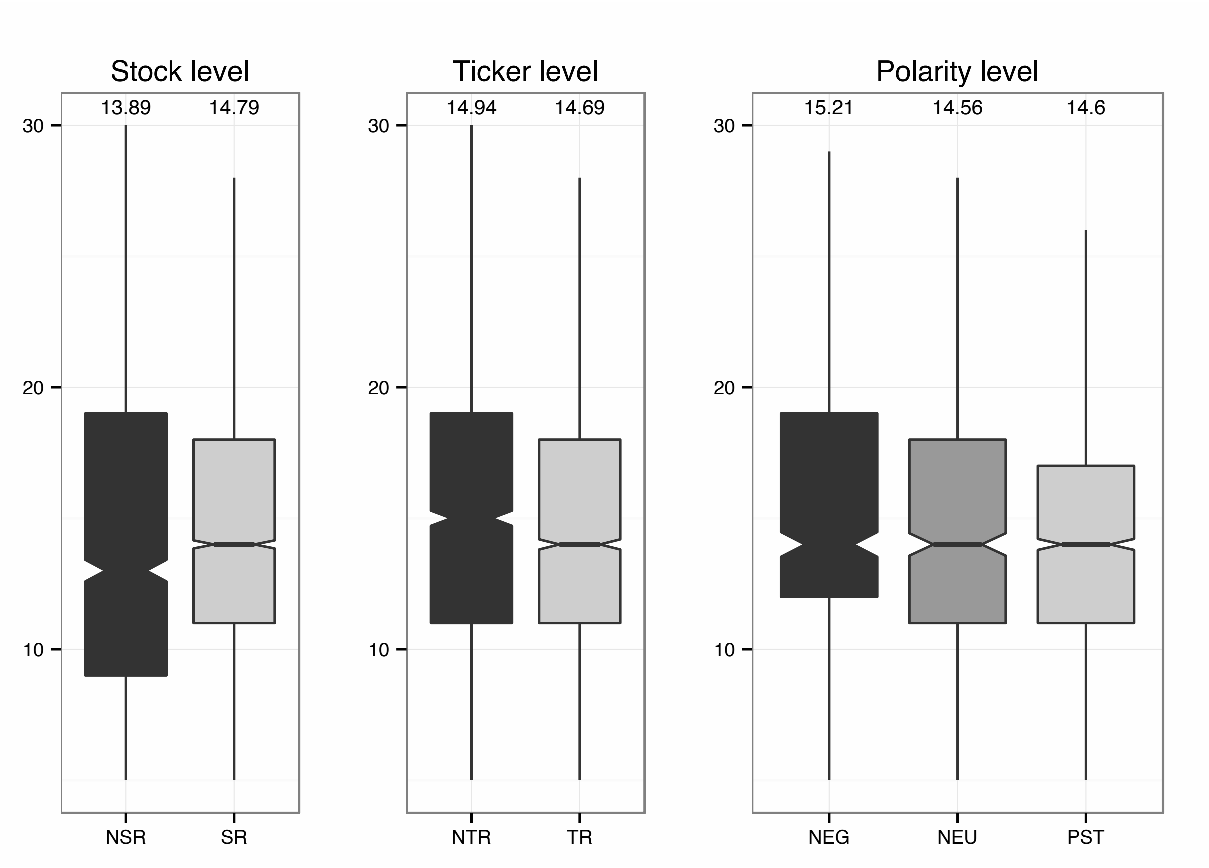


Figure 8.1.4 Boxplot of word count excluding URLs

In a nutshell, removing URLs, word count still differentiates the NSR tweets from the stock-related tweets, and the NTR tweets from the ticker-related tweets. In addition, at the polarity level, between the PST and NEG tweets, a significant difference is observable.

8.1.4 Character count excluding URL

Similarly, removing the URL tags, the NSR group still has the shortest average character count (71.18 words), though the difference between the NSR and stock-related tweets is smaller than when the URL tags are included (see Figure 8.1.5).

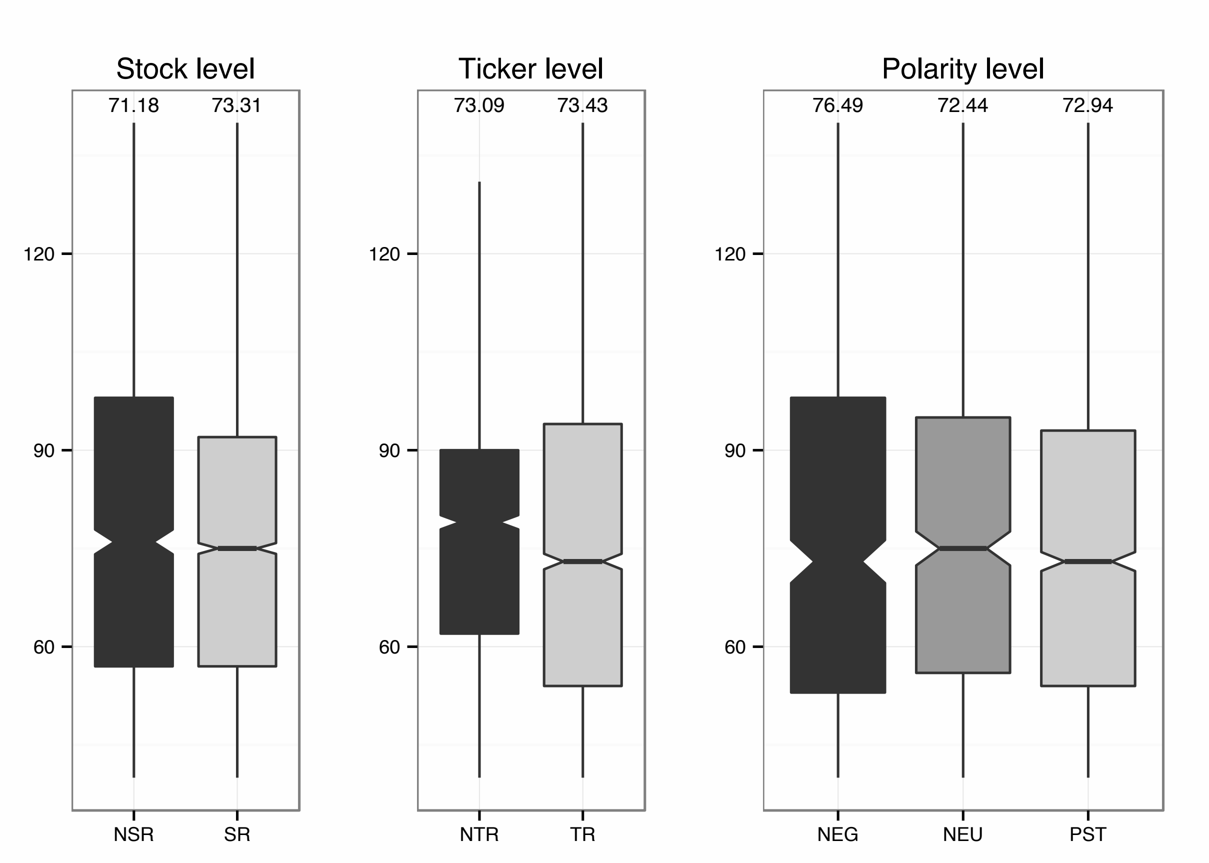


Figure 8.1.5 Boxplot of character count excluding URLs

Again, the Shapiro and the Kolmogorov-Smirnov tests showed that the data were not normally distributed (see Tables in Appendix C.2). Applying the U-test, the stock level displayed a significant effect, (Z = -3.177, p = 0.001488), but not the ticker level (Z = 1.8749, p = 0.06081). As a result, the stock-related tweets were significantly longer than the non-stock-related tweets in terms of the average character count excluding URLs. Applying the Kruskal-Wallis rank sum test to the ticker-related tweets, the difference was not significant ($\chi^2(2)$ = 3.182, p = 0.2037).

Thus, character count excluding URLs has a significant effect only at the stock level. This suggests that the classification of the tweet types at the stock level might use character count excluding URLs as a useful feature.

8.1.5 Hashtag count

In terms of the average count of hashtags, seven groups of tweets present a similar trend (see Figure 8.1.6).

Both the Shapiro test and Kolmogorov-Smirnov test suggested that the data were not normally distributed (see Tables in Appendix C.2). The U-tests indicated that there was no significant effect at either the stock level (Z = -0.7225, p-value = 0.47), or the ticker level (Z = 0.0972, p-value = 0.9226). Also, the difference in the ticker-related tweets was not significant, as the Kruskal-Wallis rank sum test indicated ($\chi^2(2)$ = 5.35, p = 0.06891).

Simply put, the hashtag count does not display any significant variation at any level, so this feature cannot indicate the type of stock tweets.

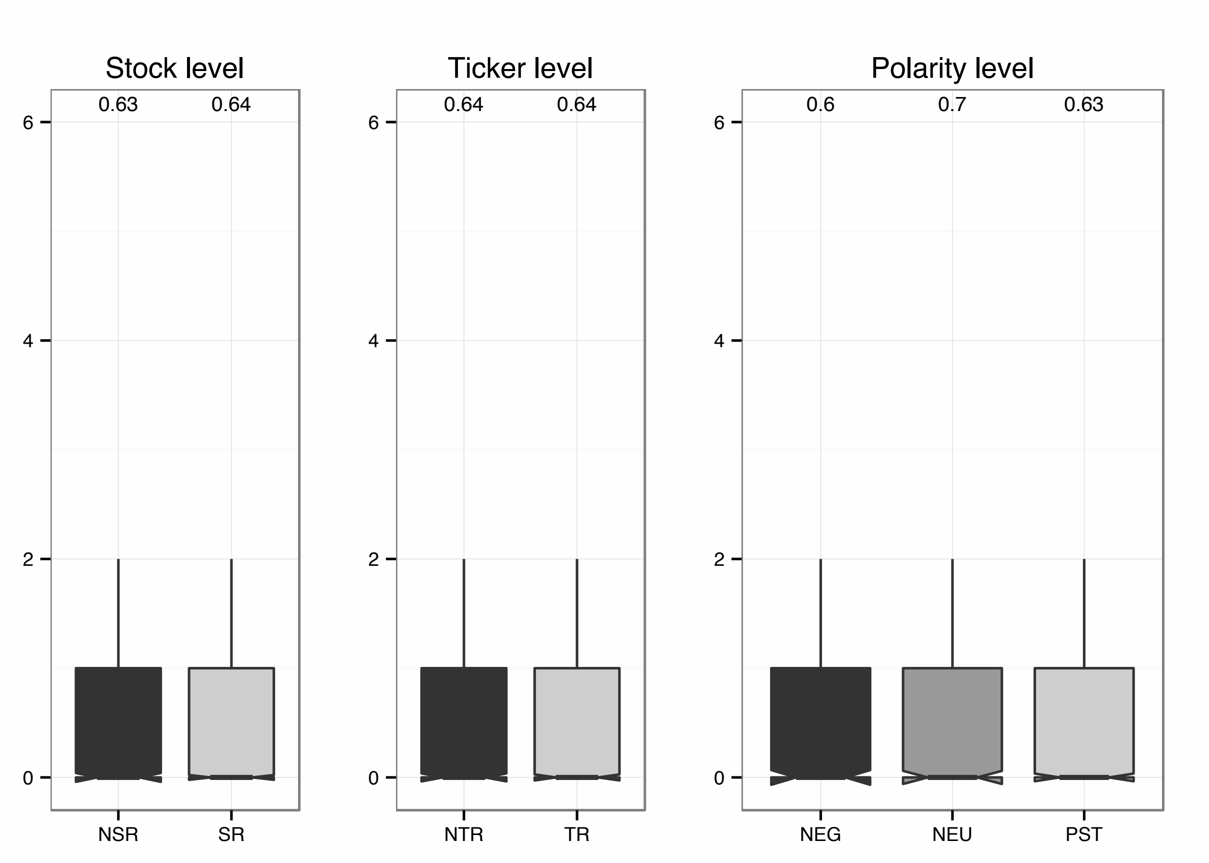


Figure 8.1.6 Boxplot of hashtag counts in the annotated tweets

8.1.6 Cashtag count

Obviously, each tweet in the annotated sample contains at least one cashtag *$GE* as a consequence of the collection criteria suggested in Chapter 4. As shown in Figure 8.1.7, the stock-related tweets contain more cashtags than the NSR tweets on average. The differences between the NTR and ticker-related groups, and among the three polarities are much smaller.

