Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

School of Computer Science and Statistics

Investigating the Effect of Sentiment in High-Frequency Financial Markets

Darragh Mc Kay
13318094

May 4, 2018

Supervisor: Professor Khurshid Ahmad

A Dissertation submitted in partial fulfilment
of the requirements for the degree of
MAI (Computer Engineering)

Declaration

I hereby declare that the following dissertation, except where otherwise stated, is entirely my own work; that it has not previously been submitted as an exercise for a degree, either in Trinity College Dublin, or in any other University; and that the library may lend or copy it or any part thereof on request.

I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at http://www.tcd.ie/calendar.

I have also completed the Online Tutorial on avoiding plagiarism 'Ready Steady Write', located at http://tcd-ie.libguides.com/plagiarism/ready-steady-write.

Signed: ______________________ Date: ______________________

Summary

The aim of this dissertation is to investigate the relationship between sentiment extracted from social media data and returns in financial time series at a firm-level for different time frequencies, in particular examining its effect at high-frequency. In doing so it leverages computational linguistics to automatically generate a domain-specific affect dictionary from a corpus of tweets for the purpose of domain-specific sentiment analysis. Furthermore, this dissertation aims to investigate the difference in the relationship between sentiment extracted from social media and formal media.

Sentiment analysis is the identification and categorisation of opinions expressed in a piece of text with the aim of determining the writer's feelings toward a particular topic, product or firm through the use of natural language processing, text analysis and computational linguistics. A popular application of sentiment analysis is the analysis of the expectation of the price of a financial asset based on the sentiment expressed about the asset. Thus, sentiment analysis often plays an important role in high-frequency trading (HFT) algorithms which aim to forecast the movement of stock prices at an incredibly high-frequency based on many external variables, including sentiment.

To reach the research goals firm-specific Twitter data was collected over a 6 month period, for analysis. The resulting tweets were aggregated at different frequencies and negative sentiment values were calculated based on the text in the tweets using the Rocksteady affect analysis system. Similarly, a financial time series for the same period was acquired and logarithmic returns were calculated. The resultant time series were then aggregated and aligned by date and time and examined using four different vector autoregressive (VAR) models, which individually examine the independent effect of different exogenous variables such as negative sentiment and media volume on financial returns.

An analysis of approximately 360,000 tweets about Ryanair, the focus of the case study, collected between October 17th, 2017 and April 2018 found that negative sentiment plays a small role in explaining returns of an asset. It shows that the explanatory power of sentiment varies at different frequencies and is highest at a daily frequency due to the aggregation of sentiment throughout the day and because changing volatility across the 6-month period made it difficult to model at high-frequency. However, a week-by-week analysis found that negative sentiment is indeed significant during certain weeks and is most significant during periods of consistent volatility. It shows strong significance when volatility is consistently high such as during a threat of strike action and also when it is consistently low such as the period leading up to Christmas.

A comparison of sentiment in formal media versus sentiment in social media found that it is easier to extract sentiment from formal media such as news articles and newswires due to their formal language and larger volume of text. However, negative sentiment extracted from formal media was not significant in the vector autoregressive models examined but a first-order lag of article volume was. This suggests that article volume could be used as a proxy for investor sentiment instead.

Future research will be carried out into the investigation of when and why high-frequency sentiment is significant, by further analysing its effect on shorter time periods and by varying the conditions. If sentiment is to be used as part of a high-frequency trading strategy it would need to be possible to determine when to take sentiment into account and how much influence it should have. In addition, this work focused mostly on a firm-specific analysis in the commercial airline industry. It would be interesting to determine whether the results are consistent with other firms in the commercial airline industry, and how much they might differ in another industry altogether.

Abstract

High-frequency trading (HFT) is a financial trading strategy that is growing in popularity among major trading firms. At the turn of the 21st century, HFT trades had an execution time of several seconds, whereas by 2010 this had decreased to milli- and even microseconds. This recent growth in HFT is paralleled by a massive increase in user-generated social media data, where every day hundreds of thousands of opinions are shared in the form of discussion, reviews and social media statuses. This research aims to leverage the rapid growth in social media, specifically Twitter, to investigate the effect of sentiment in high-frequency financial markets.

This thesis presents a system that automatically retrieves and aggregates firm-specific Twitter data for the purpose of sentiment analysis. Leveraging computational linguistics, it computes a domain-specific affect dictionary from the corpus of tweets for use during sentiment analysis. It then extracts a negative sentiment time series using the Rocksteady affect analysis system at a daily, 5-minute and 1-minute frequency. The resultant time series are then aggregated with firm-specific financial stock returns at their respective frequencies. Finally, the time series are examined using four different vector autoregressive (VAR) models, which individually examine the independent effect of different sentiment variables on returns.

The analysis of the system, using Ryanair as a case study, found that sentiment extracted from social media, specifically Twitter, plays a small but significant role in explaining the returns of an asset but its explanatory power varies at different frequencies. The research shows that the VAR models fit better at a daily frequency than at high-frequency due to the volatile nature of the time series. However, a week-by-week analysis shows that high-frequency sentiment extracted from tweets contributes significantly to modelling the returns in certain situations, in particular during periods of consistent volatility. Thus, negative sentiment found in tweets can be very appropriate for use in high-frequency trading strategies but the circumstances under which it is significant need to be thoroughly examined in future work.

Acknowledgements

I would like to thank a number of people who have made this dissertation possible. First, I would like to sincerely thank my supervisor Professor Khurshid Ahmad for introducing me to the topic of sentiment analysis and financial markets as well as for all his help, guidance and valuable time throughout the project. I would also like to include a special note of thanks to all those involved with the Master's programme at Trinity College Dublin, in both the Engineering department and Computer Science department. In particular, Dr Mike Brady, MAI coordinator in the Department of Computer Science.

I am also very grateful for all the help and encouragement I received from my friends and classmates over the years, in particular, I would like to thank Sid Gupta for his support and advice both in conducting the research and during the writing of the dissertation.

Finally, I would like to express my sincere gratitude to my family for their advice and endless encouragement in my academic endeavours, without them this dissertation would certainly not have been possible.

Contents

List of Figures

1 Introduction

The widespread adoption of the Internet, in recent years, has changed the way people express their views and opinions. More than ever before, people are expressing themselves online through blog posts, online discussion forums, product review websites and social media sites. Through the analysis of this data, the general opinion of a product, company, brand or public personality can be established. However, the amount of user-generated content is much too large for a user to analyse, so to automate this various sentiment analysis techniques can be employed. A recent, and growing, application of such sentiment analysis techniques is mining micro-blogging data to forecast stock market behaviour, often at high-frequency, which has yielded promising results[1, 2, 3]. Such systems are based on the argument that a financial model that accounts for investor sentiment and media attention can potentially be used to predict stock market variables such as returns, volatility and volume. Investor sentiment is the view or opinion that is held or expressed by an investor, in particular, it is the belief about future cash flows and investment risks that are not justified by the facts at hand[4]. A common proxy for investor sentiment is sentiment extracted from traditional media such as articles from financial news sources including the *Financial Times* and *Bloomberg*. However, a new form of proxy for investor sentiment is spontaneous sentiment harvested from high-frequency social media sites such as Twitter. Spontaneous sentiment is the opinion or view that is expressed as a result of a sudden impulse or inclination, without premeditation. This type of sentiment surfaces faster than that which is published by news sources and traditional media, allowing for high-frequency analysis. With 1/60th of the world's population on Twitter, accounting for over 330 million monthly active users, publishing more than half a billion tweets every day[5], Twitter has emerged as the perfect source of high-frequency sentiment.

1.1 Sentiment Analysis

Sentiment analysis is the identification and categorisation of opinions expressed in a piece of text, with the aim of determining the writer's feelings toward a particular topic, product or firm through the use of natural language processing, text analysis and computational linguistics. Sentiment analysis tools have proven particularly useful in applications such as marketing, reputation management and financial analysis. Typically sentiment analysis tools examine a large collection of text, often in real time, to make decisions about the writer's affect towards the topic in question. Such collections could include formal media such as news articles and newswires, press releases, blogs - both personal and corporate, online forums and message boards, online product reviews or social media streams such as Twitter and Facebook.

The sentiment of a piece of text can be expressed numerically in several ways. The simplest of which is a linear scale between -1 and 1 where -1 is entirely negative, 0 is neutral and 1 is entirely positive. A more complex representation of sentiment is a 3-dimensional vector of *polarity*, *strength* and *activeness*. *Polarity* measures how negative or positive the text is, *strength* measures how strong or weak it is, and *activeness* measures whether it is active or passive.

1.2 High-Frequency Sentiment Analysis

A recent application of sentiment analysis is the analysis of the expectation of the price of a financial asset based on the sentiment expressed about the asset. Typically, asset price dynamics are investigated based on quantitative data using historical prices and other variables such as sentiment and media-attention. Sentiment analysis in financial markets is most frequently used by large financial and trading institutions such as *Virtu Financial*, *Tradebot* and *Citadel LLC* to perform *automated* high-frequency trades. Such algorithmic trading systems aim to forecast the movement of stock prices at an incredibly high-frequency based on many external variables, including sentiment. High-frequency trading (HFT) is characterised by high speeds, high turnover rates, and high order-to-trade ratios with very short-term investment horizons, aiming to capture sometimes a fraction of a cent in profit on every trade. At the turn of the 21st century, HFT trades had an execution time of several seconds, whereas by 2010 this had decreased to milli- and even microseconds[6].

This recent growth in HFT requires high-frequency sentiment analysis, among other things. Traditionally sentiment analysis is performed on formal media such as newspapers and online articles but the rate of publication of such media is not high-frequency and thus can not contribute to high-frequency sentiment analysis. However, the widespread adoption of social media sites and abundance of user-generated content published through such sites allows for automated high-frequency information monitoring[7]. Recent estimates indicate one in five tweets discuss products or brands[8], which makes Twitter an incredibly effective and abundant source of high-frequency sentiment.

1.3 Twitter

Twitter is one of the most popular micro-blogging services which allows users to publish short messages with up to 280 characters (previously 140 characters before November 2017). Such messages are called *tweets*. Tweets are visible on a public message board, via twitter.com, Twitter's mobile applications and various other third-party applications. Twitter allows users to follow (subscribe to) a selection of their favourite authors/friends or search for tweets containing a specific keyword or hashtag. A hashtag, (e.g., *'#earnings'*) is a token, included in many tweets, that allows users to associate that tweet with a relevant topic or category. This allows such tweets to be more easily discovered. Similarly, traders have adopted the convention of tagging stock-related messages with a dollar sign followed by the relevant ticker symbol (e.g. *'$RYAAY'*). These are commonly known as *cashtags*. Twitter treats *cashtags* exactly like *hashtags* in terms of categorising tweets.

When a user publishes a tweet it appears in the timeline's of all of the users who *follow* the user, in reverse chronological order (i.e. most recent first). A given user, *UserB* can *retweet* a particular tweet of *UserA* which prepends 'RT' to the tweet and UserA's username, prefixed with the '@' symbol, and publishes the tweet to the followers of UserB. This (potentially dramatically) increases the visibility of the original tweet and can lead to the viral spread of content, information and thus sentiment. Finally, a given user can *mention* another user by specifying the username of a user (prefixed with a '@' symbol) in the body of a tweet. This directs the tweet to the mentioned user, but it still appears in the timelines' of all users who follow the original publisher.

1.4 Research Objectives

The aim of this dissertation is to investigate the relationship between sentiment extracted from social media data and returns in financial time series at a firm-level for different time frequencies, in particular examining its effect at high-frequency. It aims to do this leveraging computational linguistics to automatically generate a domain-specific affect dictionary from a corpus of tweets for the purpose of domain-specific sentiment analysis. The primary objective is to determine whether high-frequency sentiment has an effect on asset prices or vice-versa and to investigate how that effect changes at different frequencies. The research further aims to determine the relationship between the volume of trades and the volume of social media data, and whether that correlates to volatility in returns. Finally, this dissertation aims to investigate the difference in the relationship between sentiment extracted from social media data and formal media. This work does not attempt to predict market prices through sentiment, but simply assesses the role of sentiment in explaining returns, and/or the role of returns in explaining sentiment at different frequencies.

To achieve these goals, firm-specific Twitter data is collected over a 6 month period, for analysis. The resulting *tweets* are aggregated at different frequencies and negative sentiment values are calculated based on the text in the tweets. Similarly, a financial time series for the same period is acquired and logarithmic returns are calculated. Vector autoregression is used to determine the relationship between the financial time series and the sentiment time series, and that relationship is analysed at different frequencies to meet the research objectives outlined above.

1.5 Key Findings

The results found that the vector autoregressive models fit better at a daily-frequency than high-frequency due to the volatile nature of the time series. The changing volatility across the 6-month analysis period made it difficult to model at high-frequency. However, a week-by-week analysis shows that high-frequency sentiment extracted from tweets contributes significantly to modelling the returns in certain situations, in particular during periods of consistent volatility. Furthermore, the sentiment found in formal media was insignificant across a 10-year analysis but it was found that previous day article volume could act as a proxy for investor sentiment instead. Further research will be carried out to examine the exact conditions for which high-frequency sentiment is significant.

1.6 Dissertation Structure

This thesis begins by outlining the motivation of the research, in the context of existing work in the field (Chapter 2). It provides an account of previous academic research in the area, describing and discussing existing solutions. This is followed by a discussion of the research methods and the approach taken to evaluate the effect of sentiment at different frequencies (Chapter 3). This discusses the methodology and the rationality for the design choices made, outlining in detail, both the systems used to collect and analyse the data. It then presents a case study of Ryanair Holdings PLC (Ryanair) by examining the financial history of Ryanair, and the effect social media had in forecasting its stock price behaviour (Chapter 4). This includes the presentation of the findings of the study. Finally, it follows with a discussion of the results and the strengths, weaknesses and limitations of the approach before concluding by describing the implications of the research, and outlining recommendations for further research (Chapter 5).

2 Motivation and Existing Work

Using public data to proxy investor sentiment for the purpose of stock price prediction is not a new concept and whether stock returns can be predicted has long been a debate. This chapter presents existing work in the field of sentiment analysis, behavioural finance and content analysis and outlines the motivation behind the research objectives and the methods employed to reach those objectives.

Early research based on the Efficient Market Hypothesis (EMH) argues that stock returns are indeed random and cannot be predicted[9]. However, there is a small, but notable set of related work that shows that new information, especially news, has a major influence on stock returns and quickly leads to stock price changes[10, 11]. Behavioural finance shows that high negative tone expressed in the media in relation to a particular firm causes investors to rationally revise downwards their estimates of fundamental firm value[11]. This work extracted sentiment from a corpus of over 5.5 million news articles on 20 large US firms over a 10 year period to determine the relation between media-expressed firm-specific sentiment and firm-specific returns. It found that, at times, media-expressed tone had a significant impact on firm-specific returns, and that in some instances the impact endured, confirming that media-comment can contain both sentiment and news.

Similar work shows that alternative measures of sentiment such as the US-based *Michigan Consumer Sentiment Index* (MSCI) have a strong correlation with the financial markets, which suggests that consumers anticipate changes in the wider economy through feelings and emotions[12]. The same research shows that the frequency of negative affect words in news articles, over a given period, accounts to statistically significant changes in the Danish Consumer Index and in the value of stocks of individual companies.

2.1 Media Volume

Other research has investigated the relationship between the volume of media-attention on a firm-specific level and the firm-specific returns of stock's trading levels. These works aim to prove the concept that *'No news is good news'*. There are arguments that traffic volume on internet message boards is correlated with major changes in stock prices[13] and that stock prices fluctuate unpredictably just before and after macroeconomic announcements such as GDP, inflation rates and the release of financial reports. It is argued that a proxy for this uncertainty could be the volume of firm-specific articles, posts on internet message boards, newswires or indeed tweets on Twitter. Corea shows that the posting volume on Twitter about specific firms has a greater forecasting power when augmented into the existing models[14]. Not only did it have a strong impact on price forecasting but also a directional prediction, which suggests that it is indeed more valuable to consider how much people talk about a firm rather than what they think about it. Similarly, Oliveira et al. found a positive correlation between tweets posting volume and stock market variations[1] and Alanyali et al. found a positive correlation between the number of mentions of a company in the Financial Times and its stock's trading volume[15].

2.2 Sentiment Analysis of Tweets

As mentioned above, recent work has begun using Twitter data to proxy investor sentiment. Tweets differ from traditional news articles because they are constrained to 280 characters (previously 140) which encourage users to be brief and get to the point[16]. This makes it a good platform to foster the sharing of emotion[17]. As Twitter is a social network, information can spread quickly or slowly depending on the reach of the author, i.e. how many followers the user has. Sul et al. found that firm-level sentiment that spreads rapidly through social media is more likely to be quickly incorporated into stock prices while sentiment that spreads more slowly takes longer to be incorporated and thus is more useful in predicting stock prices in future days[18]. The research showed that sentiment in tweets from users with fewer than 171 followers (the median number), and which were not retweeted, had the greatest impact on future stock returns. This shows that the network effect has a significant effect on stock returns. Other methods, such as machine learning[19], have been used to extract sentiment from tweets and again has found it to be productive of stock returns several days later. Corea also examined the effect a user's reach had on tweets by taking the users *Klout Score*[20] into account[14]. The Klout Score is a number between 1-100 that represents a user's influence. The more influential they are, the higher their Klout Score. Intuitively a user with a high influence (Klout Score) would affect the opinion of more investors than a user with low-influence. The study found that the Klout Score was significant when used in the financial model of stock returns which implies that a user's reach has an effect on the impact of their sentiment.

To date, almost all of the work which extracts sentiment scores from Twitter data to model stock returns has examined only certain users or considered only users which are considered *financially literate*. Corea et al. examine the minute-by-minute sentiment of tweets and only includes tweets which contain the companies stock exchange ticker symbol. The investment community, when discussing financial news on Twitter, have adopted to tagging stock-related tweets with a dollar sign followed by the firm's ticker symbol i.e. $*GOOG*. Both [14] and [18] have limited its analysis to tweets which have adopted this convention, as a way of limiting the user base to financially literate users. Arguably this limitation excludes a large number of users that may be expressing sentiment, both positive or negative. The effect of including the larger population into the analysis has not yet been studied in detail. Future work should examine the impact of the sentiment of 'financially illiterate' users in financial models. Influential people, in the context of social-media, such as Rihanna or Kylie Jenner, while considered financially illiterate can have a huge effect on the financial markets. It is claimed that one tweet made by Kylie Jenner, an American reality television personality, model and social media personality, in which she casually claimed to no longer use Snapchat, caused the share price of Snap Inc. (SNAP) to fall by almost 8%, a total market loss of $1.3bn[21]. While Snapchat's stock-price was falling prior to her tweet, it almost certainly caused the remaining landslide. The tweet was published to her 24.5 million followers, *liked* over 375,000 times and *retweeted* over 75,000 times. Similarly, singer/songwriter Rihanna published a *story* on her Instagram account criticizing an offensive advertisement that was published on Snapchat in which she was personally mentioned. Her remark sent SNAP stock down nearly 4%, decreasing its market value by nearly $800 million[22]. These examples show that even 'financially illiterate' users, can have a massive effect on stock prices due to their social impact. Further research can be done in this field, as the impact of social media personalities grows, starting with the inclusion of 'financially illiterate' users.

2.3 Large-scale Sentiment Analysis

Behavioural finance literature argues that there is a need to create a system that can mine sentiment in large volumes of multi-modal (and perhaps multi-lingual) data and aggregate it with financial and economic data. It argues that traditional prediction methods have failed in recent years. If you examine some of the elections of the last 5 years in the USA and the UK, the results returned have not always been anticipated, sometimes by very wide margins. Traditionally, opinion pollsters have told us what the public thinks about a political party, or leader and its policies. These pollsters have suffered significant criticism as of late due to their inaccurate predictions. This suggests that it is more important in how a leader or party is perceived than their specific policies or intentions. This same argument can be applied to stock markets as of recent. Due to the increased accessibility of the stock market to the general public and the increase in availability of news sources, the opinion of investors is perhaps more significant in modelling the return of an asset than before. Arguably, this sentiment is only calculable through mining large volumes of spontaneous high-frequency sentiment perhaps in combination with traditional news sources.

2.4 Sentiment Analysis Methods

As previously mentioned, sentiment analysis of text is typically performed on news articles, which often consists of several (mostly) grammatically correct paragraphs of text providing plenty of context. Performing sentiment analysis on shorter texts, such as single sentences, as found in reviews or tweets is challenging because of the limited contextual information[23], the presence of emoticons, slang words and often intentional misspellings[24]. Preethi et al. explore the use of *Recursive Neural Networks* (RNN) within a deep learning system for the sentiment analysis of short online reviews. The resulting system increased the accuracy of the sentiment analysis compared to previous methods, such as *knowledge base* techniques. Sentiment analysis can be performed using two strategies, a *Knowledge Base* approach or a *Machine Learning* approach. A knowledge base approach requires a dataset of predefined emotions for terms and an efficient knowledge representation for identifying sentiments. A machine learning approach develops a sentiment classifier that classifies sentiments using a training set[24]. This approach does not require a predefined database of emotions and is thus simpler than knowledge base techniques.

There are several different approaches to knowledge base techniques. Turney used a *bag-of-words* approach for sentiment analysis in which the relationships between individual words are not considered[25]. The sentiment of every word is calculated and the document's sentiment score is an aggregation of the value for each word. The lexical database, WordNet[26], was used by Kamps et al. to determine the emotional content of a word along different dimensions[27]. WordNet is a lexical database of English words, where nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms. Kampes et al. developed a distance metric based on WordNet and determined the semantic orientation (positive or negative) of adjectives. It becomes clear that the knowledge base approach is difficult due to the requirement of a huge lexical database. As twitter consists of a huge amount of data, with content spanning many domains sentiment analysis using knowledge base techniques becomes tedious and erroneous[24].

In contrast, machine learning techniques don't depend on such a database. Techniques such as *Naive Bayes* (NB), *Maximum Entropy* (ME) and *Support Vector Machines* (SVM) are used to classify the input feature vectors into corresponding classes of sentiment. Many features exist, including Term Presence, Term Frequency, negation, n-grams and Part-of-Speech[28]. It is argued that Naive Bayes works well for certain problems when used with highly dependent features[29]. A new model was introduced by Niu et al. in which efficient approaches are used for feature selection, weight computation and classification[30]. Using the Bayesian algorithm, classifier weights are adjusted by making use of a representative feature and a unique feature. A 'Representative feature' is one which the information represents a class, whereas a 'Unique feature' is one which the information helps distinguish the class. The probability of each classification is calculated using those weights, improving the Bayesian algorithm.

Another system, developed by Barbosa et al. consists of a 2-stage automatic sentiment analysis method for classifying tweets as subjective or objective, after which they are further classified as positive or negative[31]. A probability model based on the user's influence to determine sentiment analysis in tweets was proposed by Wu et al.[32]. For each tweet that contains an `@username` in the body, it is considered to have an influencing action and it contributes to influencing probability. Similarly, any tweet that begins with `@username` is considered a retweet, representing an influenced action, contributing to the influenced probability. The study found a strong correlation between these probabilities.

The *Rocksteady* system, used in this project, depends on a lexical database or set of dictionaries with accompanying sentiment scores. As described above, this might not be accurate in calculating the sentiment in short bodies of text such as tweets. Further research might be necessary using a machine learning technique, or after developing a dictionary of frequently used Twitter words.

2.5 Twitter-Specific Text Analysis

Work by Neethu et al. proposes the preprocessing of tweets before performing sentiment analysis, particularly when using a feature vector with a machine learning approach[24]. The preprocessing steps included removing any URLs and avoiding misspellings and slang words. All misspellings are avoided by replacing the repeated characters with the correct spelling. Slang words are not removed, as they contribute heavily to the emotion and sentiment of a tweet. They propose maintaining a *slang word dictionary* to replace slang words with their associated meaning. Domain information would need to be learnt to form such a dictionary. Furthermore, Twitter-specific features are extracted from the tweet's body and included in the feature vector. Emoticons are both positive and negative and thus are given different weights. Existing work by Hogenboom et al. could be exploited to calculate the sentiment of emoticons in text[7]. Hashtags are further extracted and sentiment analysis is performed separately, as they too can be both positive or negative. However, without spaces, hashtags can be difficult to decipher. After preprocessing the tweets and extracting the features they are left with simple plain text, which is much easier to process and leads to more accurate results.