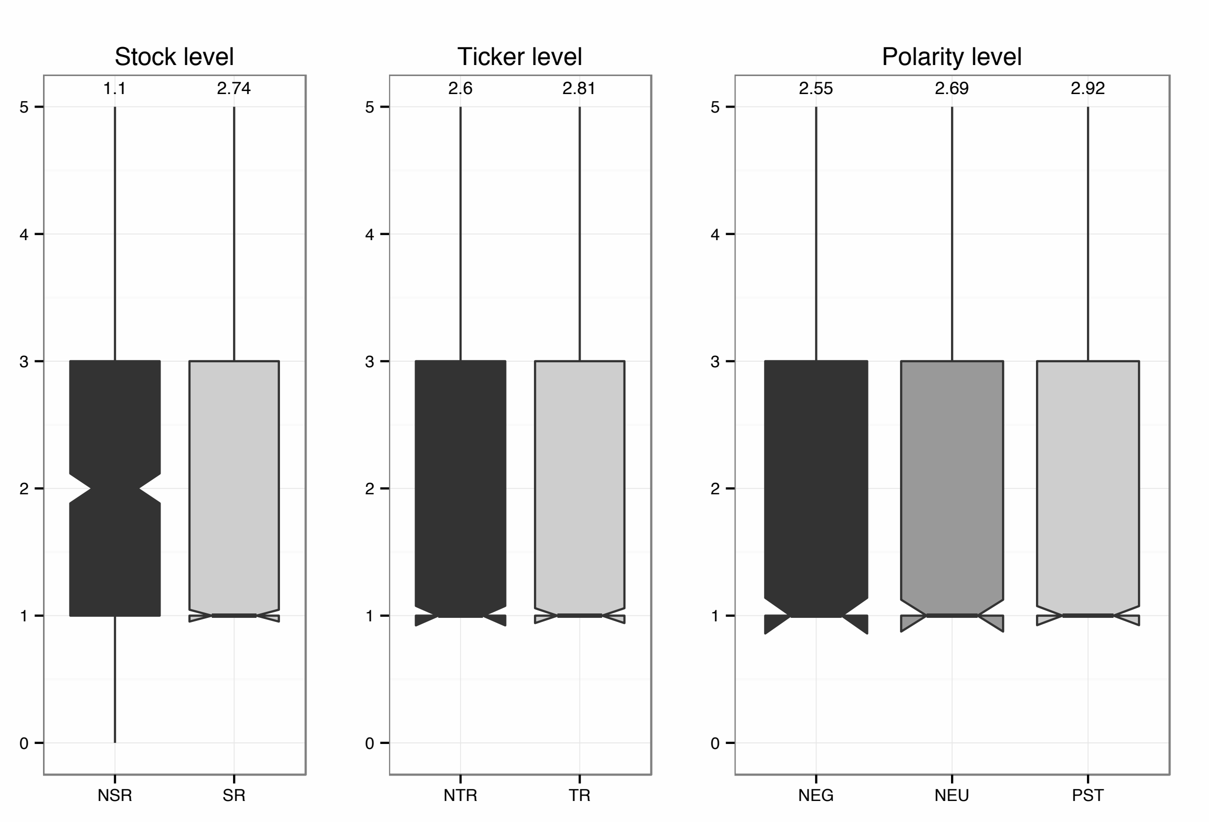


Figure 8.1.7 Boxplot of cashtag counts in the annotated tweets

Applying the Shapiro test and Kolmogorov-Smirnov test, all sets of data were not normally distributed (see Tables in Appendix C.2). Applying the *U*-test to the stock and ticker level, cashtag count was statistically different at both levels (Z = 46.1552, p < 2.2e-16, Z = 43.053, p-value < 2.2e-16). Thus, on average, both the stock-related tweets and ticker-related tweets contained more cashtags than their counterparts. However, the difference between the ticker-related tweets was not as significant as the Kruskal Wallis test suggested ($\chi^2(2) = 1.5472$, $p = 0.4614$).

Therefore cashtag count has a significant effect at both stock and ticker levels. This suggests that cashtag count might be useful in classification of the tweet types at both the stock and ticker level.

8.1.7 Retweet count

Figure 8.1.8 illustrates the statistics of retweets in different types of tweets. The NSR group contains much more retweets than other groups on average. A possible reason for this observation is that some of these tweets particularly focus on similar topics, for example, a music group, or Twitter's announcement in August, 2012. As a result, a number of tweets have reposted the original tweets.

Applying the Shapiro tests and the Kolmogorov-Smirnov tests, none of the data was normally distributed (see Tables in Appendix C.2). The U-test suggested that the differences at the stock level was significant ($Z = 81.4765$, $p <$ 2.2e-16), but not at the ticker level ($Z = -1.5764$, $p = 0.1149$). Therefore, the non-stock-related tweets contained significantly more retweets than stock-related tweets. In addition, the differences among ticker-related tweets were significant as the Kruskal-Wallis rank sum test suggested ($\chi^2(2) = 21.7277$, $p =$ 1.914e-05). Furthermore, a post-hoc test consisting of a Kruskal-Wallis rank sum test – Mann-Whitney test with Bonferroni correction showed a significant difference between PST and NEU groups ($p =$ 5.678e-06, $r = 0.0865835$).

Therefore, retweet count has a significant effect at the stock level. This suggests that cashtag count might be useful in classification of the tweet types at the stock level. Additionally, there is a significant effect between the PST tweets and NEU tweets at the polarity level.

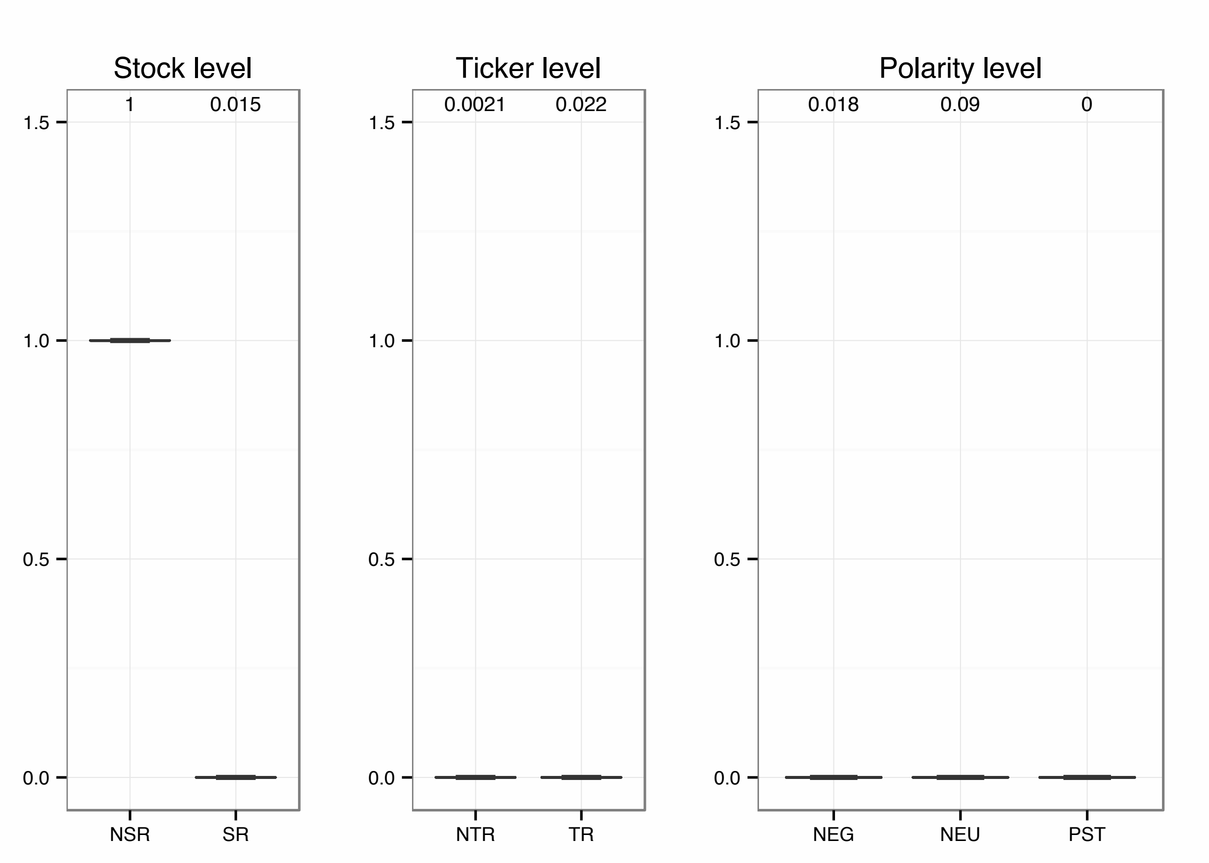


Figure 8.1.8 Boxplot of retweet counts in the annotated tweets

8.1.8 Discussion

To summarise, this section covered seven external linguistic features of the annotated tweets: word count including and excluding URLs, character count including and excluding URLs, retweet count, hashtag count and cashtag count. Among different sets of features, three groups of ticker-related tweets at the polarity level tend to be similar, while the NSR tweets differ from the stock-related tweets in terms of these seven aspects at the stock level, and the NTR tweets differ from ticker-related tweets in terms of most external linguistic features at the ticker level. Therefore, these external linguistic features will be used to design automatic classifiers, in order to identify the stock-related and ticker-related tweets (see Table 8.1.1).

Table 8.1.1 Summary of using the external features to identify the type of tweets

Features	**Stock level**	**Ticker level**	**Polarity level**
Word count including URLs	r=-0.119 ***	r=0.066 ***	N.S.
Character count including URLs	r=-0.184 ***	r=0.084 ***	N.S.
Word count excluding URLs	r=-0.003 ***	r=0.048 ***	PST & NEG r=0.049 *
Character count excluding URLs	r=-0.039 **	N.S.	N.S.
Hashtag count	N.S.	N.S.	N.S.
Cashtag count	r=-0.001 ***	r=0.598 ***	N.S.
Retweet count	r=0.994 ***	N.S.	PST & NEU r=0.087 *

* indicates that $p < 0.05$, ** indicates that $p < 0.01$, and *** indicates that $p < 0.001$

From the above analyses, NTR tweets have the longest average word count and character count (see Table 8.1.2), and they also have the most URL links (calculated as subtraction of the word count including URLs from word count excluding URLs). Hashtag count does not display much difference, but the cashtag count shows a distinct difference among different groups. In particular, the NSR group has the highest average cashtag count. This seems unexpected because it is usually believed that stock tweets would contain more cashtags than other categories of tweets. However, as introduced in Chapter 4, a number of cashtags were not used in accordance with their original purpose (to indicate a ticker), so this may explain why the NSR tweets contain more cashtags than other groups. Finally, retweet count also presents clear variance across different groups of tweets. The NSR group contains far more retweets than the other three groups, and the NTR group contains the least retweets. This shows that the retweeting mechanism is less often used in stock conversation on Twitter, particularly when the conversation does not focus on a specific ticker. Finally, at the polarity level, there is not much difference across three polarities. This suggests that different polarities of tweets tend to be similar in terms of external linguistic features.

Table 8.1.2 Average counts of the external linguistic features

Features	**NSR**	**Stock-related**	**NTR**	**Ticker-related**	**NEG**	**NEU**	**PST**
Word count including URLs	14.07	15.56	15.87	15.39	15.84	15.24	15.32
Word count excluding URLs	13.89	14.79	14.94	14.69	15.21	14.56	14.6
Character count including URLs	74.68	88.99	91.67	87.32	88.99	85.96	87.34
Character count excluding URLs	71.18	73.31	73.09	73.43	76.49	72.44	72.94
Hashtag count	0.63	0.64	0.64	0.64	0.6	0.7	0.63
Cashtag count	1.1	2.74	2.6	2.81	2.55	2.69	2.92
Retweet count	1	0.015	0.0021	0.022	0.018	0.09	0

8.2 Automatic Classification of the Ticker-related Tweets

The previous section showed that various external linguistic features differ across tweet types at different levels. The following sections further explore whether and how these differences might be used in the automatic classification of tweets at the stock level, ticker level and polarity level.

Four different machine learning methods introduced in Chapter 6 are used to classify the annotated GE tweets based on their external linguistic features. The training uses a 10-fold cross validation to access the effectiveness, and the test uses untrained GE tweets, and an additional test is conducted at the ticker level by using CAT (Caterpillar Inc.) tweets.

8.2.1 Data preparation for the classification

The following analyses also use the annotated GE tweets as the training set to compare four different machine learning methods as in Chapter 7 – the decision tree, random forests, Naïve Bayes, and support vector machine methods. For each of the tweets, the external linguistic features described in the previous sections were extracted and used to create a tweet-feature matrix - a schematic of the matrix is shown in Table 8.2.1. In this matrix, the rows represent individual tweets, and the columns represent the external linguistic features (e.g. word count, character count, cashtag count, hashtag count, link count, retweet count, character count without URLs, and word count without URLs). Each cell shows the value of the feature for a specific tweet. Hashtag count is excluded as it did not show any significant effect in the foregoing statistical analysis, but link count is included as it can change the length of a tweet.

Table 8.2.1 An extract of the external features of the annotated GE tweets

Type	Word	Character	Cashtag	Link	Retweet	Char. w/o URL	Word w/o URL
NSR	6	35	1	1	1	15	5
NTR	12	80	2	1	0	60	11
NTR	10	64	2	1	0	44	9
NTR	18	105	2	1	0	85	17
NTR	18	84	2	1	0	64	17
NTR	13	78	2	1	0	58	12

At the stock level, 1000 non-stock-related and 1000 stock-related tweets were randomly extracted, and used as the training set, and the other 533 non-stock-related and 4183 stock-related tweets were used as the test set. Similarly, at the ticker stage, 1000 non-ticker-related and ticker-related tweets were randomly extracted for use as the training set, and the other 884 non-ticker-related and 2299 ticker-related tweets were used as the test set. Finally, at the polarity stage, 500 negative, neutral and positive tweets were randomly extracted, and used as the training set, and the remaining 52 negative, 198 neutral and 1549 positive tweets were used as the test set.

Separate tweet-feature matrices were generated for the stock, ticker and polarity levels. The features (i.e. columns) included for each of these analyses varied according to the level. At the stock level, the classifiers used six features excluding hashtag count. At the ticker level, the classifiers used four external features, as the above statistical analysis recommended. At the polarity level, all features were used, even though only word count excluding URLs and retweet count were found to be indicative in the foregoing analysis. This was done because only using two distinct external features results in difficulties whilst training machine learning classifiers.

8.2.2 Stock level training

The first training process was conducted at the stock level. All four classifiers achieved similar results; each of them yielded a very high accuracy. Carrying out a 10-fold cross validation to the four supervised classifiers, the results of the best accuracy are presented in Table 8.2.2. These findings with the best accuracy present a high accuracy – almost a perfect classification. Why is the accuracy for the stock level so high? Probable clues can be found in the analysis of the external linguistic features in the previous section. As can be seen in Table 8.1.1, at the stock level, six of seven features show a significant difference between the stock-related and non-stock-related tweets. These features provide the automatic classifiers sufficient information for classifying tweets perfectly. Applying the trained classifiers to the test set, they yielded a similar result as presented in Table 8.2.2.

Table 8.2.2 The best results of four classifiers at the stock level

Method	**Best training accuracy**	**Test accuracy of GE tweets**	**Test accuracy of CAT tweets**
Decision tree	0.998	0.999	0.873
Random forests	1	0.999	0.873
Naïve Bayes	0.998	0.995	0.859
Support vector machine	1	0.999	0.873

8.2.3 Ticker-level training

The second phase of training was undertaken at the ticker level. Applying a 10-fold cross validation to the training of the ticker-level decision tree classifier, the results of best accuracy are presented in Table 8.2.3. At this level, the four classifiers exhibited considerably lower accuracy than at the stock level. The random forests classifier slightly outperformed the other three, while the Naive Bayes had the worst performance. This accuracy drop may be related to the smaller differences among these features, as discussed in Section 8.1.8. Compared to six significant differences among seven features at the stock level, only four significant differences were found at the ticker level as the foregoing statistical tests suggested. Applying the trained classifiers to the GE test set, slight drops of accuracy were observed, especially from the Naïve Bayes classifier.