One approach to analysing text is *Part-of-Speech* tagging. Part-of-speech tagging (POS tagging) is the process of automatically marking up a token in a text as corresponding to a particular part of speech based on both its definition and its context[33]. The different POS tags include nouns, verbs, participles, articles, pronouns, prepositions, adverbs and conjunctions. Most POS taggers are trained from treebanks in the formal media/newswire domain, thus standard POS taggers perform poorly on out-of-domain data such as tweets. Tweets published on twitter pose additional challenges due to their conversational nature, lack of conventional orthography and strict 280 character limit[34]. For this reason, Gimpel et al. developed a POS tagger for Twitter that handles twitter specific language structures such as emoticons, URLs, twitter specific tags, *at-mentions*, hashtags, as well as slang, abbreviations and multi-word abbreviations that would not be found in formal media and can not be mapped to a single traditional POS tag. To do this additional POS tags were included in their system in addition to the traditional tags described above. This allows for a better understanding of individual tweets and the word distribution in a collection of tweets.

As mentioned above, emoticons and emoji are crucial in calculating sentiment in short bodies of informal text such as tweets. Emoticons are a pictorial representation of a facial expression using characters, usually punctuation marks, numbers and letters, to express a person's feelings or mood. For example, the emoticon for a smiley face is $:-)$ or simply $:)$, similarly, a sad face is represented by $:-($ or $:($. Studies have found that whenever emoticons are used their associated sentiment dominates the sentiment conveyed by the textual cues, and forms a very good proxy for *intended* sentiment. Research has demonstrated that humans are strongly influenced by non-verbal cues in face-to-face communication[7], and non-verbal cues such as facial gestures, body language have been shown to dominate verbal cues in face-to-face communication in the case when both types of cues are equally strong. Hogenboom et al. found that emoticons are used in three different ways.

- Emoticons can be used to express sentiment when sentiment is not conveyed by with positive or negative words in a text segment
- Emoticons can further stress sentiment by intensifying the sentiment that is conveyed by sentiment-carrying words
- Emoticons can disambiguate sentiment in cases where sentiment associated with sentiment-carrying words needs to be negated

This emphasises the importance of understanding the effects of emoticons and emoji in social media. Until recently most lexicon-based sentiment analysers discarded all non-text characters. In short casual texts such as tweets this could entirely change the sentiment of the text and thus work such as that by Hogenboom et al. will need to be exploited.

3 Method

This chapter describes the implementation and development of a system for generating a sentiment time series from social media and traditional news articles and the incorporation of those time series into an autoregressive statistical model with financial data. The system consists of four main phases, *Data Retrieval*, *Data Processing*, *Content Analysis* and *Statistical Analysis*. Each phase and component is discussed in detail and the rationale for the design choices made are provided. An overview of the system can be seen in Figure 3.1 below.

3.1 Data Retrieval

The first phase of the system deals with collecting, structuring and aggregating text documents and financial data from online sources. Collecting structured and unstructured digital text using a computational process means a large volume of text can be collected relatively easily[35]. Collecting and storing any accompanying metadata and structuring the text in a consistent manner such as through a JSON schema allows a corpus of text to be built and processed more efficiently than when left unstructured. For the purpose of this research, it was necessary to collect three forms of data; social media data, formal media data and financial data. The process of collecting and aggregating each is discussed in the following sub-sections.

3.1.1 Social Media Data Retrieval

Social media data in the form of tweets from Twitter were collected and aggregated using the Twitter Search and Streaming API. Twitter provides a free Search API that searches against a sampling of recent Tweets published in the past 7 days. A Python client was developed to retrieve tweets based on different keyword searches. The client retrieves as many tweets as the API returns, while respecting the API limits of 180 queries every 15 minutes, after which it uses the Streaming API. The Streaming API returns tweets published

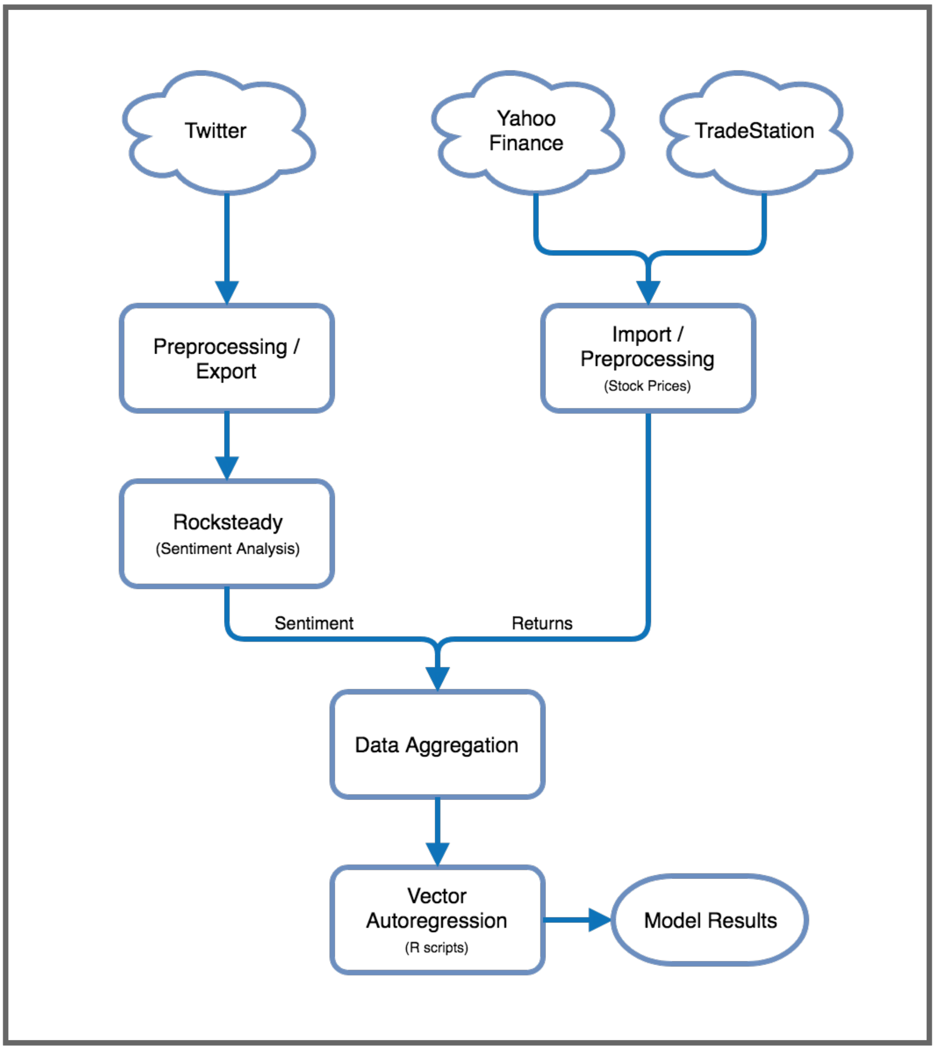

Figure 3.1: An overview of the system implementation consisting of the three main components; data retrieval and preprocessing, content analysis and statistical analysis

in real-time based on certain filters/keywords in a low-latency manner. Using a combination of the Search API and the Streaming API ensures the system is as exhaustive as possible given the constraints. Each tweet retrieved is processed and saved alongside its relevant metadata in a MongoDB database. MongoDB is a free and open-source document-oriented database program which uses JSON-like documents. This allows the tweets to be queried and aggregated by date, text, user, length, number of retweets and more.

3.1.2 Formal media Data Retrieval

News articles and web publications were collected and downloaded in bulk from *LexisNexis* for the specified query. In this case, all English language publications were queried where the specified keywords appeared in the title and duplicate publications were removed. Access to this database was covered under an academic license obtained by Trinity College Dublin

```
{
  "text" : "Ryanair grows in the Middle East with new Jordan
     flights https://t.co/WicJlswzfZ",
  "created_at" : ISODate("2018-02-04T17:45:04.000Z"),
  "id" : NumberLong(960207594190262273),
  "favorite_count" : 116,
  "user" : {
      "users_statuses_count" : 168438,
      "user_id" : 759251,
      "user_followers_count" : 39307093,
      "user_screen_name" : "CNN",
      "user_name" : "CNN",
      "user_verified" : true,
  }
}
```

Listing 3.1: Example of a tweet by CNN that has been retrieved and structured in JSON and accompanying metadata

(TCD). Access to LexisNexis is available via an API but only with a special licence, not covered under the academic licence held by TCD, thus this collection and aggregation was performed manually. The publications are downloaded in chunks of 500 documents in unstructured text files, a requirement of the sentiment analysis system that is used and discussed in Section 3.3. To simplify the task of processing and aggregation, a LexisNexis parser was developed such that the contents of each article could be saved into a MongoDB database in JSON format. As before, this allows the articles to be easily queried or grouped by date, author, publisher and text.

3.1.3 Financial Data Retrieval

Firm-specific financial stock data was required for different time periods and at different frequencies. For daily-frequency data, the historical stock prices could be downloaded from *Yahoo Finance*. For convenience, all daily historical data was downloaded for the given firm such that it could be analysed over different periods if necessary. This included the *open* and *closing* prices for each trading day as well as the daily *high*s, *low*s and trading volume. To allow for high-frequency analysis, 5-minute frequency and 1-minute frequency data was downloaded from *TradeStation*. This data included the same values per time period.

3.2 Data Preprocessing

Before performing sentiment analysis or statistical analysis on either corpus of text some pre-processing was required to aggregate the data and filter out any *off-topic* data points.

3.2.1 Language Detection

The metadata that arrived with each tweet from the Twitter APIs included the language of the user who published the tweet, but after a quick analysis of the data, it was discovered that this was not always consistent with the language of their tweets. Given that this research is only interested in English language text, each tweet needed to be reclassified into its respective language before being stored. The language was determined using the *langid* library developed at the University of Melbourne by Lui & Baldwin[36]. The library identifies the most probable language for a given tweet and accompanies it with a confidence score in the range of 0 to 1. The system accepted all tweets determined to be English with a confidence score above 0.9. This was found to be highly effective in determining the language of individual tweets and the number of valid tweets discarded was negligible.

3.2.2 Tweet Corpus Export

The corpus of tweets needed to be exported in a format that could be processed by the sentiment analysis system described in Section 3.3. This required the development of a script that would process the database of tweets and print them to text files, with the relevant meta-data in the same unstructured format as LexisNexis. Tweets were exported in chronological order (oldest first) in blocks of up to 150,000 tweets so that the system wasn't overloaded with too much data. Before writing the tweet's contents to the file the body of the tweet was cleaned to remove any URLs, mentions, and newline characters as suggested by Neethu et al[24]. Hashtags and emoticons/emojis were left *as is*.

3.3 Content Analysis

Content analysis is used in the text analysis component of the system to extract and measure information from both text corpora. An existing sentiment analysis solution, named Rocksteady, developed by the Trinity College Dublin is used for sentiment analysis. Rocksteady is a lexicon based, affect analysis system that uses a combination of general purpose *affect* dictionaries, like Stone's General Inquirer (GI) Dictionary, and an optional domain-specific dictionary[37]. It is based on the work done by computational linguists in the past half-century, whereby it identifies grammatical categories of words within text to a high degree of accuracy. The system understands that *rose*, *rise*, *risen* and *rising* are the morphological variants of the same root *rise*. Similarly, Rocksteady can identify that the word *share*, when used in financial texts, is a noun rather than a verb. This process is known as automatic syntactic analysis[11]. It is necessary to also include a domain-specific dictionary because, in some cases, the meaning of a term changes depending on the context. A term that is typically negative in a general language context may have another meaning in a domain text[35]. For example, the word *crude* is categorised as negative by the GI dictionary, but when used in the context of oil, the term *crude oil* is neither negative or positive. Thus, the inclusion of a domain-specific affect dictionary overwrites the sentiment value of the standard GI dictionary. Rocksteady is predominately used to evaluate the frequency of words categorized as *negative* in the GI dictionary. This category is the main sentiment or affect category studied in the domain of finance, primarily due to theories of overreaction from investors to negative news, the so-called asymmetric response to news[38]. The GI dictionary contains a unique list of 2,005 single word terms. The list of financial negative and positive words by Loughran and McDonald[39] is included as a domain glossary to reduce the interpretation of financial terms.