Table 8.2.3 The best results of four classifiers at the ticker level

Method	**Best training accuracy**	**Test accuracy of GE tweets**	**Test accuracy of CAT tweets**
decision tree	0.650	0.601	0.770
random forests	0.664	0.642	0.752
Naïve Bayes	0.628	0.499	0.758
Support vector machine	0.642	0.579	0.774

Applying the trained classifiers to the CAT test set, overall, the decision tree and support vector machine classifiers outperformed the other three classifiers, and followed by the random forests and Naïve Bayes classifiers, each of which achieved a slightly lower accuracy. Compared with the training accuracy of the GE test set, for each method, the accuracy of the validation is much higher. This suggested that the differences of the external features in the CAT tweets are more pronounced than that observed in the GE ticker tweet set. Thus, this validation showed that the classifiers based on the features of the GE tweets work well on CAT tweets too. The results also showed that the decision tree, random forests and support vector machine methods outperformed the Naïve Bayes method.

8.2.4 Polarity-level training

Although the above statistical analysis showed that only two features have a significant effect at the polarity level (word count excluding URLs and retweet count), sentiment classification remains the ultimate goal of this research. Therefore, the last training phase attempts to tackle this problem, and is consequentially conducted at the polarity level (see table 8.2.4). All four methods performed poorly with an accuracy of around 0.4. Although the accuracy is low, it is better than would be expected (i.e. 0.33 according to three polarity categories). Among them, the random forests classifier achieved the best accuracy of 0.446. The other three had similarly lower accuracies. In addition, applying the trained classifier to the GE test set, only the random forests and support vector machine classifiers achieved a similar accuracy to that found during the training stage, but both the decision tree and Naïve Bayes classifiers failed to classify the polarity. This finding concured with the findings reported during the analysis of the external linguistic features, where the differences at the polarity level were minimum as shown in Section 8.1.8. These low accuracies suggest that the external linguistic features offer limited help in identifying the polarity of ticker-related tweets. Applying the trained classifiers to the CAT tweet set, they achieved a similarly poor performance as was found during the test stage.

Table 8.2.4 The best results of four classifiers at the polarity level

Method	**Best training accuracy**	**Test accuracy of GE tweets**	**Test accuracy of CAT tweets**
decision tree	0.367	0.115	0.234
random forests	0.446	0.466	0.352
Naïve Bayes	0.383	0.173	0.318
Support vector machine	0.415	0.427	0.299

8.3 Discussion

This chapter focused on two research questions, particularly QuestQuestion i

> Question 2. Are stock tweets a linguistically distinct type of tweet? And what specific linguistic features do they have?

> Question 3. If stock tweets have explicit linguistic features, is it possible to automatically identify the stock tweets based on their linguistic features?

As shown in Table 8.3.1, this chapter trained four classifiers at each of several levels based on external features. At the stock level, all these four classifiers could well identify the category of tweets. At the ticker level, of the four supervised learning classifiers, the random forests classifier achieved the best accuracy with an accuracy of 0.664. The other three classifiers achieved a slightly lower accuracy of around 0.6 at the ticker level.

Table 8.3.1 Statistics of the accuracy of each level of four classifiers

Method	**Stock level**	**Ticker level**	**Polarity level**
Decision tree	0.998	0.650	0.367
Random forests	1	0.664	0.446
Naïve Bayes	0.998	0.628	0.383
Support vector machine	1	0.642	0.415

These differences of classification accuracy at each level may relate to the number of features that differ significantly between sets, as tested in Section 8.1. At the stock level, six out of seven external linguistic features presented a significant effect of difference, while at the ticker level, only four out of seven features had a significant effect of difference. At these two levels, all features differed only in two categories, but at the polarity level, the statistical tests were used to evaluate variation across three categories. In other words, at the polarity level, there were 21 pair-to-pair differences, instead of 7 pair-to-pair differences at the prior two levels. However, as shown in Table 8.1.1, only two out of 21 pair-to-pair features presented a significant effect of difference. Considering the number of significant variations at different levels, the differences of performance of classifiers at different levels seems understandable.

8.4 Summary

This chapter reported on the external linguistic features of classified stock tweets, such as word count including or excluding URLs, character count with or without URLs, hashtag count, cashtag count and retweet counts. By applying different significance tests, these features showed to be useful in differentiation of non-stock-related tweets from stock-related tweets, and non-ticker-related tweets from ticker-related tweets.

These distinct external linguistic features were then used to train four supervised learning classifiers – decision tree, random forests, Naïve Bayes, and support vector machine – to classify the annotated GE tweets at different levels. Four classifiers achieved a high accuracy at the stock level, and a moderate accuracy at the ticker level, but none of them performed well at the polarity level.

Compared with the sentiment classification conducted in Chapter 7, the classification result at the polarity level in this chapter has been shown to perform poorly. This shows that using the external linguistics features solely to classify sentiment does not work as well as expected. Therefore, the following analyses combine the external linguistic features with the bag of words model to see if the classification accuracy can be improved.

Chapter 9 Sentiment Classification Based on the Hybrid Features

To answer the overarching question of this research – to establish whether the use of of linguistic features results in improved accuracy in a sample sentiment classification task – this chapter combines different features to test the performance of each classifier. The first section reviews the sentiment classifications achieved in previous chapters. Then, an initial classification task combines the bag-of-words model with part-of-speech features. The second task combines part-of-speech features with external linguistic features. The third task combines the bag-of-words model with external linguistic features. In general, each of these steps demonstrates a slight improvement over the baseline accuracy. Based on these classifications, the final classification combines all sets of features. As it will be shown, the best accuracy is achieved by using a random forest classifier based on the combination of all sets of features. The best accuracy achieved is 0.685.

9.1 Review of Sentiment Classifications in the Previous Chapters

The two foregoing analyses, reported respectively in Chapters 7 and 8, investigated the internal and external linguistic features of a corpus of the annotated stock tweets. Individual sentiment classifications based on these features achieved different results. The automatic classification based on the bag-of-words model in Chapter 7 produced a reasonably good result. All four classifiers achieved an accuracy of around 0.6. These results were therefore regarded as the baseline accuracy level for this study (see Table 9.1.1). In this baseline result, the support vector machine classifier had the best accuracy of 0.610, and the decision tree classifier had the poorest accuracy of 0.583. However, a classification based on part-of-speech features of tweets (reported in Chapter 7) and a classification based on the external linguistic features of tweets (reported in Chapter 8) did not perform as well as classification based on the bag-of-words model.

Table 9.1.1 Sentiment classification accuracy at the training stage based on different sets of features in previous chapters

Method	bag-of-words	Part-of-speech	External features
Decision tree	0.583	0.397	0.367
Random forests	0.592	0.468	0.446
Naïve Bayes	0.591	0.397	0.383
Support vector machine	0.610	0.425	0.415

These differences in classification performance may relate to the number of features exhibiting significant variation between categories of tweet, as reported in previous chapters. As shown in Table 9.1.2, at the top, the set of bag-of-words scores present a significant effect at the polarity level, and across each pair of data. At the middle, the set of part-of-speech features, although four out of five features have a significant effect of differences at the polarity level, only 7 out of 15 pairs of data present a significant effect of differences among the NEG, NEU and PST group. At the bottom, the set of external linguistic forms, only two out of seven features present a significant effect of differences at the polarity level, and only 2 out of 21 pairs of data present a significant effect of differences. Although the set of part-of-speech features present more pairs of differences than the set of external linguistics features, only *verbs* has a significant effect of differences across all three pairs of the NEG, NEU and PST data. This may explain why the set of part-of-speech features only has a slightly better result than the set of external linguistics features.

Therefore, to improve the sentiment classification accuracy, this chapter combines these three sets of features to explore if sentiment classifications based on the hybrid features can improve on the accuracy level achieved in the baseline result.

Table 9.1.2 Summary of significance tests of three sets of features

Features	**Polarity level**	**PST & NEG**	**PST & NEU**	**NEG & NEU**
Overall score	$\chi^2(2)$ = 1166.531 ***	*r* = .612 ***	*r* = .332 ***	*r* = .571 ***
Positive score	$\chi^2(2)$ = 701.391 ***	*r* = .483 ***	*r* = .279 ***	*r* = .302 ***
Negative score	$\chi^2(2)$ = 632.414 ***	*r* = .481 ***	*r* = .154 ***	*r* = .464 ***
Adjectives	$\chi^2(2)$ = 14.433 ***	*r* = 1.263e-05 ***	*r* = .0003 ***	N.S.
Common Nouns	$\chi^2(2)$ = 7.276 *	N.S.	*r* = .0003 *	N.S.
Prepositions	$\chi^2(2)$ = 24.700 ***	*r* = 2.473e-06 ***	*r* = 1.353e-06 ***	N.S.
Proper Nouns	N.S.	N.S.	N.S.	N.S.
Verbs	$\chi^2(2)$ = 25.549 ***	*r* = 2.287e06 ***	*r* = .0002 ***	*r* = 6.278e-08 ***
Word count including URLs	N.S.	N.S.	N.S.	N.S.
Character count including URLs	N.S.	N.S.	N.S.	N.S.
Word count excluding URLs	$\chi^2(2)$ = 6.5805 *	PST & NEG r=0.049 *	N.S.	N.S.
Character count excluding URLs	N.S.	N.S.	N.S.	N.S.
Hashtag count	N.S.	N.S.	N.S.	N.S.
Cashtag count	N.S.	N.S.	N.S.	N.S.
Retweet count	$\chi^2(2)$ = 21.7277 ***	PST & NEU r=0.087 *	N.S.	N.S.

* indicates that $p < 0.05$, ** indicates that $p < 0.01$, and *** indicates that $p < 0.001$

9.2 Classification Based on the Combined Internal Features

The first classification in this chapter focuses on the combination of the internal linguistic features – bag-of-words model is combined with part-of-speech features. Combining the most frequently observed part-of-speech features with the bag-of-words scores reported in Chapter 7, a matrix was generated, containing 8 columns and 1500 rows (see Table 9.2.1). This matrix was to be used for training the classifier. The first column indicates the tweet type, and the second to the fifth columns represent the part-of-speech features. The last three columns represent the scores of the bag-of-words model: the negative scores are for the negative keywords, the positive scores for the positive keywords, and the overall scores for the sum of the negative and positive scores. Each row represents the above features of an individual tweet, and each cell shows the occurrence or the score of a particular feature.

Table 9.2.1 Sample of the training model based on the combined internal features

Type	**Common noun**	**Adjective**	**Verb**	...	**Overall score**	**Positive score**	**Negative score**
PST	2	1	1	...	3	5	-2
PST	1	2	1	...	3	4	-1
PST	1	0	2	...	-7	1	-8

Applying the four supervised learning classifiers, Table 9.2.2 shows the classification results. The support vector machine classifiers performed best. The random forests classifier and the Naïve Bayes classifier had a 4% improvement over the baseline, while the decision tree classifier and the support vector machine had a 3% improvement. However, applying the trained classifiers to the GE test set, except the Naïve Bayes classifier, the other three classifiers had a slightly poorer performance. In addition, these four classifiers displayed a even poorer performance when being applied to the CAT test set.

Table 9.2.2 Classification results based on the combined internal features

Method	Best training accuracy	Test accuracy of GE tweets	Test accuracy of CAT tweets
Decision tree	0.614	0.542	0.352
Random forests	0.633	0.617	0.366
Naïve Bayes	0.632	0.671	0.389
Support vector machine	0.645	0.617	0.347

9.3 Classification Based on the POS and External Features

The second classification investigates whether combining part-of-speech features and external linguistic features can improve sentiment classification accuracy. Table 9.3.1 shows an extract of the matrix to be used for training the classifier: It has 13 columns and 1500 rows. Similarly, the first column represents the tweet type, the second to fifth columns represent the part-of-speech features, and the remaining columns represent the external linguistic features.

Table 9.3.1 Sample of the training model based on POS and external features

Type	Common noun	Adjective	Verb	...	Word	Character	Cashtag
PST	2	1	1	...	12	72	3
PST	1	2	1	...	11	67	3
PST	1	0	2	...	13	79	3

All classifiers displayed decreased accuracies, compared to the baseline accuracy (see Table 9.3.2). The Naïve Bayes classifier had the most obvious drop. This shows that neglecting the bag-of-words model, none of the classifiers performed well. However, applying the trained classifiers to the GE test set, all of them achieved a better accuracy than that in the training stage. Similarly, they performed poorly with the CAT test set, as did the classifiers based on the combined internal features above.

Table 9.3.2 Classification results based on POS and external features

Method	**Best training accuracy**	**Test accuracy of GE tweets**	**Test accuracy of CAT tweets**
Decision tree	0.394	0.560	0.347
Random forests	0.467	0.498	0.333
Naïve Bayes	0.389	0.389	0.349
Support vector machine	0.431	0.429	0.278

Compared with the sentiment classification results based individually on part-of-speech features or external linguistic features, the four classifiers displayed a slight improvement over the classifiers based solely on part-of-speech features, but all of them performed similarly to classifiers based on external linguistic features (see Table 9.3.3).

Table 9.3.3 Comparison of the classification results based only on part-of-speech or external features

Method	**POS features only**	**External features only**	**POS and External features**
Decision tree	0.377	0.397	0.394
Random forests	0.438	0.468	0.467
Naïve Bayes	0.383	0.397	0.389
Support vector machine	0.413	0.425	0.431

9.4 Classification Based on the Bag-of-words and External Features

The third classification tests the combination of the bag-of-words model and external linguistic features. Table 9.4.1 shows an extract of the matrix to be trained. This matrix has 11 columns.

The first column is the tweet type, the second to fourth columns represent are the scores of the bag-of-words model and the rest are the external features.

Table 9.4.1 Sample of the training model based on combination of the bag-of-words model and external linguistic features

Type	Overall score	Positive score	Negative score	...	Word	Character	Cashtag
PST	3	5	-2	...	12	72	3
PST	3	4	-1	...	11	67	3
PST	-7	1	-8	...	13	79	3

All classifiers improved in accuracy, with the random forests classifier and support vector machine classifier performing particularly strongly (see Table 9.4.2). Applying the trained classifiers to the GE test set, only the decision tree showed a clear drop in accuracy. However, with the CAT test set, none of the classifiers performed well.

Table 9.4.2 Classification results based on the combination of bag-of-words and external features

Method	Best training accuracy	Test accuracy of GE tweets	Test accuracy of CAT tweets
Decision tree	0.620	0.542	0.352
Random forests	0.641	0.647	0.383
Naïve Bayes	0.619	0.647	0.400
Support vector machine	0.651	0.637	0.392

9.5 Classification Based on the Combined Internal and External Features

The above three analyses showed that combining different sets of features, can result in improved sentiment classification accuracy. To maximise the accuracy, this final classification tests the combination of all four sets of features. The matrix being trained has 15 columns and 1500 rows. It contains three sets of features: the part-of-speech, external linguistic features and bag-of-words scores (see Table 9.5.1).

Table 9.5.1 Sample of the training model based on the combined internal and external features

Type	Noun	Adverb	...	Word	Character	...	Overall score	Positive score	Negative score
PST	2	1	...	12	72	...	3	5	-2
PST	1	2	...	11	67	...	3	4	-1
PST	1	0	...	13	79	...	-7	1	-8

All four classifiers trained in this manner showed an improved performance. The random forests topped the accuracy again; followed by the support vector machine classifier. Both decision tree and Naïve Bayes classifiers displayed similar performance. Compared with the baseline accuracy, the random forests classifier in this experiment displayed the most significant improvement among all experiments: 9.7%.

However, applying the trained classifiers to the GE test set resulted in decreased accuracy in all four cases. Even worse, the test on the CAT test set displayed poorer performance.

Table 9.5.2 Classification results based on combined internal and external features

Method	Best training accuracy	Test accuracy of GE tweets	Test accuracy of CAT tweets
Decision tree	0.611	0.545	0.352
Random forests	0.685	0.663	0.377
Naïve Bayes	0.616	0.404	0.404
Support vector machine	0.660	0.629	0.360

9.6 Discussion

It is clear from the above experiments that all four classifiers showed an accuracy increase when applied to the training set by combining different sets of features (see Table 9.6.1). The decision tree, random forests and support vector machine classifiers exhibited an obvious improvement against the baseline, and the Naïve Bayes classifier exhibited a slight improvement against the baseline. Comparing the best accuracy in four experiments in this chapter with the baseline of each classifier, the random forests classifier based on the all features had the best improvement, 9.7%, while the other three classifier displayed improvements ranging from 3.1% to 4.9%. Most importantly, the random forests classifier based on all features achieved the best accuracy of 0.685 among all classifiers in five experiments.

However, only random forests classifiers achieved the best accuracy when all sets of features are included, while the other three classifiers achieved improved accuracy based on the combination of the bag-of-words model and external linguistic features. This is rather unexpected as the previous experiments suggested that inclusion of more features with significant between-group variance would increase the accuracy of sentiment classification.

Table 9.6.1 The best accuracy (in bold) of each classifier based on different features

Method	Baseline	BoW + POS	POS + external	BoW + external	All features
Decision tree	0.583	0.614	0.394	**0.620**	0.611
Random forests	0.592	0.633	0.467	0.641	**0.685**
Naïve Bayes	0.591	0.632	0.389	**0.619**	0.616
Support vector	0.610	0.645	0.431	**0.651**	0.630

Finally, the support vector machine classifier displayed the best performance among the four classifiers in the three experiments – baseline, based on BoW and POS, and based on Bow and external features, while the random forests classifier had the best performance in the other two experiments – based on all features and POS and external features. The decision tree and Naïve Bayes classifiers have similarly lower accuracies. Regarding the overall classification accuracy, the support vector machine is the best one among the four, whilst Naïve Bayes is the worst.

9.7 Summary

This chapter focused on the overarching research question in this research: **Do linguistic analysis improve the sentiment identification accuracy of the stock tweet sentiment**. It combined three sets of features – bag-of-words scores, part-of-speech features and external linguistic features – in order to train four supervised learning classifiers using different combinations of features. All four classifiers presented improved accuracies when different sets of features were combined; in particular, a random forests classifier presented a 9.7% increase over the baseline accuracy. The decision tree, Naïve Bayes and support vector machine classifiers achieved their best accuracy when trained using a combination of bag-of-words scores and external linguistic features, while the random forests classifier achieved its best score using the combination of all three sets of features. In a nutshell, the analyses reported in this chapter show that combining different sets of features can improve the accuracy of sentiment classification on annotated stock tweets.

Chapter 10 Conclusion

The final chapter first reviews the research context, and then discusses the overarching question with the six specific research questions in detail, including methodologies, results and limitations. Finally, it provides an overview of possible implementations in the future.

10.1 Research Context

The one question began with the observation that current applications of sentiment analysis for stock prediction based on corpora of tweets do not exhibit or apply a comprehensive understanding of the linguistic features of tweets. Some previous studies in this specific area have been conducted in the computational linguistics field, with an emphasis on algorithms, while others have been carried out in economics, with a focus on the application of economics theories. Both fields have contributed to this interdisciplinary field, but neither has paid enough attention to the essential aspect of this field: the language that is used in tweets.

As a linguistics-focused project, this research posed an overarching question in Chapter 1, proposing that the observation of linguistic features is capable of improving the accuracy of the sentiment classification process in stock tweets. Then, to delve more deeply into this argument, six specific research questions were raised:

> Question 1. Does a clearer definition of stock tweets improve the quality of an analysis of such tweets?
>
> Question 2. Are stock tweets a linguistically distinct type of tweet? What specific linguistic features do they have?
>
> Question 3. If stock tweets have explicit linguistic features, is it possible to automatically identify the stock tweets based on their linguistic features?

Question 4. How can a robust sentiment word list for tweet sentiment classification be designed?

Question 5. Does a more precise definition of a positive, neutral, or negative stock tweet in accordance with market values help to improve the quality of stock tweet sentiment analysis?

Question 6. How can the neutral sentiment category of stock tweets be defined and processed?

These six specific research questions were answered in turn, providing supporting material that was then used to answer the overarching question of sentiment analysis in Chapter 9. To answer these questions, the research applied different analytical methods according to the domain of inquiry.

10.2 Discussion

This section discusses the six specific research questions first, and then reviews the overarching question: **Do linguistic analysis improve the sentiment identification accuracy of stock tweet sentiment analysis?** The section presents and discusses the results of these questions with their methodologies, and compares them with previous work.

10.2.1 Linguistic features of stock tweets

In chapter 7 and 8, this research divided the linguistic features of stock tweets into two categories: the internal and external features. Features relating to the meaning of tweets were regarded as the internal features, while other low-level linguistics features, such as word count, were viewed as the external features.

Chapter 7 focused on the internal features, using a frequency list to investigate the most and least frequent words in the annotated tweets, and then applied stemming and keyness analysis to identify the keywords in different polarities. Finally, it investigated the part-of-speech features by frequency list and different significance tests in statistical analysis. Chapter 8 focused on the external features, and used different statistical methods to ascertain whether the identified external linguistic features are sufficient to differentiate stock tweets from noise, and whether their use can improve identification of tweet sentiment polarity.

In terms of the internal features, evaluation of frequently-appearing unigrams did not show a great deal of explicit differences across the three polarities. Each polarity contained a number of ticker names and stock-related words, and presented a short-term discussion about the market, including some time-related words. In other words, all polarities tended to exhibit similar characteristics. On the other hand, unigrams with low frequency of appearance did demonstrate differences; however, a large percent of them only occurred once in the corpus, so they were not sufficiently stable referents by which to identify the polarity of tweets. An analysis of part-of-speech features showed that four out of five most frequent part-of-speech categories significantly differ between stock tweets and non-stock tweets, including *adjectives*, *common nouns*, *proper nouns*, and *verbs*.

Analysis of the external features – word count including/excluding URLs, character count including/excluding URLs, hashtag count, cashtag count and retweet count – within annotated stock tweets demonstrated that several distinct features were of interest. First, stock-related tweets contained more characters than non-stock-related tweets, so stock-related tweets were longer on average than other tweets in the dataset. Second, stock-related tweets contained more hashtags and cashtags, but fewer retweets. Most importantly, statistical significance tests suggested that six of these differences are vary significant at the stock level, four at the ticker level, and only two at the polarity level.

In short, stock tweets are linguistically distinct to other tweets: Stock tweets share a number of similarities in the internal features among different polarities, while the external features differentiate them from other tweets from a number of aspects.

Section 3.4 summarised four main linguistic aspects of tweets which have been explored in previous studies, namely vocabulary features, conversation features, hashtag and cashtag features, and topic features. But only a few of these previous studies focused on stock tweets in particular, and none of them has provided a thorough analysis of the linguistic features of stock tweets. In this sense, this research provided the first comprehensive linguistic analysis of stock tweets, which would contribute to the future analysis in this specific area.

10.2.2 Definition of stock tweets

To provide an in-depth definition of stock tweets, this research used manual annotation (Chapter 4), time series analysis (Chapter 5), corpus analysis (Chapter 7 and 8) and significance analysis (Chapter 7 and 8) to investigate the tweet data.

Although the data collection in Chapter 4 adopted a narrow definition of stock tweets to collect tweets from Twitter's Search API, the results of the data collection ('crawling') process still contained a large proportion of irrelevant data. Based on manual annotation, the definition of stock tweets became much clearer: stock tweets are tweets discussing stock market related topics, and containing at least one cashtag, which refers to the stock market.

The time series analysis in Chapter 5 first showed that stock-related and ticker-related tweets have much stronger correlations with the corresponding stock prices than non-stock-related and non-ticker-related tweets. The analysis of internal linguistic features in Chapter 7 and external linguistic features in Chapter 8 demonstrated that, linguistically, stock tweets differ from other tweets as summarised in Section 10.2.1.

As a consequence of the temporal analysis in Chapter 5 and the analysis of internal and external linguistic features in Chapter 7 and 8, this thesis offers a clear and comprehensive definition for stock tweets: **the stock tweets are tweets discussing stock market related topics, and containing at least one cashtag, which refers to the stock market. They are longer than general tweets, and have a closer temporal relationship with the corresponding stock price**. Based on these, it is reasonable to summarise that stock tweets are temporally and linguistically different from general tweets, so they can be considered as a specific type of tweet.

As discussed in Section 3.3, there are three approaches to collect data for stock prediction based on tweet data: among others, these include collection of general tweets and collection of ticker-specific tweets. The second approach is the only option that attempts to filter for stock tweets; however, even in this category, previous studies did not provide any detailed definition of stock tweets. Some of them took the view that tweets containing ticker names of S&P 500 companies with a dollar sign in front are de facto stock tweets (Brown, 2012; Y. Mao et al., 2012; Oliveira et al., 2013; Smailović et al., 2012; Sprenger & Welpe, 2010), whilst some also considered tweets containing company names, company leader names, or even informal company names as stock tweets (Chakoumakos et al., 2011; Ruiz et al., 2012).

The main consequence of such loose definitions is that they may include a great deal of noise in the collected data. This thesis represents the first research to provide a comprehensive definition of stock tweets, which would be useful for later automatic identification of stock tweets and their sentiments.

10.2.3 Identification of stock tweets

As shown in the manual annotation stage described in Chapter 4, nearly half of annotated tweets are not ticker-related tweets, so including them to conduct a sentiment classification runs the risk of inaccurate evaluation. Thus, based on the analysis of external linguistic features, automatic classifiers were trained by using four machine learning methods in order to identify stock tweets, enabling unrelated tweets to be filtered from the dataset.

As shown in Table 10.2.1, at the stock level, all four classifiers performed perfectly, in other words, they could identify stock-related tweets well. At the ticker level, all of them performed moderately well. The difference in performance between these two stages may relate to the prior significance tests. Six external linguistic features had a significant effect of difference at the stock level, but only four at the ticker level.

Table 10.2.1 Comparison of the best training accuracy at the stock level and the ticker level

Method	**Stock level**	**Ticker level**
Decision tree	0.998	0.650
Random forests	1	0.664
Naïve Bayes	0.998	0.628
Support vector machine	1	0.642

Automatically identifying stock tweets is critical to the later sentiment classification as it introduces the possibility of an effective capability for filtering noise from the sheer quantities of tweet data. However, only few of the previous studies have identified this problem; moreover, none of them has done this extensively. To the author's knowledge, the sole previous discussions of this approach can be found in Nann et al. (2013) and Bar-Haim et al. (2011)'s studies, both of which used simple word lists to clean up their data. Therefore, this study is the first to establish a simple and effective solution to automatically identify stock tweets from raw data.

10.2.4 Definition of sentiment polarities in stock tweets

As discussed in Chapter 3, there was a disagreement between tweet sentiment and the market trend, which concurs with Sprenger & Welpe (2010)'s argument. Moreover, there are two main classification strategies in sentiment analysis, either polarity or fine-grained, but the more suitable choice of approach for stock tweet classification has not been discussed extensively before. Manually annotating collected tweets in Chapter 4 provided a closer demonstration of this disagreement.

As the market includes 'buying', 'selling' and 'holding', this relationship was retained for manual annotation, in order to identify correspondence between tweet sentiment and these actions. Therefore, the annotation work described in Chapter 4 sought to classify stock tweets into three polarities: negative, neutral and positive. Specifically, the three polarities of stock tweets can be summarised as:

Negative tweets

1. Tweets express negative thoughts about the ticker.
2. Tweets express the idea of 'selling the ticker'.
3. Tweets discuss about the negative news of the ticker.
4. Tweets express disappointments or other negative emotions about events of relevance of the ticker.

Neutral tweets

1. Tweets ask a question, though they probably have a positive or negative tendency.

2. Tweets express the idea of 'holding' the ticker.
3. Tweets report the ticker news, but contain no explicit sentiment.
4. Tweets advertise the ticker, focus on the ticker exclusively, but do not contain any obvious sentiments.
5. Tweets contain conflicting opinions in an single tweet.