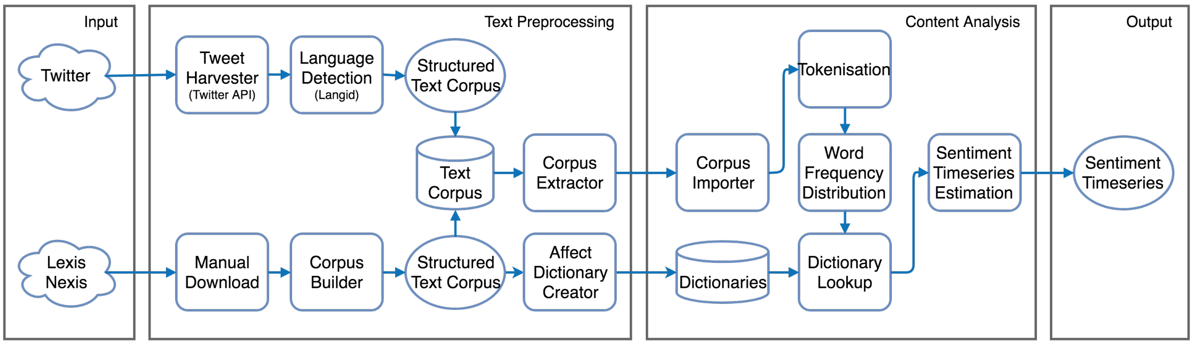

Figure 3.2: System diagram of the text analysis component that outputs a sentiment time series from a corpus of structured text

3.3.1 Automatic Affect Dictionary Creation

To develop an *affect* dictionary for the airline industry, and in particular, Ryanair - the focus of the case study in Chapter 4, the corpus of collected tweets is analysed. Each tweet is parsed with the part-of-speech tagger developed by Gimpel et al. described in Section 2.4. This splits each tweet into words/tokens and identifies the part-of-speech (POS) of each token. Then the frequency of each token for a given POS is calculated. A sample tweet is presented in Table 3.1 illustrating how a tweet is tokenised (split into words, emojis, punctuation, URLs etc) and tagged with a POS tag and the confidence of the tag. To gauge the specificness of a certain word to the domain in question the weirdness ratio of each token is calculated. The weirdness ratio of a word, a concept developed by Ahmad, is a comparison of the relative frequency of a word in a given corpus to the relative frequency of the word in a general corpus, such as the *British National Corpus* (BNC).

$$W_t = \frac{\left(\frac{F_{t,c}}{N_c}\right)}{\left(\frac{F_{t,BNC}}{N_{BNC}}\right)} \tag{1}$$

The formula for *weirdness* is defined above in Equation 1 where $F_{t,c}$ is the absolute frequency of token t in the corpus c, N_c is the number of tokens in c. $F_{t,BNC}$ and N_{BNC} are the absolute frequency of the token and the total number of tokens in the BNC. A weirdness value equal to one suggests that the relative frequency of the token is equal in both corpora. A weirdness value much greater than one suggests the token is used much more frequently (relatively) in corpus c and thus that token could be a domain-specific token and a carrier term. Tokens with a weirdness ratio of *infinity* do not appear in the BNC and are considered *new* words, slang words or misspellings.

The common nouns, identified by the POS-tagger, are then extracted from the distribution of words. High-frequency common nouns with high weirdness values are considered the domain-specific *carrier terms*. To identify these carrier terms the *z-scores* of both frequency (z_f) and weirdness (z_w) are calculated for each term and terms with $z_f > 0$ and $z_w > 0$ are extracted and included in the domain-specific affect dictionary. This allows for the automatic creation of an affect dictionary allowing Rocksteady to analyse domain specific affect words and overwrite negative terms which are neutral in the domain context.

3.3.2 Sentiment Analysis

Rocksteady takes, as input, a corpus of text in the format supplied by LexisNexis downloads and aggregates the 'articles' by date at a chosen frequency. It produces a time series of the percentage of negative terms (or any other affect category) for each given time period. Thus a sentiment time series of the twitter corpus was constructed from the twitter corpus export described in Section 3.2.2 for both a daily frequency, a 5-minute frequency and a 1-minute frequency. Similarly, a sentiment time series of the news corpus was constructed at a daily frequency. A sample of Rocksteady's input/output can be seen in Figure 3.2 below, where three tweets were preprocessed and analysed for negative sentiment.

Table 3.1: Sample tweet by *BALPApilots* tagged by part-of-speech

Token	POS-Tag	Description	Confidence
RT	~	Discourse marker	0.9979
@eu_cockpit	@	At-mention	0.9983
:	~	Discourse marker	0.9800
#Ryanairpilots	#	Hashtag	0.6261
ask	V	Verb	0.9817
for	P	Pre/postposition	0.9966
something	N	Common noun	0.9843
eminently	R	Adverb	0.9998
reasonable	A	Adjective	0.9987
&	&	Coordinating conjunction	0.9950
simple	A	Adjective	0.8252
:	,	Punctuation	0.9757
social	A	Adjective	0.9698
dialogue	N	Common noun	0.9985
+	&	Coordinating conjunction	0.6648
permanent	A	Adjective	0.8784
,	,	Punctuation	0.9983
direct	A	Adjective	0.9277
local	A	Adjective	0.6886
contracts	N	Common noun	0.9995
https://t.co/WicJlswzfZ	U	URL or e-mail address	0.9920

3.4 Statistical Analysis

The pricing data obtained through Yahoo Finance and TradeStation is used to calculate the returns for each time period (daily, 5-minute, 1-minute). Logarithmic returns are used instead of comparing closing prices for normalisation purposes as it allows assets of unequal value to be transformed into a metric that is comparable[35] and logarithmic returns are symmetric. This means positive and negative percent logarithmic returns of equal magnitude cancel each other out and result in no net change, whereas ordinary returns of equal magnitude but opposite signs will not cancel each other out. The logarithmic return at time t is calculated as follows:

$$r_t = ln\frac{c_t}{c_{t-1}} \tag{2}$$

where c_t is the closing price of a financial asset at time t. The pricing time series was then aggregated and aligned with the sentiment time series produced by Rocksteady at the relevant frequencies by the statistical modelling system developed in R, adjusting for time zone differences. The *z-score* of the sentiment variable is calculated, to make it easier to

Table 3.2: Example output after prepossessing and performing sentiment analysis on 3 sample tweets using *Rocksteady*, where terms identified as negative are highlighted in red

Text	Terms	%Negative
Belgian Minister of Mobility confirms Ryanair strategy to avoid paying Brussels Region noise fines https://t.co/qRl9Gn7sV3 https://t.co/WcL8h0u1hE		
belgian minister of mobility confirms ryanair strategy to avoid paying brussels region noise fines	14	21.43%
40 min delay. No communication #ryanair #fail #worstairlineever		
40 min delay. no communication ryanair fail worstairlineever	8	25.00%
flying @tapairportugal is an expensive @Ryanair experience | the most inefficient boarding system I've experienced, they need @Peguha to fix that | in the meantime avoid avoid avoid #frequentflyer		
flying is an expensive experience the most inefficient boarding system i ve experienced they need to fix that in the meantime avoid avoid avoid frequentflyer	25	24.00%

interpret any relationships with the financial variables. Calculating the z-score standardises the series to have zero mean and unit variance, with a distribution centred on zero. The z-score for a time series is calculated as so:

$$z = \frac{x - \mu}{\sigma} \tag{3}$$

where x is the time series or vector of data, μ is the average of x, σ is the standard deviation of x. The statistical analysis system starts by exploring the individual data series by calculating the correlation coefficients between different variables in the time series, including returns, the percentage of negative tokens (sentiment), trade volume and tweet volume. This uses *Pearson's Correlation Coefficient* method to measure the linear correlation between two variables X and Y. It produces a correlation coefficient between -1 and +1, where +1 is a total positive linear correlation, 0 is no linear correlation, and -1 is a total negative linear correlation. The equation used to calculate the coefficients is given below in Equation 4.

$$r = \frac{\sum_{i=1}^{n}(x_i - \hat{x})(y - \hat{(x)})}{\sqrt{\sum_{i=1}^{n}(x_i - \hat{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \hat{y})^2}} \tag{4}$$

Furthermore, the system calculates the autocorrelation of returns for 5 lags using Pearson's method. Autocorrelation measures the linear dependence of a variable with lagged values of itself. The autocorrelation function (ACF) is important in identifying and characterising linear effects that may be present in a time series.

The 5-minute time series is divided into days such that the volatility for a given day can be calculated. In finance, volatility, σ, can be calculated in many ways. Basic volatility, in finance, is defined as the degree of variation of a trading price series over time as measured by the standard deviation of logarithmic returns. According to Rogers and Satchell, in certain situations, such as intraday trading, volatility methods based on high/low, open/close prices are preferred[40]. Thus, Rogers' and Satchell's volatility estimator (Equation 5 below) is calculated for each trading day. Similarly, the 1-minute data is divided into hours and the volatility of each hour is calculated. This volatility time series is a proxy for investor uncertainty, and thus a spike in volatility should represent a period of large investor uncertainty. The respective time series are used to compare volatility against the volume of tweets and trades at different frequencies.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\ln\frac{h_i}{l_i})(\ln\frac{h_i}{o_i}) + (\ln\frac{l_i}{c_i})(\ln\frac{l_i}{o_i})} \tag{5}$$

Finally, a multivariate analysis of the financial time series is performed. Vector autoregression (VAR) is chosen as the method for multivariate analysis, to determine any inter-relationships between financial variables and sentiment variables. The advantage of VAR modelling is that it does not require as much knowledge about the forces influencing a variable as with structural models such as simultaneous equations. The only prior knowledge required is a list of variables which can be hypothesized to affect each other intertemporally. However, before calculating the VAR model, the univariate model, defined in Equation 6 below, is analysed.

$$r_t = \alpha_0 + \alpha_1 r_{t-1} + \epsilon_t \tag{6}$$

where r_t is the return at time t, α is a coefficient, ϵ is the residual term assumed to have constant mean and variance.

Typically autoregressive financial models consider 5-time lags of variables, thus the first autoregressive model examined incorporates a lag of 5 time periods of returns;

$$r_t = \alpha_0 + \alpha L_5 r_t + \rho V_{t-1} + \epsilon_t \tag{M1}$$

Also included in this model is the volume of trades in the previous time-period, V_{t-1}, with ρ, its coefficient. VAR estimates are based on two types of variables; endogenous variables and exogenous variables. Endogenous variables are those that are internal to the system, such as lagged values of returns. Exogenous variables are those that are external to the system such as sentiment or volume of tweets/articles. Both sentiment and volume of media attention are outside factors which are expected to affect the price of an asset. VAR allows for the calculation of a value, such as returns, based on lagged values of itself while incorporating exogenous variables. This allows the relationship between sentiment or media volume on returns to be determined and vice versa. The following 3 VAR models, in addition to Model M2, are examined and analysed at a daily, 5-minute and 1-minute frequency;

$$r_t = \alpha_0 + \alpha L_5 r_t + \beta L_5 s_t + \rho V_{t-1} + \epsilon_t \tag{M2}$$

where s is the negative sentiment time series of either tweets or the news corpus, and β and is a coefficient.

$$r_t = \alpha_0 + \alpha L_5 r_t + \lambda L_5 v_t + \rho V_{t-1} + \epsilon_t \tag{M3}$$

where v is the time series of volume of tweets/news articles, and λ is a coefficient.

$$r_t = \alpha_0 + \alpha L_5 r_t + \beta L_5 s_t + \lambda L_5 v_t + \rho V_{t-1} + \epsilon_t \tag{M4}$$

Model M2 incorporates a sentiment time series as an exogenous variable while M3 includes media volume of either tweets or news articles. Finally, Model M4 includes both sentiment and media volume as exogenous variables.

Vector autoregression captures the linear interdependencies among multiple time series, such as a sentiment time series and a financial time series, by generalising the univariate autoregressive model (Equation 6) and allowing for more than one evolving variable. The evolving variables are introduced independently in Models M2 and M3 to examine their independent impact on the model. In Model M4 the two independent time series, negative sentiment and tweet/article volume are included together to examine their combined impact.

The adjusted R-squared value, $\hat{R}^2$, is calculated and used to compare the *explanatory power* of the different models. It calculates the explanatory power of regression models that contain different numbers of predictors, as in the four models defined above. The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model[41]. The adjusted R-squared increases only if the new terms improve the model more than would be expected by chance. Whereas the R-squared value, R^2, always increases when a predictor is added to the model. Thus, a model with more terms may appear to have a better fit simply because it has more terms, as would be the case with Model M4. In contrast, the adjusted R-squared value decreases when a predictor improves the model by less than expected by chance.

The statistical analysis methods described above are performed on the high-frequency Twitter corpus at a daily, 5-minute and 1-minute frequency to examine the effect of sentiment at different frequencies. In addition to that, the methods are performed on the corpus of formal media at a daily frequency only, to compare the effect of sentiment in formal media against the effect of sentiment in social media.

4 Evaluation and Case Study

Chapter 3 describes the system which combines methods of content analysis and statistical analysis, automatically collecting and aggregating text and financial time series. This chapter presents a case study that is used to evaluate the system.

4.1 Ryanair Holdings PLC

Ryanair Holdings PLC (Ryanair), is an Irish low-cost airline founded in 1984. By 2016, it was the largest European airline by the number of scheduled passengers flown and carried more international passengers than any other airline[42]. Ryanair has been listed on the Dublin Stock Exchange and the NASDAQ Stock Market since 1997. It operates under the RYAAY ticker symbol on the NASDAQ where it opened at $2.18 in June 1997. Ryanair stock currently trades for around $120 ($120.14 as of April 2nd, 2018). The growth over the past 21 years can be seen in Figure 4.1. In the last 5 years they have grown from $40 per share to $120 per share, a percentage increase of 200% and in the last year, they have seen an increase of 42.9%, rising from $84 per share (Figure 4.2). This consistent growth over the last 21 years of trading suggests that Ryanair are in a stable financial position and investors continuously believe in them and their future prospects. However, there are at least three critical dates where Ryanair's share price has fallen a considerable amount; almost as much as 25% in some cases. This is visible in Figures 4.1 and 4.2. Ryanair was chosen as a case study due to the large and consistent amount of media attention they receive, their large social media presence/activity and their stable financial position.

The Summary statistics for time series of returns for Ryanair are shown in Table 4.1. The unconditional mean, μ, shows a low positive average return and the unconditional standard deviation, σ, does not account for the changing variance that is characteristic of prices and returns. Skewness, δ, and kurtosis, γ indicate the shape and symmetry of the returns distribution. *Skewness* is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. Unintuitively, skew does not refer to the direction the curve appears to be leaning; in fact, it is the opposite. *Kurtosis* is a measure of the 'tailedness' of the probability distribution of a real-valued random variable. Kurtosis is often used as a comparison to the kurtosis of a normal distribution, which has a kurtosis of 3. Thus, a value greater than 3 suggests the probability distribution has *fatter* tails than a normal distribution. Ryanair's returns have a small, negative skewness, which suggests that the tail on the left side of the probability density function is longer and fatter than the right side. The returns also have large positive kurtosis, larger than that of the normal distribution, which is typical of financial returns[35]. The statistics suggest the series is *leptokurtic* with long tails and a high peak. This is an observation that is seen in all series of daily financial returns across asset classes[43].

Table 4.1: Summary statistics for time series of returns for RYAAY, where μ is the mean, σ the standard deviation, δ skewness and γ kurtosis. Values *L*1 to *L*5 are values of the autocorrelation function for five lags. *N* denotes the number of observations (trading days).

$\mu(10^{-4})$	$\sigma(10^{-2})$	δ	γ	L1	L2	L3	L4	L5	*N*
29.88	10.76	-0.56	16.14	-0.0104	0.0056	0.0011	-0.0067	-0.0181	5253

Autocorrelation is the correlation of a signal with a delayed copy of itself as a function of delay. In financial data, the autocorrelation function (ACF) of lag returns are typically near-zero, as evident in the ACF of the series in Table 4.1. A stylised fact, according to Taylor, states that returns of an asset are not predictable and there is almost no correlation between returns for different days[43], although, in this series, a small degree of (mostly negative) correlation can be seen. In literature, this slight correlation has been attributed to the calculation of observed returns and issues surrounding how prices are determined in an exchange[43]. Of the five autocorrelation values, first-order autocorrelation is negative, while second- and third-order are positive, and fourth- and fifth-order are again negative. This suggests that mean-reversion occurs, which is the assumption that a stock's price will tend to move to the average price over time.

An understanding of the autocorrelation is important when using autoregressive (AR) models as the AR component includes lagged variables of the dependent variable to account for these dynamics and any potential multicollinearity. The summary statistics indicate that the lagged variables account for the possibility that an observation is influenced by the previous one, despite the low correlation values.

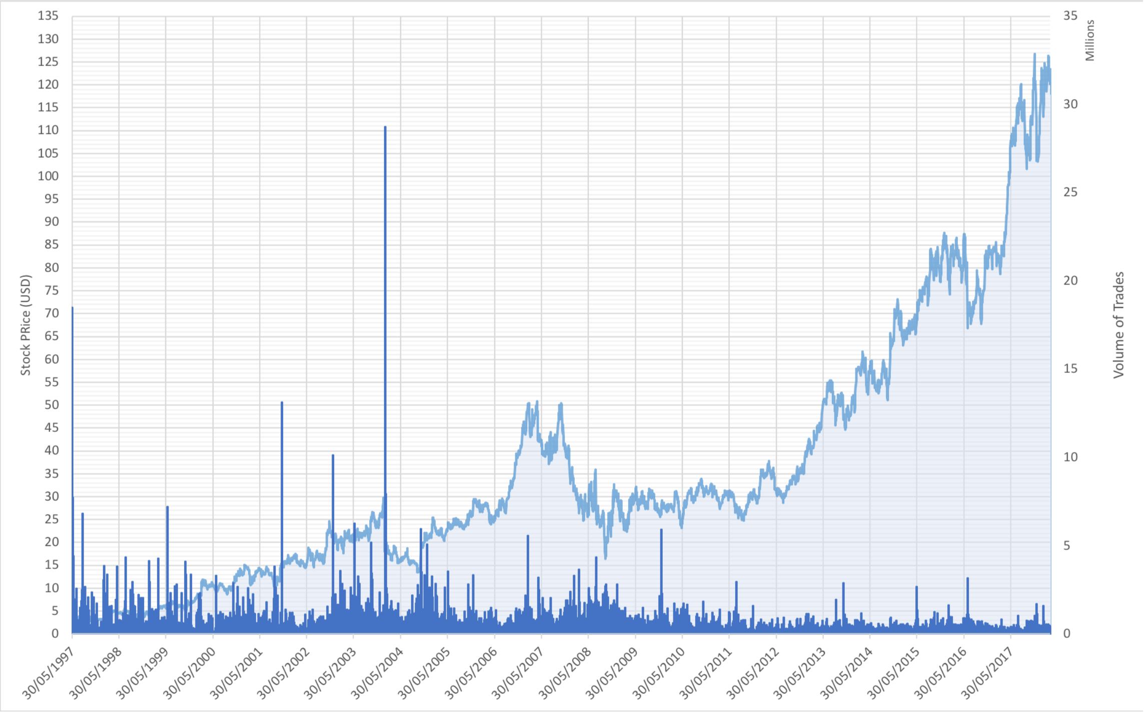

Figure 4.1: Ryanair All Time Share price (light blue line) and Trade Volume (dark blue bars) (n=5235)

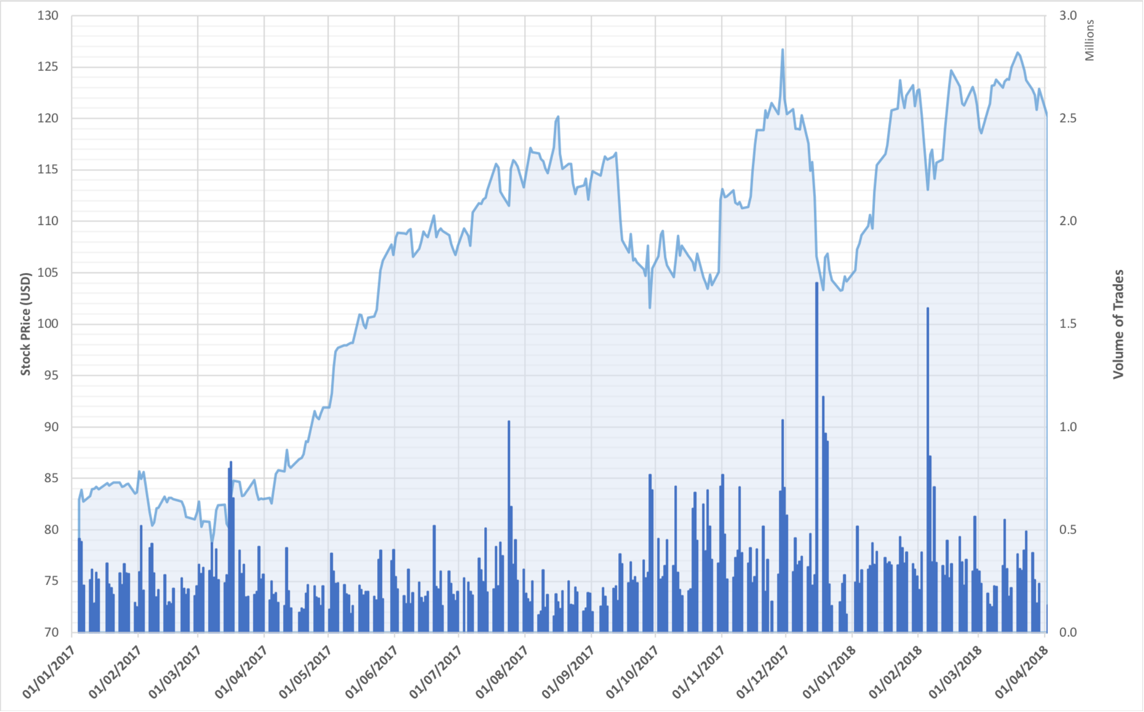

Figure 4.2: Ryanair 1 Year Share price (light blue line) and Trade Volume (dark blue bars) (n=312)

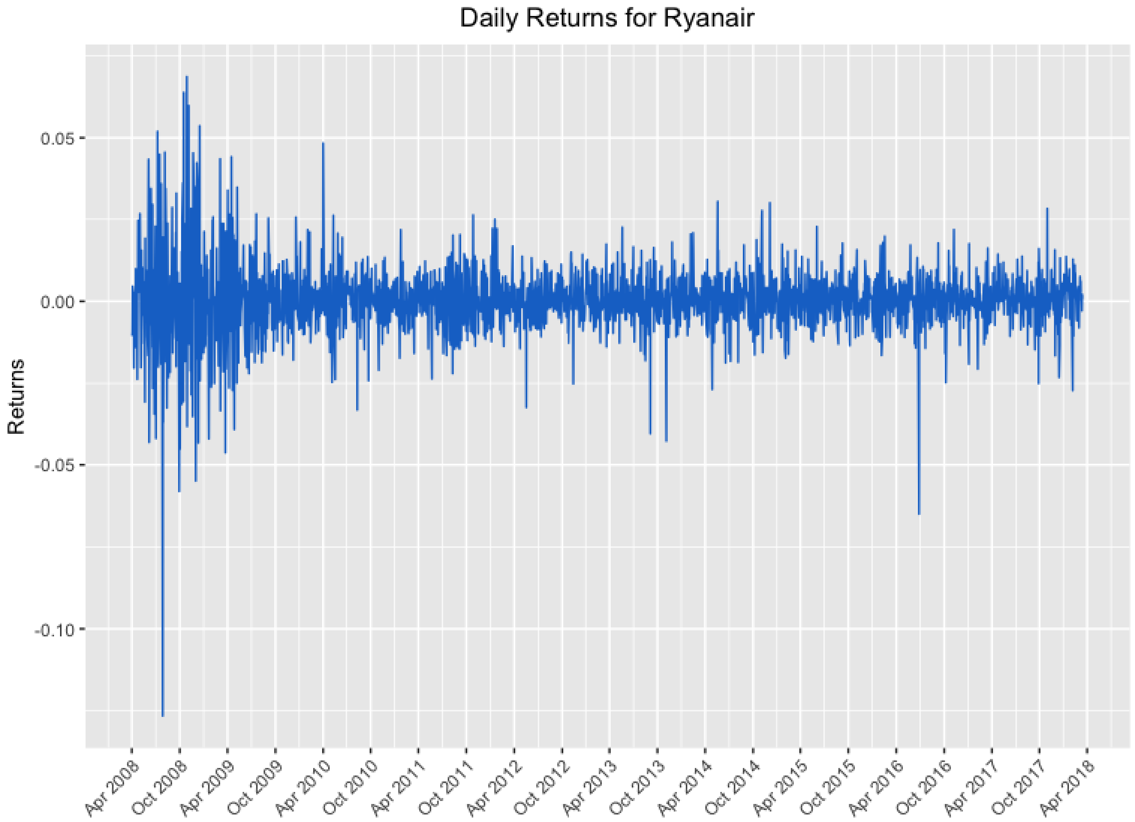

Figure 4.3: Ryanair Daily Returns over 10 years (n=2519)

4.2 Text Data

Between October 16th, 2017 and April 2018 over 585,000 tweets were collected, of which, approximately 360,000 are in English, as determined by the *langid* system. Further analysis of the English tweets shows that approximately 150,000 of those collected are *retweets*. Table 4.2 presents a descriptive summary of the time, in seconds, between sequential tweets. The average time between tweets during the collection period is 43.17 seconds, with a standard deviation, σ, of 114.38 seconds (01:54.34). Excluding *retweets* (RTs) from this analysis provides an insight into the frequency of new information being published on Twitter. Naturally, the average time between sequential *new* tweets rises to 73.66 seconds (01:13.66) with a standard deviation of 179.28 seconds (02:59.28). This implies that, on average, tweets about Ryanair are published at a frequency of approximately 3 tweets every 2 minutes. If tweets published outside trading hours/daytime are excluded, as in the second half of Table 4.2 where only tweets published in the 12-hour period between 9.00am and 9.00pm Monday to Friday were included, the average time between tweets decreases significantly to 26.16 seconds with a standard deviation of 77.60 seconds. The period of (average time between) original tweets decreases similarly, to 44.01 seconds, which is almost in-line with the period of all tweets including *retweets*, with a lower standard deviation of

106.04 seconds (01:46.04), compared to almost 3 minutes when tweets at all times of day are considered. This time period, 9.00am-9.00pm, was chosen because it includes the morning in Greenwich Mean Time (GMT), where Ryanair and its customers are based, and includes the NASDAQ trading hours, 2.30pm-9.00pm GMT or 9.30am-4.00pm Eastern Time (EST).