Positive tweets

1. Tweets repeat positive news related to the ticker, for example, new investments, new product launches, or a new leader's good performance.
2. Tweets report positive news of the ticker's movements.
3. Tweets mention an investor's interest in or confidence about the ticker.
4. Tweets express a positive sentiment explicitly, even when the market displays opposite motion.

Previous studies have very limited discussions of how to define sentiments in stock tweets. It appears that the only discussion can be found in Sprenger & Welpe (2010) and Sprenger et al. (2013)'s work (the first paper is the working paper of the second one). Their discussions mainly focused on the relationship between market and tweet sentiment, but have not presented a detailed definition of each sentiment category. This research provided a series of detailed definitions, which can be useful for further automatic sentiment classification, especially based on supervised learning methods.

10.2.5 Processing of neutral tweets

As shown in the data annotation phase described in Chapter 4, the analysis used a hierarchy model to filter out noise from conventionally neutral tweets, and classified them as non-stock-related tweets and non-ticker-related tweets.

Specifically, these two types of noise meet the following standards:

Non-stock-related tweets

1. Tweets borrow the cashtag, but do not discuss any relevant topic related to the concept of the crawling keyword.
2. Tweets use the cashtag according to its original purpose, but the main topic of these tweets is not about the stock market, the ticker company, or any related concepts.
3. Tweets contain no text, other than a single cashtag.

Non-ticker-related tweets

1. Tweets advertise stock-related products, such as investment reports.
2. Tweets discuss a similar topic, but the foci are not the keyword.
3. Tweets do not contain a precisely identifiable cashtag.
4. Tweets contain the cashtag, but are automatically generated by a website.
5. Retweets of the previously listed four types of non-ticker-related tweets.

Furthermore, based on this hierarchy model, the research then applied time series analysis to compare the temporal correlation between this model and the conventional model in Chapter 5. The results showed that the hierarchy model can improve the positive correlation strength by 1%-4%.

10.2.6 Generating a localised word list for sentiment classification

Based on detailed analyses of the internal linguistic features of tweets, the research then used keyness rank of stemmed unigrams to generate word lists for positive and negative polarities respectively. This is a new approach, because all previous research used pre-designed sentiment word lists, for example, a tweet sentiment word list by Davies and Ghahramani (2011). However, with a relatively small unnormalised tweet dataset, it was found that these pre-designed word lists did not perform well.

This localised word list was less statistically biased than previous word lists and provided a better coverage than previous word lists did. With such a localised word list, Chapter 7 achieved a relatively good result in subsequent sentiment classification. In addition, it was easy to apply. Therefore, an approach which focuses on generation of a localised sentiment word list can provide benefits at the analysis stage, especially for sentiment classification on small datasets or applying semi-supervised machine learning methods to train automatic classifiers.

10.2.7 Sentiment classification of stock tweets

In Chapter 7, an analysis stage was described that used word lists based on keyness analysis to train the bag-of-words model, in order to provide a baseline accuracy against which to compare subsequent sentiment classification methods. The baseline accuracy was 0.583 by decision tree classifier, 0.592 by random forests classifier, 0.591 by Naïve Bayes classifier, and 0.610 by support vector machine classifier.

The research then selected different sets of features individually to classify the sentiment in stock tweets, including the part-of-speech features and external features, but it was found that neither of these methods outperformed the baseline approach described in Chapter 7. Nevertheless, in Chapter 9, the research combined these features with the bag-of-words model, and it was found that this hybrid approach provided a better result than the baseline (see Table 10.2.2). The random forests classifier provided the best performance among the four classifiers, with an accuracy of 0.685 based on the combination of all sets of features. However, the other three achieved improved accuracy based on the combination of bag-of-words model and external features: the decision tree classifier with its best accuracy of 0.620, the Naïve Bayes classifier with its best accuracy of 0.619, and a support vector machine classifier with its best accuracy of 0.651.

Table 10.2.2 The accuracy comparison at the training stage

Method	**Best accuracy in the training**
Decision tree	0.620
Random forests	0.685
Naïve Bayes	0.619
Support vector machine	0.651

In addition, comparing the best accuracies of the nominated classifiers with the baseline, it was found that the random forests classifier had the most obvious improvement of 0.094, whilst the other three showed slightly less improvement: the support vector machine classifier with an improvement of 0.041, the decision tree classifier with an improvement of 0.037, and the Naïve Bayes classifier with an improvement of 0.028.

The varying performance of each classifier based on different features may link to the number of significant effects of difference in the prior statistical analysis. Although it is interesting to see why these classifiers have different improvements based on different combinations of features, it is an algorithmic rather than a linguistic question. Therefore, it is beyond the scope of this research, although future study may focus on this question.

As listed in Table 10.2.3, only three out of ten stock prediction projects based on tweet data reported the sentiment classification accuracy achieved. Sprenger & Welpe (2010)'s classification based on the bag-of-words model achieved an accuracy of 0.643, while Oh & Sheng (2011)'s classification based on the combination of the bag-of-words model and a lexical scorer achieved an accuracy of 0.663. Yet Smailović et al. (2012) claimed that their sentiment classification achieved an accuracy of 0.810, which was a high result. Their classification was based on a general dataset collected by Go et al. (2009), which was completely different from their own stock tweet dataset. Therefore, it suggested that this result is utterly misleading and may not be fairly comparable with sentiment classification on more recently collected tweet datasets. Thus, compared with Sprenger & Welpe (2010) and Oh & Sheng (2011)'s results, this research outperformed both of them.

Table 10.2.3 Statistics of sentiment classification accuracy in previous studies (same as Table 3.6.1)

Project	**Best accuracy**	**Project**	**Best accuracy**
Sprenger & Welpe (2010)	0.643	Oh & Sheng (2011)	0.663
W. B. Zhang & Skiena (2010)	NA	Xue Zhang et al. (2011)	NA
Xue Zhang et al. (2010)	NA	Smailović et al. (2012)	NA
Bollen et al. (2011)	NA	Oliveira et al. (2013)	NA
Mao et al. (2011)	NA	Nann et al. (2013)	NA

In short, the sentiment classification in this research not only outperformed the baseline, but also beat previous research. In other words, combining linguistic features with the bag-of-words model **can** improve the sentiment classification accuracy. This is a particularly relevant finding to future practical applications of sentiment classification on stock tweets.

10.3 Conclusion

As shown in the previous chapters, the main contribution of this research includes five aspects, not only to sentiment analysis, but also to stock prediction, and as well as to the analysis of social media data.

It provided an in-depth linguistic analysis of stock tweets. The in-depth analysis covered the internal and external linguistic features of stock tweets, including various length counts, the most and least frequent unigrams and part of speech features. When these features were incorporated into the sentiment analysis procedure, it showed an improvement in sentiment classification accuracy. Such linguistic analyses can also be applied to other sets of textual data as preliminary analyses, providing an insight into the data. For example, a prediction of political elections based on social media data can first conduct an analysis on the linguistic features of such type of data, and based on the result of this linguistic analysis result, the prediction can thus categorize different groups of voters, in order to predict their voting intentions and behaviours.

It gave a clearer and more comprehensive definition of stock tweets. As summarised above, the stock tweets are tweets discussing stock market related topics, and containing at least one cashtag, which refers to the stock market. They are longer than general tweets, and have a closer temporal relationship with the corresponding stock price. Such a definition not only contributes to a specified data collection approach of stock tweets, but also improves the detection of fraudulent tweets with a focus of stock market. The future analyses focusing on stock tweets can apply this definition, and thus simplify their work. Additionally, other analyses of social network data could borrow the idea of defining their data and then applying further investigations, which may improve the relevance of the data if a classification based on the definition is applied accordingly.

It presented a simple but effective way to automatically identify stock tweets, based on the in-depth analysis of the external linguistic features of stock tweets. This procedure can significantly reduce the computing cost of the later sentiment analysis as a large proportion of irrelevant tweets can be excluded. In addition, researchers working on the identification of other types of tweets may benefit from the example provided by this approach. For example, commercial social network analysis has attracted more interests than before, and identifying marketing tweets of a specific company has become much important in this area. If this kind of analysis apply such a tweet recognition process, the potential of achieving more accurate results could be extended, because this strategy can significantly reduce the redundancy and noise in the data.

It proposed a simple but effective solution enabling the generation of a localised sentiment keyword list. Although the bag-of-words model has been one of the most popular approaches in current sentiment analysis, most of the applications used a pre-designed word list that may not fit the specific data well. Designing a localised word list for the further analysis can improve the accuracy of sentiment classification, and this approach can be applied to other similar classifications based on the occurrence of the words.

It demonstrated a significant improvement of sentiment classification accuracy within stock tweet datasets. As demonstrated in the previous chapters, the sentiment analysis based on the stemmed stock tweets by adding different linguistic features to the bag-of-words model has improved its accuracy. Such an approach has a wide application, so it can be applied to not only the stock tweets, but also other domains of data. For example, an analysis of product reviews or restaurant reviews can apply this strategy, as the data to be analysed are similarly concise and full of spelling variants.

Based on the various contributions reviewed above, this thesis provides a clear direction for future sentiment analysis, adding that the linguistic features of the textual data are the essence of such analyses, therefore, analysing the linguistic features of such textual datasets should receive more attention.

In a broader sense, this thesis also sets an example for future analysis of social network data by showing that analysing the textual data is a must: to ignore the linguistic features of textual data would be to miss much of the value of working with social network data.

Furthermore, it provides a blueprint for bridging linguistics and natural language processing, and shows that an interdisciplinary approach to analysis can yield better results than would be achieved by either discipline working alone. Therefore, there should be more collaborations between these two areas, because such interdisciplinary studies do require a great deal of high-level knowledge from both areas, and working alone indeed takes much more time and effort to solve problems in these overlapping areas.

10.4 Limitation and Future Work

This study was not without limitations. Given the limited time and resources, this research focused on a fairly small scale dataset. In the future, it is worth scaling up this research in the following ways.

The research only analysed GE (General Electronics) tweets, and used a small proportion of CAT (Caterpillar, Inc.) tweets to validate the result. However, data collection obtained a much larger dataset, which can be used to evaluate the research result.

Tweet data were achieved from the Search API, and this only provided a limited set of data. It is worth trying to analyse real-time data from the Streaming API.

One question that remained unanswered in this research is why the four machine learning methods perform very differently based on different features. Although it is an algorithmic problem rather than a linguistic problem, it is worth further exploration in the future.

This research used the bag-of-words model to identify only sentiment in tweets, but it may also be possible to identify non-stock-related or non-ticker-related tweets using this model. Future studies can implement this.

The research did not conduct a time series analysis based on the automatic sentiment classification results. It is worth carrying out one to see whether it is possible to reach Bollen et al. (2011)'s high correlation.

Appendix A Scripts

Appendix A.1 Scripts for Data Collection

Appendix A.1.1 Scripts for crawling ticker tweets

```
#!/usr/bin/env bash


DIR=tweets/T`date "+%d-%m-%y-%H:%M"`

mkdir -p $DIR

wget -i tweet/ticker_address.txt  -np -r -N -l1 -P $DIR
```

Appendix A.1.2 Scripts for crawling ticker tweets

```
#!/usr/bin/env bash


DIR=tweets/M`date "+%d-%m-%y-%H:%M"`

mkdir -p $DIR

wget -i tweet/media_address.txt -np -r -N -l1 -P $DIR
```

Appendix A.1.3 Scripts for crawling retweets

```
#!/usr/bin/env bash


DIR=tweets/R`date "+%d-%m-%y-%H:%M"`

mkdir -p $DIR

wget -i tweet/retweet_address.txt  -np -r -N -l1 -P $DIR
```

Appendix A.2 Scripts for Data Analysis

All the scripts for data analysis can be found at: `https://github.com/illy/Codes_for_dissertation`

Appendix B Address Book for Data Collection

Appendix B.1 Address Book for Ticker Tweets

http://search.twitter.com/search.json?q=%24INTC&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%24MMM&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%24AA&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%24AXP&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%24SBC&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%24NB&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%24BA&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%24CAT&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%24CHV&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%24CSCO&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%24KO&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%24DD&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%24XON&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%24GE&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%24HWP&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%24HD&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%24IBM&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%24JNJ&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%24CHL&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%24KFT&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%24MCD&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%24MRK&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%24MSFT&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%24PFE&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%24PG&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%24SPC&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%24UTX&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%24BEL&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%24WMT&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%24DIS&result_type=recent&lang=en&rpp=100&

Appendix B.2 Address Book for Media Tweets

http://search.twitter.com/search.json?q=%40msnbc_breaking&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40Breakingnews&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40cnnbrk&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40BBCBreaking&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40wsjbreakingnewsg&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40SkyNewsBreak&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40BreakingNewz&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40CNBCbrk&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40USABreakingNews&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40ABSCBNBreaking&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40news_of_the_day&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40breakingstorm&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40aubrk&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40Breaking_News&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40Reuters&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40thomsonreuters&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40RLStream&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40ReutersAgency&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40ReutersMarkets&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40reutersuk&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40EUReuters&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40ReutersTech&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40ReutersUS&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40RtrsHedgeFunds&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40Reuters_TopNews&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40reuters_co_uk&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40BloombergNews&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40BloombergJapan&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40BloombergGov&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40BloombergLaw&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40BloombergLP&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40BloombergMrkts&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40BloombergMuse&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40BloombergNEF&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40BloombergNews&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40BloombergNow&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40BloombergRadio&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40BloombergTech&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40BloombergTV&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40BloombergView&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40BloombergWest&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40BSURVEILLANCE&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40BW&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40OnTheEconomy&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40BloombergWest&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40thelexcolumn&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40ftasia&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40ftchina&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40connectedbiz&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40ftenergy&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40ftfinancenews&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40ftcomment&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40ftcommodities&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40ftcompanies&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40ftmoney&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40ftmedianews&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40fttechnews&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40ftuknews&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40ftuseconomy&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40ftconferences&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40FTWeekendMag&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40financialtimes&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40ftfirehose&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40WSJ&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40WSJMarkets&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40wsjusnews&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40WSJNY&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40WSJHeard&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40WSJDeals&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40WSJWorldMarkets&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40WSJTech&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40WSJopinion&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40WSJRealEstate&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40stocktwits&result_type=recent&lang=en&rpp=100&

http://search.twitter.com/search.json?q=%40msnbc_breaking&result_type=recent&lang=en&rpp=100&

Appendix B.3 Address Book for Reweets

http://search.twitter.com/search.json?q=RT%40msnbc_breaking&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40Breakingnews&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40cnnbrk&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40BBCBreaking&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40wsjbreakingnewsg&result_type=recent&lang=en&rpp=50

http://search.twitter.com/search.json?q=RT%40SkyNewsBreak&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40BreakingNewz&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40CNBCbrk&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40USABreakingNews&result_type=recent&lang=en&rpp=50

http://search.twitter.com/search.json?q=RT%40ABSCBNBreaking&result_type=recent&lang=en&rpp=50

http://search.twitter.com/search.json?q=RT%40news_of_the_day&result_type=recent&lang=en&rpp=50

http://search.twitter.com/search.json?q=RT%40breakingstorm&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40aubrk&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40Breaking_News&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40Reuters&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40thomsonreuters&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40RLStream&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40ReutersAgency&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40ReutersMarkets&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40reutersuk&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40EUReuters&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40ReutersTech&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40ReutersUS&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40RtrsHedgeFunds&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40Reuters_TopNews&result_type=recent&lang=en&rpp=50

http://search.twitter.com/search.json?q=RT%40reuters_co_uk&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40BloombergNews&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40BloombergJapan&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40BloombergGov&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40BloombergLaw&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40BloombergLP&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40BloombergMrkts&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40BloombergMuse&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40BloombergNEF&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40BloombergNews&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40BloombergNow&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40BloombergRadio&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40BloombergTech&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40BloombergTV&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40BloombergView&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40BloombergWest&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40BSURVEILLANCE&result_type=recent&lang=en&rpp=50

http://search.twitter.com/search.json?q=RT%40BW&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40OnTheEconomy&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40BloombergWest&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40thelexcolumn&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40ftasia&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40ftchina&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40connectedbiz&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40ftenergy&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40ftfinancenews&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40ftcomment&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40ftcommodities&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40ftcompanies&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40ftmoney&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40ftmedianews&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40fttechnews&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40ftuknews&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40ftuseconomy&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40ftconferences&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40FTWeekendMag&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40financialtimes&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40ftfirehose&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40WSJ&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40WSJMarkets&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40wsjusnews&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40WSJNY&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40WSJHeard&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40WSJDeals&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40WSJWorldMarkets&result_type=recent&lang=en&rpp=50

http://search.twitter.com/search.json?q=RT%40WSJTech&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40WSJopinion&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40WSJRealEstate&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40stocktwits&result_type=recent&lang=en&rpp=50&

http://search.twitter.com/search.json?q=RT%40msnbc_breaking&result_type=recent&lang=en&rpp=50&

Appendix C Normality test results

Appendix C.1 Normality Test Results for Chapter 7

Appendix C.1.1 Analysis of bag-of-words scores

Table C.1.1 The Shapiro test of the overall BoW scores in NEG, NEU, and PST tweets

Type	W value	p-value
NEG	0.9788	3.49e-07
NEU	0.9849	1.229e-06
PST	0.9907	3.131e-10

Table C.1.2 The Shapiro test of the positive BoW scores in NEG, NEU, and PST tweets

Type	W value	p-value
NEG	0.9127	2.2e-16
NEU	0.9506	1.402e-14
PST	0.9803	3.189e-16

Table C.1.3 The Shapiro test of the negative BoW scores in NEG, NEU, and PST tweets

Type	W value	p-value
NEG	0.9649	3.063e-10
NEU	0.9147	2.2e-16
PST	0.8877	2.2e-16

Appendix C.1.2 Analysis of POS tags

Table C.1.4 The Shapiro test of adjectives in NEG, NEU, and PST tweets

Type	W value	p-value
NEG	0.8019	<2.2e-16
NEU	0.8066	<2.2e-16
PST	0.8313	<2.2e-16

Table C.1.5 The Shapiro test of common nouns in NEG, NEU, and PST tweets

Type	W value	p-value
NEG	0.9509	<1.337e-16
NEU	0.949	<7.583e-16
PST	0.9417	<2.2e-16

Table C.1.6 The Shapiro test of prepositions in NEG, NEU, and PST tweets

Type	W value	p-value
NEG	0.9178	<2.2e-16
NEU	0.9005	<2.2e-16
PST	0.8901	<2.2e-16

Table C.1.7 The Shapiro test of proper nouns in NEG, NEU, and PST tweets

Type	W value	p-value
NEG	0.8423	<2.2e-16
NEU	0.8327	<2.2e-16

Type	W value	p-value
PST	0.8644	<2.2e-16

Table C.1.8 The Shapiro test of verbs in NEG, NEU, and PST tweets

Type	W value	p-value
NEG	0.8877	<2.2e-16
NEU	0.858	<2.2e-16
PST	0.8638	<2.2e-16

Appendix C.2 Normality Test Results for Chapter 8

Table C.2.1 The Shapiro test of word count including URLs in NSR, NTR, NEG, NEU, and PST tweets

Type	W value	p-value
NSR	0.9692	2.2e-16
NTR	0.9854	2.902e-12
NEG	0.989	0.0003372
NEU	0.9911	0.0001535
PST	0.9886	3.337e-12

Table C.2.2 The Kolmogorov-Smirnov test of word count including URLs in ticker-related and stock-related tweets

Type	W value	p-value
Ticker-related	0.9987	< 2.2e-16
Stock-related	0.999	< 2.2e-16

Table C.2.3 The Shapiro test of character count including URLs in NTR, NSR, NEG, NEU, and PST tweets including URLs

Type	W value	p-value
NTR	0.9441	< 2.2e-16
NSR	0.9676	< 2.2e-16
NEG	0.9623	8.61e-11
NEU	0.9674	5.305e-12
PST	0.9694	< 2.2e-16

Table C.2.4 The Kolmogorov-Smirnov test of character count including URLs in ticker-related and stock-related tweets

Type	W value	p-value
Ticker-related	1	< 2.2e-16
Stock-related	1	< 2.2e-16

Table C.2.5 The Shapiro test of word count excluding URLs in NSR, NTR, NEG, NEU and PST tweets

Type	W value	p-value
NSR	0.973	< 2.2e-16
NTR	0.9865	2.487e-12
NEG	0.9827	3.727e-06
NEU	0.9861	3.217e-06
PST	0.9759	< 2.2e-16

Table C.2.6 The Kolmogorov-Smirnov test of word count excluding URLs in ticker-related and stock-related tweets

Type	W value	p-value
Ticker-related	0.9987	< 2.2e-16
Stock-related	0.9987	< 2.2e-16

Table C.2.7 The Shapiro test of character count excluding URLs in NSR, NTR, NEG, NEU and PST tweets

Type	W value	p-value
NTR	0.967	< 2.2e-16
NSR	0.9844	1.782e-13
NEG	0.983	4.598e-06
NEU	0.9879	1.511e-05
PST	0.992	3.702e-09

Table C.2.8 The Kolmogorov-Smirnov test of character count excluding URLs in ticker-related and stock-related tweets

Type	W value	p-value
Ticker-related	1	< 2.2e-16
Stock-related	1	< 2.2e-16

Table C.2.9 The Shapiro test of hashtag count in NSR, NTR, NEG, NEU and PST tweets

Type	W value	p-value
NTR	0.4075	< 2.2e-16
NSR	0.7421	< 2.2e-16

Type	W value	p-value
NEG	0.6346	< 2.2e-16
NEU	0.6569	< 2.2e-16
PST	0.672	< 2.2e-16

Table C.2.10 The Kolmogorov-Smirnov test of hashtag count in ticker-related and stock-related tweets

Type	W value	p-value
Ticker-related	0.5	< 2.2e-16
Stock-related	0.5	< 2.2e-16

Table C.2.11 The Shapiro test of cashtag count in NSR, NTR, NEG, NEU and PST tweets

Type	W value	p-value
NSR	0.2469	< 2.2e-16
NTR	0.8365	< 2.2e-16
NEG	0.6391	< 2.2e-16
NEU	0.6582	< 2.2e-16
PST	0.6604	< 2.2e-16

Table C.2.12 The Kolmogorov-Smirnov test of cashtag count in ticker-related and stock-related tweets

Type	W value	p-value
Ticker-related	0.9772	< 2.2e-16
Stock-related	0.9772	< 2.2e-16

Table C.2.13 The Shapiro test of word count including URLs in NSR, NTR, NEG, NEU and PST tweets

Type	W value	p-value
NTR	0.6572	< 2.2e-16
NSR	0.4304	< 2.2e-16
NEG	0.4054	< 2.2e-16
NEU	0.3796	< 2.2e-16
PST	0.3724	< 2.2e-16

Table C.2.14 The Kolmogorov-Smirnov test of retweet count in ticker-related and stock-related tweets

Type	W value	p-value
Ticker-related	0.5	< 2.2e-16
Stock-related	0.5	< 2.2e-16

Reference

Ahmad, K., Cheng, D., & Almas, Y. (2006). Multi-lingual sentiment analysis of financial news streams. In *Grid technology for financial modeling and simulation*. Palermo.

Ammirati, S. (2007, September). Twitter's Open Platform Advantage. Retrieved from `http://readwrite.com/2007/09/05/twitter_open_platform_advantage`

André, P., Bernstein, M., & Luther, K. (2012). Who gives a tweet?: evaluating microblog content value. In *Proceedings of the aCM 2012 conference on computer supported cooperative work* (pp. 471–474).

Anthony, L. (2004). AntConc: A learner and classroom friendly, multi-platform corpus analysis toolkit. In *Proceedings of an interactive workshop on language e-learning* (pp. 7–13).

Antweiler, W., & Frank, M. (2004). Is all that talk just noise? The information content of internet stock message boards. *The Journal of Finance*, *59*(3), 1259–1294.

Argan, M., Sevil, G., & Yalama, A. (2011). Word-of-Mouth Communication Effect in the Holdings and Trades of Stocks: Empirical Evidence from Emerging Market. In *Academic and business research institute international conference, international conference*. Las Vegas.

Asur, S., & Huberman, B. (2010). Predicting the future with social media. In *WI-iAT '10 proceedings of the 2010 iEEE international conference on web intelligence and intelligent agent technology*.

Aw, A., Zhang, M., Xiao, J., & Su, J. (2006). A phrase-based statistical model for SMS text normalization. In *Proceedings of the cOLING/ACL 2006 main conference poster sessions* (pp. 33–40). Association for Computational Linguistics.

Awan, A. (2010). *Sentiment Analysis Over Newswire Final Report*.

Baayen, R. (2008). *Analyzing linguistic data: A practical introduction to statistics using R* (1st ed.). Cambridge: Cambridge University Press.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Seventh conference on international language resources and evaluation, malta.* (p. 2010).

Bar-Haim, R., Dinur, E., Feldman, R., Fresko, M., & Goldstein, G. (2011). Identifying and following expert investors in stock microblogs. In *EMNLP '11 proceedings of the conference on empirical methods in natural language processing*. Edinburgh.

Barber, B. M., & Odean, T. (2007). All That Glitters: The Effect of Attention and News on the Buying Behavior of Individual and Institutional Investors. *Review of Financial Studies*, *21*(2), 785–818.

Barbosa, G., Silva, I., Zaki, M., Meira, W., Prates, R., & Veloso, A. (2012). Characterizing the effectiveness of twitter hashtags to detect and track online population sentiment. In *CHI '12 extended abstracts on human factors in computing systems.*

Baron, A. (2014). **VARD 2**. *Proceedings of the Postgraduate Conference in Corpus Linguistics.*

Benevenuto, F., Magno, G., Rodrigues, T., & Almeida, V. (2010). Detecting spammers on twitter. In *Proceedings of the 7th annual collaboration, electronic messaging, anti-abuse and spam conference (cEAS).*

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. New York: Springer.

Bloom, K. (2011). *Sentiment Analysis Based on Appraisal Theory and Functional Local Grammars* (PhD thesis). lingcog.iit.edu.

Bollen, J., Mao, H., & Zeng, X. J. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1–8.

Bollen, J., Pepe, A., & Mao, H. (2009). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *Fifth international aAAI conference on weblogs and social media*.

Bontcheva, K., Dercrynski, L., Funk, A., Greenwood, M., Maynard, D., & Aswani, N. (2013). TwitIE: An Open-Source Information Extraction Pipeline for Microblog Text. In *Proceedings of the international conference on recent advances in natural language processing (rANLP 2013)*.

Bordino, I., Battiston, S., Caldarelli, G., Cristelli, M., Ukkonen, A., & Weber, I. (2012). Web search queries can predict stock market volumes. *PLOS ONE*, *7*(7).

boyd, danah michele. (2011). *Taken Out of Context*. Proquest, Umi Dissertation Publishing.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Brown, E. (2012). Will twitter make you a better investor? a look at sentiment, user reputation and their effect on the stock market. In *Proceedings of the southern association for information systems conference, 2012*. Atlanta.

Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New Avenues in Opinion Mining and Sentiment Analysis. *Intelligent Systems*, *28*(2), 15–21.

Carey, B. (2012). Using the Twitter REST API, 1–6.

Cavicchi, T. J. (2000). *Digital Signal Processing*. Hoboken: Wiley.

Chakoumakos, R., Trusheim, S., & Yendluri, V. (2011). Automated Market Sentiment Analysis of Twitter for Options Trading.

Chalmers, D., Fleming, S., Wakeman, I., & Watson, D. (2011). Rhythms in Twitter. In *1st international workshop on social object networks, 3rd iEEE conference on social computing*. Boston.

Chan, W. S. (2003). Stock price reaction to news and no-news: drift and reversal after headlines. *Journal of Financial Economics*, *70*(2), 223–260.

Chatfield, C. (1980). *The analysis of time series: an introduction*. London: Chapman; Hall.

Chen, R., & Lazer, M. (2011). Sentiment Analysis of Twitter Feeds for the Prediction of Stock Market Movement.