Table 4.2: Comparison of descriptive statistics of time, in seconds, between sequential English tweets when *retweets* (RTs) are included and excluded, where μ is the average, σ the standard deviation and *N* the number of tweets.

	N	μ	σ	Min	25%	50%	75%	Max
All time								
All	358,411	43.17	114.38	0	6	17	43	25,684
Exc RTs	211,348	73.66	179.28	0	11	31	76	25,684
Between 09:00-21:00 Monday-Friday								
All	214,231	26.16	77.60	0	5	13	32	25,684
Exc. RTs	127,218	44.01	106.04	0	9	24	55	25,684

The breakdown of results of tokenising and analysing each tweet using the POS-tagger by *Gimpel et al.* can be seen in Table 4.3. Common nouns account for 15.7% of tokens, with 756,204 occurrences. As these were the only tokens which were considered for the domain-specific affect dictionary, this led to an exclusion list consisting of 76.88% of tokens, totalling 3,693,517 tokens out of 4,816,481. In total, 52 tokens were extracted from the corpus for inclusion in the domain affect dictionary based on their frequency and weirdness values as described in Section 3.3.1. The top 20 extracted terms, sorted by *weirdness*, are presented in Table 4.4.

Table 4.3: Part-of-Speech breakdown for Twitter Corpus

Syntactic Category	Symbol	Frequency		Percentage
Nominal				
Common Noun	N	756,204	756,204	15.70%
Nominal w/o Common Noun				
Pronoun	O	231,610		
Proper Noun	^	411,121		
Nominal + possesive	S	1,431		
Proper Noun + possesive	Z	5,565	649,727	13.49%
Other open-class words				
Verbs	V	819,115		
Adjective	A	216,399		
Adverb	R	199,545		
Interjection	!	56,551	924,850	19.20%
Other closed-class words				
Determiner	D	309,944		
Pre/postposition	P	521,096		
Coordinating conjunction	&	72,663		
Verb particle	T	15,794		
existential there	X	53,53	924,850	19.20%
Twitter/online specific				
hashtag	#	65,296		
at-mention	@	275,070		
Dicsourse marker	~	68,688		
URL or email	U	124,126		
emoticon	E	13,568	54,6748	11.35%
Misc				
numeral	$	77,433		
punctuation	,	491,045		
other abbrevations	G	35,457	603,935	12.54%
Other compounds				
nominal + verb	L	43,301		
proper noun + verbal	M	0		
X + verbal	Y	106	43,407	0.90%
Total Words		**4,816,481**		
Elimination Word List			**3,693,517**	**76.68%**

Table 4.4: Weirdness Distribution of top 20 common nouns

Token	Frequency	Weirdness
Bellew	1113	22565.04
ex-pilots	742	15043.36
CEO	29998	6269.91
DL	3869	3016.94
email	6095	2873.73
euros	689	2328.14
ops	2120	499.78
cancellations	1537	486.89
playoff	530	429.81
app	3074	402.08
info	3180	328.94
laptop	1060	294.39
refund	4876	288.21
flights	15741	257.78
check-in	583	236.40
pilots	10335	188.43
refunds	689	174.61
COO	530	173.31
flight	35457	139.48

4.3 Correlation

Correlation between time series variables is calculated for all time series to understand how the different variables relate to each other, if at all. As before, weekends and non-trading days are excluded from the analysis for all time series while data outside trading hours (9.30am EST to 4.00pm EST, Monday to Friday) is excluded in the analysis of 5-minute and 1-minute frequency time series. The correlation of the daily, 5-minute and 1-minute time series can be seen in Tables 4.5, 4.6 and 4.7 respectively. The correlation coefficient is scaled between -1 and 1 with -1 indicating a perfect negative correlation, 0 indicating no correlation and 1 indicating a perfect positive correlation. *Sentiment* refers to the percentage of *negative* terms in the text, as decided by *Rocksteady* during sentiment analysis. Similarly, *Airline* refers to the percentage of terms in the text that are included in the domain-specific airline affect dictionary. *Tweets* refer to the volume of tweets in each period and *Volume* refers to the volume of trades transacted in the period.

Table 4.5: Correlation between time series variables on a daily basis (n=102)

⌢	Returns	Sentiment	Airline	Tweets	Volume
Returns	1.0000				
Sentiment	-0.0990	1.0000			
Airline	-0.0979	0.7565	1.0000		
Tweets	-0.0028	0.1988	-0.0990	1.0000	
Volume	-0.0990	-0.0009	-0.0990	-0.0990	1.0000

As expected, at a daily level, a negative but weak correlation exists between Returns and Sentiment, which suggests that an increase in negative sentiment correlates to a decrease in returns, or a decrease in returns correlates to an increase in negative sentiment. A similar negative relationship is observed between returns and *Airline* terms, the terms identified in Section 3.3.1. This might suggest that as returns decrease the population talks more about domain specific terms, the carrier terms. The correlation between *Volume* and *Returns* is equally as strong as negative sentiment and airline terms, at -0.0990. This suggests that the number of shares traded is a proxy for uncertainty in a market, and thus a proxy for investor sentiment. Interestingly, a much weaker correlation exists between tweet volume and returns, while still negative, it is much less significant to the point where it is mostly uncorrelated. A strong correlation of 0.7565 exists between negative sentiment and airline-specific terms, although the airline terms are neutral in terms of sentiment it is interesting to see that domain-specific carrier terms are used in-line with negative sentiment. Rather unexpectedly, a weak negative correlation of -0.0990 exists between tweet volume and trade volume. Literature suggests that this should be a positive correlation regardless of strength. Such that as media-attention increases, the volume of trades and uncertainty increases.

Table 4.6: Correlation between time series variables on a 5 minute basis (n=8184)

⌢	Returns	Sentiment	Airline	Tweets	Volume
Returns	1.0000				
Sentiment	-0.0152	1.0000			
Airline	-0.0083	0.0536	1.0000		
Tweets	-0.0093	0.1293	0.1249	1.0000	
Volume	-0.0837	-0.0047	0.0276	0.0116	1.0000

The correlation coefficients calculated at a 5-minute frequency are very similar in characteristics as those discussed above for daily frequency, with the exception of tweet volume and trade volume. In line with the literature, a weak positive correlation exists between the two, such that, as the number of trades in a 5-minute period increases so do the number of tweets published in the period. A much weaker correlation of 0.0536 is seen between negative sentiment and airline terms compared to the daily correlation of 0.7565.

Similarly, the correlation between returns and negative sentiment is lower at this frequency, but still negative, as expected. As before, none of the correlations are notably strong, but their polarities do provide some insight.

Table 4.7: Correlation between time series variables on a minute basis (n=30006)

⌢	Returns	Sentiment	Airline	Tweets	Volume
Returns	1.0000				
Sentiment	0.0015	1.0000			
Airline	-0.0091	0.1626	1.0000		
Tweets	0.0004	0.2136	0.2200	1.0000	
Volume	-0.0326	0.0091	0.0086	0.0086	1.0000

The correlation coefficients calculated at a 1-minute frequency present a very different picture than those at daily and 5-minute frequency. As expected, a weak negative correlation between trade volume and returns exists, but surprisingly, the correlation between returns and negative sentiment is positive. However, this correlation is weaker than previous correlations, and thus it is almost completely uncorrelated. Almost no correlation exists between tweet volume and returns, with a coefficient of 0.004. Similarly, the correlation coefficient between tweet volume and trade volume is 0.0086, exhibiting an incredibly weak, but positive correlation. Interestingly, a weak to moderate positive correlation exists between tweet volume and negative sentiment at the 1-minute frequency, with a coefficient of 0.2136. This value is stronger than any comparable coefficient at other frequencies.

As mentioned, none of the correlations between returns and any of the other variables are particularly strong. This behaviour is, of course, expected, as the efficient market hypothesis suggests that that stock market prices evolve according to a random walk, such that price changes are random and thus cannot be predicted. Therefore, it is difficult to expect any time series variable to be highly correlated with a time-series that is best described as random. For this reason, it is the polarity of the correlations that describe the relationship between the variables, if any.

4.4 Volatility

In finance, volatility, δ, is the degree of variation of a trading price series over time-based on the high/low and open/close prices. Using the 5-minute frequency data the intraday volatility for each day is calculated to proxy uncertainty in the markets. Volatility is then plotted against 5-minute tweet volume (Figure 4.4) and the daily tweet volume (Figure 4.5). The aim of these plots is to determine whether a spike in volatility corresponds to a spike in the volume of tweets. If a spike in volume proceeds a spike in volatility, it suggests that new information on Twitter is a cause of the uncertainty in the market. Conversely, if a spike in volatility proceeds a spike in volume, it suggests that the uncertainty in the market causes media-attention and discussion on Twitter.

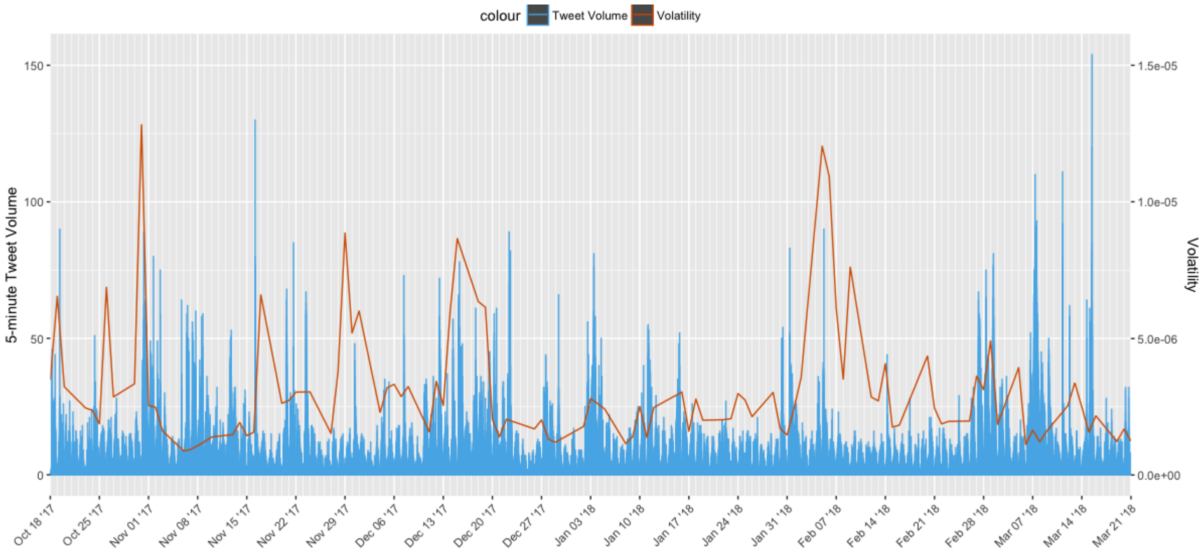

Figure 4.4: Intraday Volatility (orange, n=106) vs 5-minute tweet volume (blue, n=8184) between October 17th 2017 and April 2017

The relationship between intraday volatility and 5-minute tweet volume (Figure 4.4) provides interesting insights into social media's response to market volatility and how social media attention influences intraday volatility. There are many examples, from the graph, which show that spikes in volatility were proceeded and succeeded with an increase in tweet volume. In particular, this can be seen between December 13th and December 20th. During this time Ryanair's pilots were threatening strike action[44]. This period consists of a week of consistent high intraday volatility, in combination with clustering in large tweet volume. Similarly, on February 5th, 2018, there was a significant spike in volatility and a spike in tweet volume. Unfortunately there are also many counter examples where there is a spike in tweet volume and a less dramatic spike in volatility such as between January 3rd and 6th.