Chyan, A., Lengerich, C., & Hsieh, T. (2011). A stock-purchasing agent from sentiment analysis of twitter.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Mahwah: Lawrence Erlbaum Associates.

Conover, W. J. (1999). *Practical nonparametric statistics* (3rd ed.). Hoboken: Wiley.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*, 273–297.

Costolo, D. (2010, December). Meaningful Growth. Retrieved from `http://blog.twitter.com/2010/12/stocking-stuffer.html`

Cunningham, H., Tablan, V., Roberts, A., & Bontcheva, K. (2013). Getting more out of biomedical documents with GATE's full lifecycle open source text analytics. *PLoS Comput Biol*, *9*(2), e1002854–e1002854.

Das, S. R., & Chen, M. (2007). Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management Science*, *53*(9), 1375–1388.

Datasift. (2012, December). Data Sources - DataSift. Retrieved from `http://datasift.com/source/6/twitter`

Davidov, D., Tsur, O., & Rappoport, A. (2010). Enhanced sentiment learning using Twitter hashtags and smileys. In *Proceedings of the 23rd international conference on computational linguistics: Posters* (pp. 241–249).

Davies, A., & Ghahramani, Z. (2011). Language-independent Bayesian sentiment mining of Twitter. In *Fifth international workshop on social network mining and analysis (sNAKDD 2011).*

Davies, M. (2008, October). *The Corpus of Contemporary American English: 450 million words, 1990-present.*

Debbini, D., Estin, P., & Goutagny, M. (2011). Modeling the Stock Market Using Twitter Sentiment Analysis.

Derczynski, L., Ritter, A., Clark, S., & Bontcheva, K. (2013). Twitter Part-of-Speech Tagging for All: Overcoming Sparse and Noisy Data. In *Proceedings of recent ...*.

Dewally, M. (2003). Internet investment advice: Investing with a rock of salt. *Financial Analysts Journal*, *59*(4), 65–77.

Dichter, E. (1966). How word-of-mouth advertising works. *Harvard Business Review*, *44*(11-12), 147–166.

Dictionary, O. E. (n.d.). "ecosystem, n.". Retrieved from `http://www.oed.com/view/Entry/59402?redirectedFrom=ecosystem`

Doan, S., Ohno-Machado, L., & Collier, N. (2012). Enhancing Twitter Data Analysis with Simple Semantic Filtering: Example in Tracking Influenza-Like Illnesses. In *2012 iEEE second international conference on healthcare informatics, imaging and systems biology*. La Jolla.

Doan, S., Vo, B., & Collier, N. (2012). An analysis of Twitter messages in the 2011 Tohoku Earthquake. *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, *91*, 58–66.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, *19*(1), 61–74.

Earle, P. S., Bowden, D. C., & Guy, M. (2012). Twitter earthquake detection: earthquake monitoring in a social world. *Annals of Geophysics*, *54*(6).

Earle, P., Guy, M., Buckmaster, R., Ostrum, C., Horvath, S., & Vaughan, A. (2010). OMG Earthquake! Can Twitter improve earthquake response?, *81*(2), 246.

Enders, C. K. (2010). *Applied Missing Data Analysis*. New York: Guilford Press.

Escobar, M. (2011). How to select a social monitoring tool. Factors to consider, 1–3.

Fama, E. (1965a). Random walks in stock market prices. *Financial Analysts Journal*, *21*(5), 55–59.

Fama, E. (1965b). The Behavior of Stock-Market Prices. *Journal of Business*, *38*(1), 34–105.

Fama, E. (1970). Efficient capital markets: A review of theory and empirical work. *Journal of Finance*, *25*(2), 383–417.

Fama, E. (1996). Discounting under uncertainty. *Journal of Business*, *69*(4), 415–428.

Fielding, R. (2000). *Architectural styles and the design of network-based software architectures* (PhD thesis).

Finger, L., & Dutta, S. (2013). Who is FAKE? In *O'Reilly's strata conference, 2013* (pp. 1–43). Santa Clara.

Garrett, S. (2010, June). Big Goals, Big Game, Big Records. Retrieved from `https://blog.twitter.com/2010/big-goals-big-game-big-records`

Gilbert, E., & Karahalios, K. (2009). Widespread worry and the stock market.

Gimpel, K., Schneider, N., OConnor, B., Das, D., Mills, D., Eisenstein, J., ... Smith, N. A. (2011). Part-of-speech tagging for twitter: Annotation, features, and experiments. In *The 49th annual meeting of the association for computational linguistics: Human language technologies*.

Gnip. (2012, December). Historical Twitter Data - Gnip. Retrieved from `http://gnip.com/twitter_history/`

Gnip. (2013). Twitter Data - Gnip. Retrieved from `http://gnip.com/twitter/power-track/`

Go, A., & Bhayani, R. (2010). *Exploiting the Unique Characteristics of Tweets for Sentiment Analysis.*

Go, A., Bhayani, R., & Huang, L. (2009). *Twitter sentiment classification using distant supervision.*

Gobry, P.-E. (2010, December). BUBBLE WATCH: New Hedge Fund Uses Twitter To Pick Stocks - Business Insider. Retrieved from `http://www.businessinsider.com/new-hedge-fund-uses-twitter-to-pick-stocks-2010-12`

Goonatilake, R., & Hearth, S. (2007). The Volatility of the Stock Market and News. *International Research Journal of Finance and Economics*, (11), 53–64.

Gries, S. T. (2009). *Statistics for linguistics with R: a practical introduction*. Berlin: Walter de Gruyter.

Grimes, S. (2013). Text Analytics in 2013. In *Text analytics in 2013* (pp. 1–5). San Francisco.

Gross, M. (1982). *Constructing lexicon-grammars.*

Han, B., & Baldwin, T. (2011). Lexical Normalisation of Short Text Messages: Makn Sens a# twitter. In *Proceedings of the 49th annual meeting of the association for computational linguistics* (pp. 368–378). Portland.

Harris, Z. S. (1954). Distributional Structure. *Word*, *10*(2/3), 146–162.

Hassan, M., & Nath, B. (2005). Stock market forecasting using hidden Markov model: a new approach. In *The 5th international conference on intelligent systems design and applications (iSDA05).*

Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York: Springer.

Hatzivassiloglou, V., & McKeown, K. (1997). Predicting the semantic orientation of adjectives. In

The 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics (aCL '98) (pp. 174–181). Madrid, Spain.

Hepworth, I. (2012, October). The human face of big data | Twitter Blogs. Retrieved from http://blog.twitter.com/2012/10/the-human-face-of-big-data.html

Hogenboom, A., Bal, D., Frasincar, F., Bal, M., Jong, F., & Kaymak, U. (2013). Exploiting emoticons in sentiment analysis. In *28th annual aCM symposium on applied computing, sAC 2013*. Coimbra.

Honeycutt, C., & Herring, S. C. (2009). Beyond Microblogging: Conversation and Collaboration via Twitter. In *42nd hawaii international conference on system sciences*. Los Alamitos, CA.

Hsu, E., Shiu, S., & Torczynski, D. (2011). Predicting dow jones movement with twitter.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the aCM sIGKDD international conference on knowledge ; discovery and data mining (kDD-2004)* (pp. 168–177). Seattle: ACM.

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Indvik, L. (2013, April). Tweets Coming to Bloomberg Terminals. Retrieved from http://mashable.com/2013/04/04/bloomberg-terminals-twitter/

Ingram, M. (2011, April). Can Twitter Help You Predict the Stock Markets Tech News and Analysis. Retrieved from http://gigaom.com/2011/04/06/can-twitter-help-you-predict-the-stock-market/

Jones, D. (2011, August). Dow Jones Averages. Retrieved from http://www.djaverages.com/?go=industrial-overview

Jordan, J. (2010, December). Hedge Fund Will Track Twitter to Predict Stock Moves -

Bloomberg. Retrieved from `http://www.bloomberg.com/news/2010-12-22/hedge-fund-will-track-twitter-to-predict-stockmarket-movements.html`

Kaufmann, M. (2010). Syntactic Normalization of Twitter Messages.

Kevin, H.-Y. L., Yang, C., & Hsin-Hsi, C. (2008). Emotion classification of online news articles from the reader's perspective. In *Proceedings of the 2008 iEEE/WIC/ACM international conference on web intelligence and intelligent agent technology* (pp. 220–226). Sydney.

Kiciman, E. (2010). Language Differences and Metadata Features on Twitter. In *Web n-gram workshop at aCM sIGIR special interest group on information retrieval 2010*. Geneva.

Kilgarriff, A. (2005). Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory*, *1*(2), 263–276.

Kononenko, I., & Kukar, M. (2007). *Machine Learning and Data Mining: Introduction to Principles and Algorithms*. Cambridge: Horwood Publishing.

Koppel, M., & Shtrimberg, I. (2006). Good news or bad news? let the market decide. In *Proceedings of the aAAI spring symposium on exploring attitude and affect in text: Theories and applications* (pp. 297–301). Stanford.

Kuleshov, V. (2011). Can Twitter predict the stock market?

Kumar, A., & Sebastian, T. (2012). Sentiment Analysis on Twitter. *International Journal of Computer Science Issues*, *9*(4), 372–378.

Kumar, A., & Sebastian, T. M. (2012). Sentiment Analysis: A Perspective on its Past, Present and Future. *Intelligent Systems and Applications*, *4*(10), 1–14.

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media? *WWW 2010*.

Lake, T. (2010). Status report: Twitter nlp. *Western Michigan University, Kalamazoo*.

Lampos, V., & Cristianini, N. (2010). Tracking the flu pandemic by monitoring the Social Web. In *2nd international workshop on cognitive information*.

Leeuwen, L. van. (2011, July). *Monitoring the chatter in social media: The Amsterdam Exchange Index* (PhD thesis).

Lehrer, A. (1974). *Semantic Fields and Lexical Structure*. North-Holland.

Leinweber, D., & Sisk, J. (2010). Relating News Analytics to Stock Returns. In *5th annual cARISMA conference* (pp. 1–30). London: CARISMA, London.

Li, F. (2006). Do stock market investors understand the risk sentiment of corporate annual reports.

Litosseliti, L. (2010). *Research Methods in Linguistics*. Continuum.

Liu, K., Li, W., & Guo, M. (2012). Emoticon Smoothed Language Models for Twitter Sentiment Analysis. In *Twenty-sixth aAAI conference on artificial intelligence*. Toronto.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10Ks. *The Journal of Finance*, *66*(1), 35–65.

Malkiel, B. (2003). The efficient market hypothesis and its critics. *Journal of Economic Perspectives*, (1), 59–82.

Mao, H., Counts, S., & Bollen, J. (2011). Predicting Financial Markets: Comparing Survey, News, Twitter and Search Engine Data. *Arxiv Preprint ArXiv:1112.1051*.

Mao, Y., Wang, B., Wei, W., & Liu, B. (2012). Correlating S&P 500 Stocks with Twitter Data. *Nlab.engr.uconn.edu*.

Mason, O. (2004). Automatic processing of local grammar patterns. In *Proceedings of the 7th annual colloquium for the uK special interest group for computational linguistics, university of birmingham* (pp. 166–171).

McCrank, J., & Gaffen, D. (2013, January). Hoax tweets send Audience shares atwitter. Retrieved from http://www.reuters.com/article/2013/01/29/us-audience-shares-idUSBRE90S11T20130129

McEnery, T., & Hardie, A. (2011). *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

McNair, D. M., Lorr, M., & Droppleman, L. F. (2003). Profile of Mood States (POMS).

Mei, Q., Ling, X., Wondra, M., Su, H., & Zhai, C. (2007). Topic sentiment mixture: modeling facets and opinions in weblogs. In *International world wide web conference com- mittee (iW3C2)* (pp. 171–180). Banff.

Melanson, M. (2011, February). Twitter Kills the API Whitelist: What it Means for Developers & Innovation. Retrieved from http://readwrite.com/2011/02/11/twitter_kills_the_api_whitelist_what_it_means_for

Mitchell, T. M. (1997). *Machine Learning*. New York: Springer.

Mitixa, R. P., & Rana, A. (2013). A Survey on Opinion and Sentiment Analysis With Applications and Issues. In *International journal of computational linguistics and natural language processing* (pp. 237–240).

Mittal, A., & Arpit, G. (2011). Stock Prediction Using Twitter Sentiment Analysis. *Final Report of Machine Learning Course, 2009*.

Mizrach, B., & Weerts, S. (2009). Experts online: An analysis of trading activity in a public Internet chat room. *Journal of Economic Behavior & Organization*, *70*, 266–281.

Mizumoto, K., Yanagimoto, H., & Yoshioka, M. (2012). Sentiment Analysis of Stock Market News with Semi-supervised Learning. In *2012IEEE/ACIS 11th international conference on computer and information science*. Shanghai.

Mohammad, S., Dunne, C., & Dorr, B. (2009). Generating High-Coverage Semantic Orientation Lexicons From Overtly Marked Words and a Thesaurus. In *Proceedings* (pp. 599–608). Singapore.

Morgan, S. (2010, March). Can Twitter Predict the Market. Retrieved from `http://www.smartmoney.com/invest/stocks/can-twitter-predict-the-market/`

Mukherjee, S., Bhattacharyya, P., & Balamurali, A. (2012). Sentiment Analysis in Twitter with Lightweight Discourse Analysis. In *The 24th international conference on computational linguistics (cOLING 2012)*. Bumbai.

Naaman, M., Boase, J., & Lai, C. (2010). Is it really about me?: message content in social awareness streams. In *2010 aCM conference on computer supported cooperative work*. Savannah.

Nann, S., Krauss, J., & Schoder, D. (2013). Predictive Analytics On Public Data-The Case Of Stock Markets. In *The 21st european conference on information systems*.

Naveed, N., Gottron, T., Kunegis, J., & Alhadi, A. C. (2011). Bad News Travel Fast: A Content-based Analysis of Interestingness on Twitter. *WebSci '11*.

Nigam, K., & Hurst, M. (2004). Towards a robust metric of opinion. *AAAI Spring Symposium on Exploring Attitude and Affect in Text*, 598603.

Ogilvie, D. M., Stone, P. J., & Kelly, E. F. (1982). Computer-aided content analysis. In R. B. Smith & P. K. Manning (Eds.), *Handbook of social science methods: Qualitative methods*. Pensacola, FL: Ballinger.