Figure 4.5, below, presents the same plot of volatility but against the sum of all tweets throughout the trading day. Similar patterns exist as described above but it is equally as unpredictable. A weak positive correlation of 0.0343 exists between the two variables, indicating that a general increase in tweet volume does correlate to an increase in volatility but as before, the plot exhibits plenty of examples where a spike in tweet volume is not reciprocated with a significant increase in volatility.

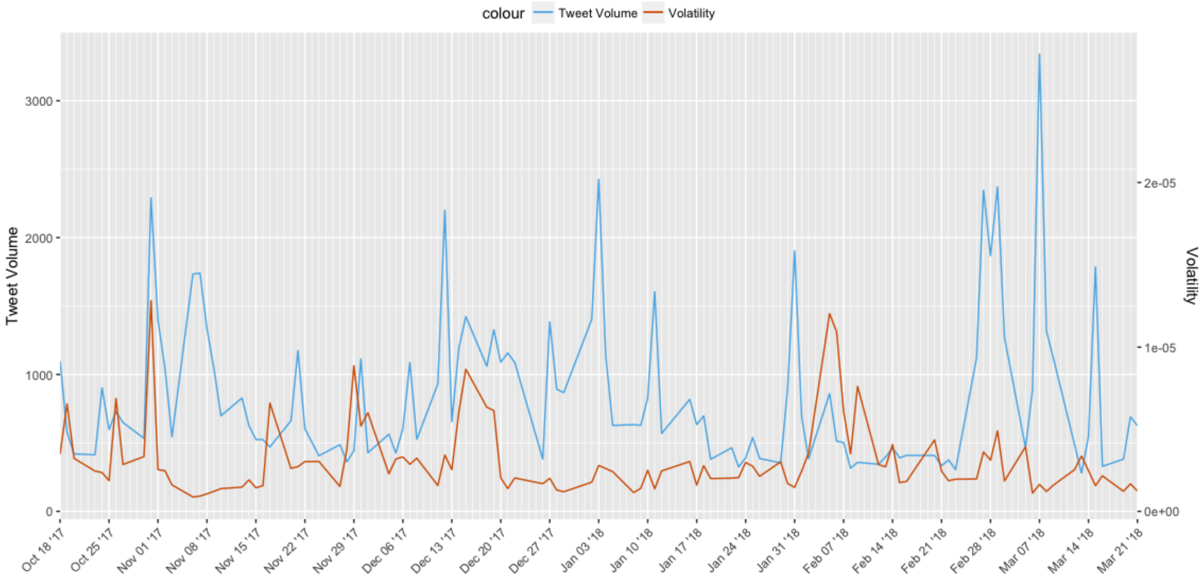

Figure 4.5: Intraday Volatility (orange) vs daily tweet volume (blue) between October 17th 2017 and April 2017 (n=106)

4.5 Impact of Sentiment on Returns

The relationships between the negative sentiment variable generated from the corpus of tweets, the volume of tweets and Ryanair returns are examined using a vector autoregression (VAR) model as described in Section 3.4. Models M1 to M4 are analysed in R for daily, 5-minute and 1-minute frequencies. The main output of the analysis is the coefficients of the sentiment variable, s_t, and the tweet volume variable, v_t. The coefficients describe any potential dependence that exists between the returns time-series and the sentiment proxy. The coefficient values and their corresponding statistical significance are tabulated in a panel analysis in Tables 4.8, 4.9 and 4.10 for daily, 5-minute and 1-minute frequency respectively. The significance for each coefficient is given at 0% (***), 1% (**), 5% (*) and 10% (.) levels and Newey-West adjusted standard errors are used to account for heteroskedasticity and autocorrelation of residuals. The adjusted-r-squared ($\bar{R}^2$) value is also reported for each model estimation to show the percentage of variance explained by the model.

The results of analysis of the daily data are presented in Table 4.8. The model is estimated incrementally as described in Section 3.4 with each independent variable being added and the model re-estimated to observe possible confounding effects with sentiment. The results show that 1-day lag of returns and 3-day lag of returns are statistically significant at the 10% level or higher, with an 1800 basis point impact on returns on average. Negative sentiment, s_t, is insignificant across all models but predominantly negative, as expected. Negative coefficients suggest that as returns decrease, the variable increases, i.e. negative sentiment increases as returns decrease. 3-day lag of tweet volume, v_{t-3}, is significant at the 5% level in Model M3, where sentiment is excluded. This suggests that tweet volume is indeed a proxy for investor sentiment. Similarly, in Model M4, 3-day and 4-day lag of tweet volume are significant at a 10% level. The coefficients for v_{t-3} are negative, which suggests that an increase in tweet volume contributes to a decrease in returns, but the 4-day lag coefficient is positive. This alternating-sign pattern suggests that mean-reversion is occurring. Similarly, the alternate coefficients for the lag of returns alternate signs, for the most part. The 1-day lag of trade volume, $Volume_{t-1}$, is insignificant and the estimates of its coefficients are incredibly small which suggests an almost negligible contribution to the model. The adjusted-r-squared value, $\bar{R}^2$, actually decreases with the inclusion of sentiment variables, explaining less variance in returns. Negative $\bar{R}^2$ values suggest that the model contains terms that negatively contribute to the explanatory power of the model. As the $\bar{R}^2$ value is positive in Model M3, it suggests that the inclusion of negative sentiment, s_t, is detrimental to the effectiveness of the model.

Table 4.8: Coefficient estimates after vector autoregression on daily data between October 17th 2017 and April 2017 (n=102). Coefficients are presented in basis points, one basis point equals a percentage point of 0.01%. The significance for each coefficient is given at 0% (***), 1% (**), 5% (*) and 10% (.) levels.

	M1		M2		M3		M4	
r_{t-1}	1862.50	.	1857.50	.	1786.60	.	1831.30	.
r_{t-2}	-1313.30		-1216.00		-1243.00		-1169.90	
r_{t-3}	1720.80	.	1687.50	.	1820.80	*	1869.10	*
r_{t-4}	727.79		546.82		697.18		568.89	
r_{t-5}	-492.50		-649.79		-541.16		-713.19	
s_{t-1}			2.4757				2.4884	
s_{t-2}			-4.9047				-3.1136	
s_{t-3}			-2.2686				0.5048	
s_{t-4}			-8.5669				-11.8620	
s_{t-5}			10.9290				9.6141	
v_{t-1}					-0.0040		-0.0046	
v_{t-2}					0.0010		0.0016	
v_{t-3}					-0.0074	*	-0.0067	.
v_{t-4}					0.0047		0.0062	.
v_{t-5}					0.0065		0.0052	
$Volume_{t-1}$	8.93e-06		1.35e-05		4.91e-06		1.09e-05	
Intercept	2.1471		45.4980		2.1085		1.5650	
$\bar{R}^2$	1.79%		-1.39%		0.36%		-3.76%	

The VAR coefficient estimates for 5-minute data are presented in Table 4.9. As before, the model is estimated incrementally, with each independent variable being added and the model re-estimated to observe possible confounding effects with sentiment. Initial impressions suggest that neither sentiment, s_t, nor tweet volume, v_t, are significant at any level. r_{t-4} (20-minute lag of returns) is significant at a 10% level across each model, with a large negative impact of 195 basis points on average. Unlike the analysis at a daily level, the sentiment coefficients are mostly positive, which is unusual. Trade volume of 1-lag is insignificant and negligibly contributes, but its coefficient is negative as expected. The adjusted-r-squared value, $\bar{R}^2$, decreases again as sentiment and tweet volume are included in the model, which suggests a reduced *goodness-of-fit* with the inclusion of the exogenous variables.

Table 4.9: Coefficient estimates after vector autoregression on 5-minute frequency data between October 17th 2017 and April 2017 (n=8184). Coefficients are presented in basis points, one basis point equals a percentage point of 0.01%. The significance for each coefficient is given at 0% (***), 1% (**), 5% (*) and 10% (.) levels.

	M1	M2	M3	M4
r_{t-1}	-86.36	-87.60	-86.98	-88.13
r_{t-2}	-234.39	-232.44	-234.23	-232.39
r_{t-3}	-97.38	-93.56	-96.08	-92.46
r_{t-4}	-196.62 .	-196.35 .	-194.80 .	-194.48 .
r_{t-5}	119.54	119.85	119.01	119.52
s_{t-1}		-0.0554		-0.0503
s_{t-2}		0.1272		0.1275
s_{t-3}		0.0883		0.0843
s_{t-4}		0.0689		0.0712
s_{t-5}		-0.1875		-0.1842
v_{t-1}			-0.0191	-0.0192
v_{t-2}			0.0006	-0.0003
v_{t-3}			0.0341	0.0322
v_{t-4}			-0.0010	-0.0024
v_{t-5}			-0.0191	-0.0160
$Volume_{t-1}$	-5.47e-05	-5.55e-05	-5.46e-05	-5.54e-05
Intercept	0.4510	0.3509	0.5013	0.3991
$\bar{R}^2$	0.08%	0.06%	0.03%	0.01%

Finally, Table 4.10 presents the VAR coefficient estimates for the 1-minute analysis. Interestingly, previous minute returns, r_{t-1}, are highly significant at the 0% level, with an impact of -822 basis points on average across each model. This suggests that the returns are very strongly related to the returns of the previous minute, this relationship is not evident at a 5-minute frequency. The coefficient is also negative which, as before, suggests mean-reversion occurs, such that positive returns in one minute are very highly likely to be followed by negative returns in the next minute and vice-versa. Tweet volume at this frequency is completely insignificant, which is most likely due to the low volume of tweets that occurs in each minute. As discussed above, the average time between tweets during the day is 26.16 seconds, which averages out to approximately 2 tweets per minute. The variation at this frequency-level would have no significant effect across such a broad time period. However, 3-minute lag of negative sentiment, s_{t-3} is significant at a 10% level with an impact of -0.0484 basis points. As before, this negative relationship suggests an increase in negative sentiment contributes to a decrease in returns. The lag of trade volume, $Volume_{t-1}$, is once again insignificant and very small, but negative as expected. The adjusted-r-squared value, $\bar{R}^2$, remains mostly consistent throughout each model at this frequency. This means the inclusion of sentiment and tweet volume does not reduce the *goodness-of-fit* as seen at other frequencies, however, since it does not increase, it suggests that the inclusion of those variables do not contribute significantly to the explanatory power of the models.

Table 4.10: Coefficient estimates after vector autoregression on 1-minute frequency data between October 17th 2017 and April 2017 (n=30006). Coefficients are presented in basis points, one basis point equals a percentage point of 0.01%. The significance for each coefficient is given at 0% (***), 1% (**), 5% (*) and 10% (.) levels.

	M1		M2		M3		M4	
r_{t-1}	-823.32	***	-822.52	***	-822.83	***	-821.73	***
r_{t-2}	-64.00		-63.88		-64.61		-64.57	
r_{t-3}	-141.52		-141.35		-142.27		-142.19	
r_{t-4}	-117.91		-118.10		-117.59		-117.59	
r_{t-5}	75.16		75.07		74.32		74.41	
s_{t-1}			-0.0028				-0.0008	
s_{t-2}			0.0192				0.0247	
s_{t-3}			-0.0484	.			-0.0505	.
s_{t-4}			0.0067				0.0107	
s_{t-5}			0.0112				0.0075	
v_{t-1}					-0.0159		-0.0152	
v_{t-2}					-0.0462		-0.0531	
v_{t-3}					0.0236		0.0376	
v_{t-4}					-0.0371		-0.0411	
v_{t-5}					0.0540		0.0530	
$Volume_{t-1}$	-4.49e-05		-4.48e-05		-4.44e-05		-4.44e-05	
Intercept	0.1118		0.1426		0.1599		0.1720	
$\bar{R}^2$	0.70%		0.70%		0.69%		0.70%	

The results of the VAR models described above vary dramatically for different frequencies. It is clear that the models fail to fit the high-frequency data, as the daily-frequency models show the greatest promise in terms of significant variables. As a whole, sentiment is rarely significant, except at a 1-minute frequency. It might be better to consider the volume of tweets as a proxy for investor sentiment rather than extracting sentiment from tweets. Of course, it might also be because *Rocksteady* is designed to extract sentiment from formal-media as opposed to social media, and thus the short, casual nature of tweets poses a challenge to the system.