Oh, C., & Sheng, O. (2011). Investigating Predictive Power of Stock Micro Blog Sentiment in Forecasting Future Stock Price Directional Movement. In *Thirty second international conference on information systems*. Shanghai.

Oliveira, N., Cortez, P., & Areal, N. (2013). Some experiments on modeling stock market behavior using investor sentiment analysis and posting volume from Twitter. In *WIMS13*.

Otto, M. (2012, August). Bootstrap 2.1 and counting. Retrieved from http://blog.twitter.com/2012/08/bootstrap-21-and-counting.html

Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Shneider, N., & Smith, N. (2013). Improved part-of-speech tagging for online conversational text with word clusters. In *The conference of the north american chapter of the association for computational linguistics: Human language technologies, 2013*.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From Tweets to polls: Linking text sentiment to public opinion time series. In *The fourth international aAAI conference on weblogs and social media* (pp. 122–129).

Page, R. (2012). *Stories and Social Media: Identities and Interaction*. London: Routledge.

Paice, C. D. (1990). Another stemmer. In *ACM sIGIR special interest group on information retrieval 1990* (pp. 56–61). ACM.

Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *The seventh international conference on language resources and evaluation (lREC)*.

Pang, B., & Lee, L. (2008). *Opinion Mining and Sentiment Analysis* (Vol. 2). Hanover: Now Publishers Inc.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the conference on empirical methods in natural language processing (eMNLP), philadelphia, july 2002* (pp. 79–86). Philadelphia, Pennsylvania.

Pass, G. (2009, January). Building on Open Source. Retrieved from https://blog.twitter.com/2009/building-open-source

Patell, J., & Wolfson, M. (1984). The intraday speed of adjustment of stock prices to earnings and dividend announcements. *Journal of Financial Economics*.

Payne, A. (2009, January). Putting a ceiling on requests from users and IPs on the whitelist - Google Groups. Retrieved from https://groups.google.com/forum/?fromgroups=#!topic/twitter-development-talk/v2WnFgqKRMk

Perreault, M., & Ruths, D. (2011). The Effect of Mobile Platforms on Twitter Content Generation. In *Fifth international conference on weblogs and social media*. Barcelona.

Petrovic, S., Osborne, M., & Lavrenko, V. (2010). The Edinburgh Twitter Corpus. In *Proceedings of the nAACL hLT 2010 workshop on computational linguistics in a world of social media* (p. 25).

Pettit, A. (2011a, October). Step 1: Collect social media data that mentions cookies. Does the kind of cookie matter? #eso3d. Retrieved from https://twitter.com/lovestats/status/129638092322779136

Pettit, A. (2011b, October). Step 2: Clean the social media data. Do we really want viral games in our dataset? It depends on the research objective. #MRX #eso3d. Retrieved from https://twitter.com/lovestats/status/129639133726515200

Pettit, A. (2011c, October). Step 3: Did you forget to code slang? Emoticons? Does it matter? YES! #eso3d #mrx. Retrieved from https://twitter.com/lovestats/status/129639914718511104

Pettit, A. (2011d, October). Step 4: Sample precisely. Or wherever you're allowed. Or wherever you remember to look. #MRX #eso3d. Retrieved from https://twitter.com/lovestats/status/12964062725525

Porta, J., & Sancho, J. L. (2013). Word normalization in Twitter using finite-state transducers. In *Tweet normalization workshop at annual conference of the spanish society for natural language processing 2013*.

Porter, M. (1980). An algorithm for suffix stripping. *Program: Electronic Library and Informa-*

tion Systems, *14*(3), 130–137.

Porter, M. (2006, October). The Porter Stemming Algorithm. Retrieved from `http://tartarus.org/martin/PorterStemmer/`

Potts, C. (2011, November). Sentiment Symposium Tutorial: Language and cognition. Retrieved from `http://sentiment.christopherpotts.net/lingcog.html`

Programs, T. A. R., & Inc. (1981). Measuring the Grapevine–Consumer Response and Word-of-Mouth, 1–25.

Rambocas, M., & Gama, J. (2013). Marketing Research: The Role Of Sentiment Analysis. *Ideas.repec.org*.

Rao, S., & Hong, J. (2012). *Analysis of hidden markov models and support vector machines in financial applications* (No. UCB/EECS-2010-63).

Rao, T., & Srivastava, S. (2012). Modeling Movements in Oil, Gold, Forex and Market Indices using Search Volume Index and Twitter Sentiments. *ArXiv*. Retrieved from `http://arxiv.org/abs/1212.1037v1`

Rao, T., & Srivastava, S. (2012). Twitter Sentiment Analysis: How To Hedge Your Bets In The Stock Markets. *ArXiv Preprint ArXiv:1212.1107*.

Rayson, P. (2003). Matrix: A statistical method and software tool for linguistic analysis through corpus comparison. *Unpublished Doctoral Thesis, Lancaster University, Lancaster*.

Razali, N., & Wah, Y. (2011). Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of Statistical Modeling and Analytics*, *2*(1), 21–33.

Read, J. (2005). Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the aCL student research workshop (aCLstudent '05)*.

Richins, M. (1984). Word of mouth communication as negative information. *Advances in Consumer Research*, *11*(1), 697–702.

Roche, E. (1999). Finite state transducers: parsing free and frozen sentences. In *Proceeding of the eCAI 96 workshop extended finite state models of language* (pp. 108–120).

Royston, P. (1995). Remark AS R94: A remark on algorithm AS 181: The W-test for normality. *Journal of the Royal Statistical Society*, *44*(4), 547–551.

Ruiz, E., Hristidis, V., Castillo, C., Gionis, A., & Jaimes, A. (2012). Correlating financial time series with micro-blogging activity. In *The fifth aCM international conference on web search and data mining* (pp. 513–522).

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on world wide web* (pp. 851–860).

Sarver, R. (2010, March). Enabling A Rush of Innovation | Twitter Blogs. Retrieved from `http://blog.twitter.com/2010/03/enabling-rush-of-innovation.html`

Schumaker, R., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The AZFin text system. In *ACM transactions on information* (pp. 1–19). New York.

Scott, M., & Tribble, C. (2006). *Textual Patterns: Key Words and Corpus Analysis in Language Education*. Amsterdam: J. Benjamins.

Segaran, T. (2008). *Programming Collective Intelligence: Building Smart Web 2.0 Applications*. Sebastopol: O'Reilly Media.

Shasha, D. (2000). Time Series in Finance: the array database approach. Retrieved from `http://cs.nyu.edu/shasha/papers/jagtalk.html`

Shumway, R. H., & Stoffer, D. S. (2006). *Time series analysis and its applications: with R examples; 2nd ed.* New York: Springer.

Siegel, S., & Castellan, N. J. (1988). Nonparametric Statistics for Behavioural Science. New York: McGraw-Hill Book Company.

Singletary, T. (2013, February). Planning for API v1s Retirement. Retrieved from `https://dev.twitter.com/blog/planning-for-api-v1-retirement`

Sippey, M. (2012, August). Changes coming in Version 1.1 of the Twitter API | Twitter Developers. Retrieved from `https://dev.twitter.com/blog/changes-coming-to-twitter-api`

Smailović, J., Grčar, M., & Znidarsic, M. (2012). Sentiment analysis on tweets in a financial domain. *Ipssc.mps.si*.

Sprenger, T., & Welpe, I. (2010). Tweets and Trades-The Information Content of Stock Microblogs. *Papers.ssrn.com*.

Sprenger, T., Tumasjan, A., Sandner, P. G., & Welpe, I. M. (2013). Tweets and trades: The information content of stock microblogs. *European Financial Management*.

Stone, B. (2006, September). Introducing the Twitter API. Retrieved from `http://blog.twitter.com/2006/09/introducing-twitter-api.html`

Sukumar, N. (2013). DCM Capital Puts Itself Up for Sale in Online Auction - Bloomberg. *Bloomberg.com*.

Täckström, O., & McDonald, R. (2011). Discovering fine-grained sentiment with latent variable structured prediction models. *Advances in Information Retrieval*, 368–374.

Team, R. (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria.

Thau, K. (2010, November). Twitter + Ping = Discovering More Music. Retrieved from `https://blog.twitter.com/2010/twitter-ping-discovering-more-music`

Timmermann, A., & Granger, C. (2004). Efficient market hypothesis and forecasting. *International Journal of Forecasting*, (20), 15–27.

Turney, P. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 417–424).

Twitter. (2011a, April). API Terms of Service: April 13, 2011. Retrieved from https://dev.twitter.com/terms/api-terms/2011-04-13

Twitter. (2011b, July). One Million Registered Twitter Apps. Retrieved from https://blog.twitter.com/2011/one-million-registered-twitter-apps

Twitter. (2011c, October). Tweets per second. Retrieved from http://yearinreview.twitter.com/en/tps.html

Twitter. (2011d, September). One hundred million voices. Retrieved from http://blog.twitter.com/2011/09/one-hundred-million-voices.html

Twitter. (2012a, August). Rate Limiting. Retrieved from https://dev.twitter.com/docs/rate-limiting

Twitter. (2012b, August). REST API v1.1 Resources. Retrieved from https://dev.twitter.com/docs/api/1.1

Twitter. (2012c, December). GET search. Retrieved from https://dev.twitter.com/docs/api/1/get/search

Twitter. (2012d, December). There are now more than 200M monthly active @twitter users. You are the pulse of the planet. We're grateful for your ongoing support! Retrieved from https://twitter.com/twitter/status/281051652235087872

Twitter. (2012e, March). Twitter turns six. Retrieved from https://blog.twitter.com/

2012/twitter-turns-six

Twitter. (2012f, September). The Streaming APIs. Retrieved from https://dev.twitter.com/docs/streaming-apis

Twitter. (2013a, April). Frequently Asked Questions. Retrieved from https://dev.twitter.com/docs/faq#6861

Twitter. (2013b, March). GET search/tweets. Retrieved from https://dev.twitter.com/docs/api/1.1/get/search/tweets

Twitter. (2013c, March). REST API Rate Limiting in v1.1. Retrieved from https://dev.twitter.com/docs/rate-limiting/1.1

TwitterEng. (2011, June). 200 million Tweets per day. Retrieved from http://blog.twitter.com/2011/06/200-million-tweets-per-day.html

Vandor, M. (2012, December). Your Twitter archive. Retrieved from https://blog.twitter.com/2012/your-twitter-archive

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S*. New York: Springer.

Vincent, A., & Armstrong, M. (2010). Predicting break-points in trading strategies with Twitter. *SSRN ELibrary*.

Vlastelica, R., Bases, D., & Flitter, E. (2013, January). Second Twitter hoax in two days smacks another stock. Retrieved from http://www.reuters.com/article/2013/01/30/us-sarepta-idUSBRE90T1CF20130130

Vohra, M., & Teraiya, J. (2013). Applications and Challenges for Sentiment Analysis: A Survey. *International Journal of Engineering*, *2*(2).

Watters, A. (2011, March). How Recent Changes to Twitter's Terms of Service Might Hurt Academic Research. Retrieved from http://readwrite.com/2011/03/03/how_recent _changes_to_twitters_terms_of_service_mi

Weil, K. (2010, February). Measuring Tweets. Retrieved from http://blog.twitter.com/2010/02/measuring-tweets.html

Whitelaw, C., Garg, N., & Argamon, S. (2005a). Using appraisal groups for sentiment analysis. In *Proceedings of the 14th aCM international conference on information and knowledge management* (pp. 625–631).

Whitelaw, C., Garg, N., & Argamon, S. (2005b). Using appraisal taxonomies for sentiment analysis. In *ACM international conference on information and knowledge management* (p. 625).

Wickham, H. (2009). *ggplot2 : elegant graphics for data analysis.* New York : Springer.

Williams, D. (n.d.). Twitter Certified Products: Tools for Businesses. Retrieved from https://blog.twitter.com/2012/twitter-certified-products-tools-for-businesses

Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., ... Patwardhan, S. (2005a). OpinionFinder: A system for subjectivity analysis. In *Proceedings of hLT/EMNLP on interactive demonstrations* (pp. 34–35).

Wilson, T., Wiebe, J., & Hoffmann, P. (2005b). Recognizing contextual polarity: An Exploration of Features for Phrase-Level Sentiment Analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing* (pp. 347–354).

Xue, Z., Yin, D., Davison, B. D., & Davison, B. D. (2011). Normalizing Microtext. In *Proceedings of the association for the advancement of artificial intelligence, 2011, workshop on analyzing microtext* (pp. 74–79). Las Vegas.

Yang, J., & Leskovec, J. (2011). Patterns of temporal variation in online media. In *The fourth aCM international conference on web search and data mining* (pp. 177–186). Hong Kong: ACM.

Yarowsky, D. (1993). One sense per collocation. In *Workshop on human language technology.*

Yi, A. (2009). Stock Market Prediction Based on Public Attentions: a Social Web Mining Approach.

Zappavigna, M. (2011). Ambient affiliation: A linguistic perspective on Twitter. *New Media & Society*, *1*(19).

Zappavigna, M. (2012). *Discourse of Twitter and Social Media: How We Use Language to Create Affiliation on the Web*. London: Bloomsbury Publishing.

Zhang, W. B., & Skiena, S. (2010). Trading Strategies To Exploit Blog and News Sentiment. In *Fourth international aAAI conference on weblogs and social media*.

Zhang, X., & Zhu, X. (2007). A New Type of FeatureLoose N-Gram Feature in Text Categorization. In *The 3rd iberian conference on pattern recognition and image analysis, part i* (pp. 378–385). Springer.

Zhang, X., Fuehres, H., & Gloor, P. (2010). Predicting stock market indicators through TwitterI hope it is not as bad as I fear. *Procedia - Social and Behavioral Sciences*, *26*, 55–62.

Zhang, X., Fuehres, H., & Gloor, P. (2011). Predicting Asset Value through Twitter Buzz. *Advances in Collective Intelligence 2011*, (113), 23–34.

Zweig, J. (2011, August). Making Sense of Market Forecasts. Retrieved from `http://online.wsj.com/article/SB10001424052748703675904576064320900295678.html`