4.6 Week-by-Week Analysis

As mentioned, the high-frequency models failed to adequately fit the time series over the 6-month period. Sentiment was mostly insignificant and the adjusted R-squared values indicated that sentiment mostly negatively contributed to the explanatory power of the models. To further investigate the effect of sentiment using high-frequency data, Model M2 is tested on a week-by-week basis, over 21 weeks between October 2017 and April 2018 to identify under which conditions the models work best. The magnitudes, in basis points, of the significant coefficients, and their respective significance level are presented below. Insignificant coefficients were omitted for the purpose of clarity. Figures 4.6 presents the values for α_t, the coefficient for returns, r_t. Figure 4.7 presents the values for β_t, the coefficients for negative sentiment, s_t, and Figure 4.8 presents the α_0, intercept value, ρ, the coefficient for lagged Volume of trades, V_{t-1} and N, the number of observations per week.

The results from Figure 4.6 shows that over the 21 weeks, the first-order lag of returns, r_{t-1}, are most significant. In total it is significant 8 times. All its coefficients but one are negative and range between -1000 basis points and -2000 basis points. This suggests that in certain weeks, perhaps in periods of consistent volatility, the returns in the previous 5 minutes were strongly negatively correlated with the current time period. In total, across all orders of lag for returns, only two coefficients were positive which implies that previous returns are most significant with a negative coefficient, possibly due to the nature of mean-reversion. Fifth-order lags were the least significant, with only one significant week, at a 10% level. This analysis presents an interesting comparison compared to the analysis over the 6-month period presented in Table 4.9. Over 6-months, only forth-order lags in returns were significant, whereas this weekly analysis shows that in different weeks the model fits differently. In particular, fourth-order lags are less significant than first-order or third-order lags on average.

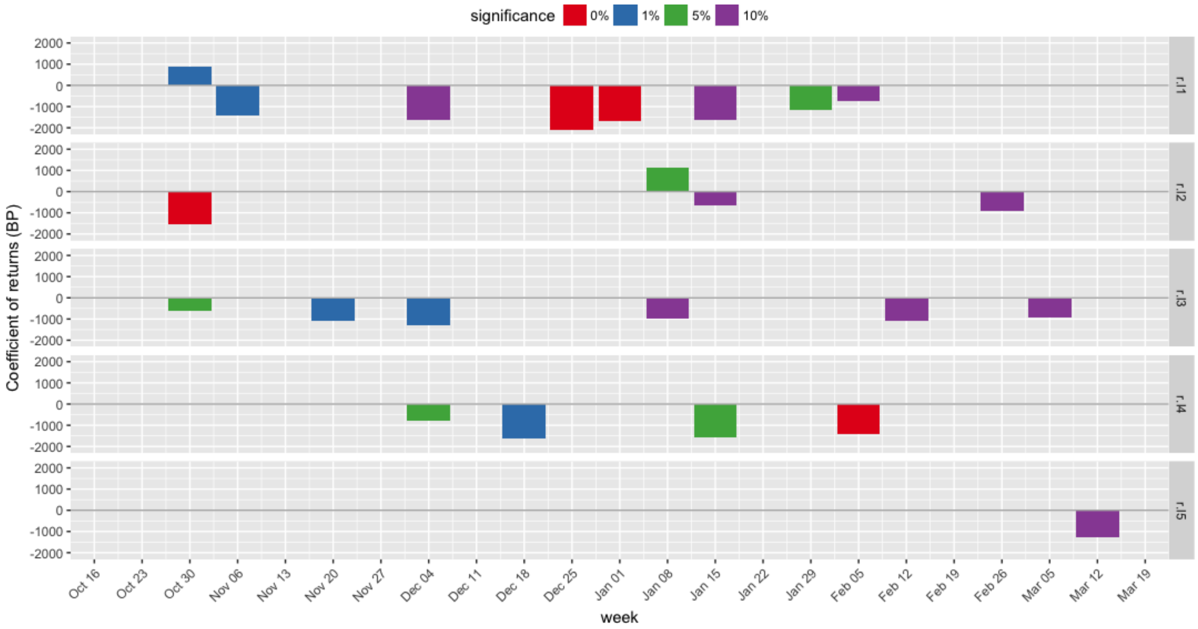

Figure 4.6: Estimates of significant coefficients for lagged returns, α_t, after analysis of Model M2 on a rolling weekly basis

Figure 4.7 presents the estimates of significant coefficients for negative sentiment, of which, none were significant when the model was fit over a 6-month period. It is immediately obvious that sentiment does indeed have a significant effect on returns under different circumstances as each order of lag of returns shows some level of significance across the 21 weeks. In particular, the 4 consecutive weeks from December 11th to January 1st present significant coefficients for both first-order lags and third-order lags at the 5% and 10% significant level. During this period, Ryanair's pilots threatened strike action and Ryanair received a lot of media attention, both negative and positive. Examining the plot of volatility in Figure 4.4 shows that this period had consistent high volatility during the week of the strike announcement and then consistent low volatility in the following weeks and the weeks leading up to Christmas. It is interesting that negative sentiment seemed to significantly and consistently contribute to the returns during this period. The coefficients for r_{t-1} are mostly positive, and the weeks in which they are negative, their magnitude is on average lower than the weeks in which it is positive. A positive coefficient implies that negative sentiment found in tweets in the previous five minutes contributes to an increase in returns, which is not expected. The negative coefficients in the week starting December 18th would again be explained by the negative sentiment surrounding the threat of pilot strikes in that week. Interestingly, second-order lags are the least significant with only one significant week out of the 21 weeks. This might be due to the fact that some sentiment is quickly absorbed into the price, as seen in the first-order lags, but very little is explained by the second order-lags. Similarly, fifth-order lags of negative sentiment are only significant in two weeks, most likely due to sentiment being quickly absorbed into the prices.

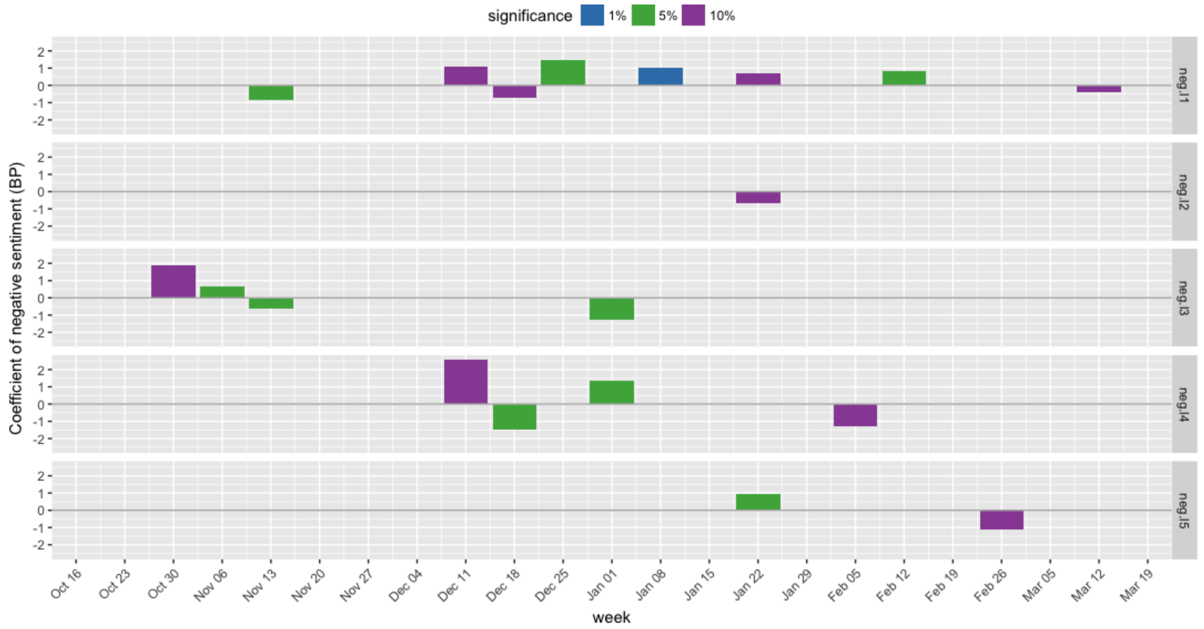

Figure 4.7: Estimates of significant coefficients for lagged returns, β_t, after analysis of Model M2 on a rolling weekly basis

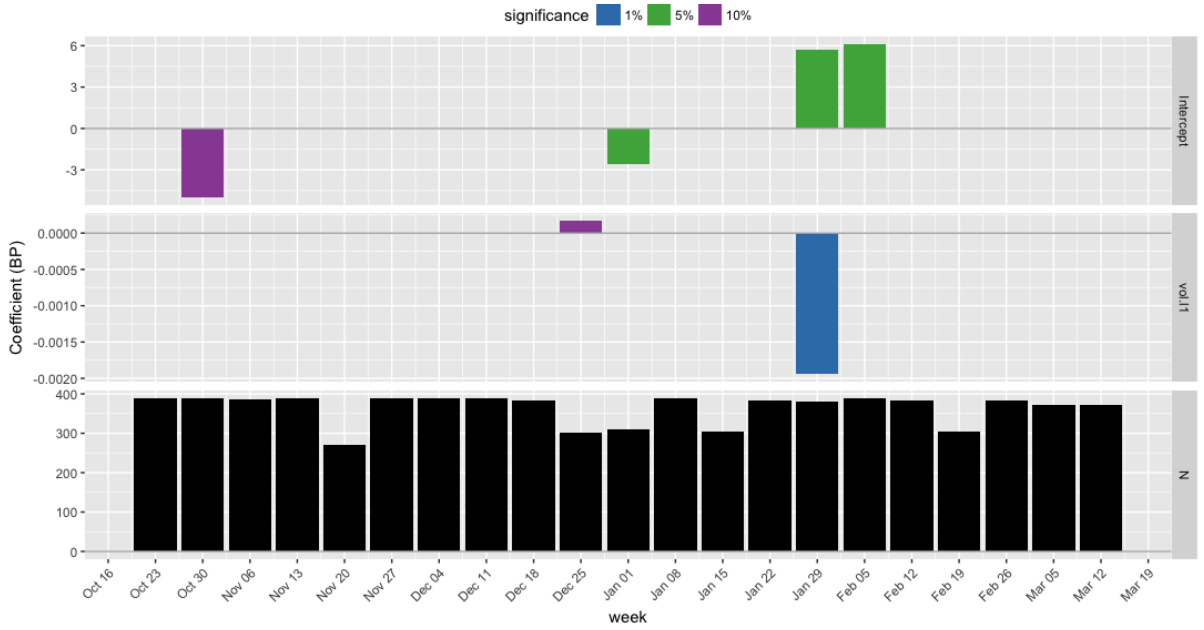

Figure 4.8: Estimates of significant coefficients after analysis of Model M2 on a rolling weekly basis, and N, the number of observations per period

Estimates for the coefficients for a first-order lag trading volume and the model's intercept value are presented in Figure 4.8. Volume was insignificant when the model was fit across 6-months of data, and even week-by-week it is only significant twice. Literature suggests that its coefficient should be negative, which it is in the week starting January 29th, but it is positive, to a much weaker extent for the week starting December 25th, although this could be due to seasonal effects.

4.7 Comparison of Formal Media and Social Media

For purpose of comparison between the impact of sentiment found in tweets versus formal media on financial returns, over 41,000 news articles, written in English, where the term 'Ryanair' appeared in the headline, were collected over a 10-year period ranging from April 2008 to April 2018.

Correlation

Table 4.11 presents the correlation coefficients between the different variables in the formal media time series. As before, *Sentiment* refers to the percentage of negative terms in the text, as decided by Rocksteady during sentiment analysis. Similarly, *Airline* refers to the percentage of terms in the text that are included in the domain-specific affect dictionary. *Articles* refer to the volume of articles published in each period and Volume refers to the volume of trades transacted in the period.

Table 4.11: Correlation between news article time series variables on a daily basis (n=2507)

	Returns	Sentiment	Airline	Articles	Volume
Returns	1.0000				
Sentiment	-0.0403	1.0000			
Airline	-0.0434	0.2531	1.0000		
Articles	0.0261	0.1433	0.0492	1.0000	
Volume	-0.0125	-0.0040	-0.0222	0.0866	1.0000

As expected the correlation between negative sentiment, *Sentiment*, and returns is weak and negative. Similarly, the relationships between trade volume and returns, and airline terms and returns are negative. This confirms that even over a much longer period (10 years versus 6 months), the trade volume is negatively correlated with returns. This implies that as the volume of trades increases, the returns decrease, albeit a weak correlation. Rather surprisingly, the correlation between article volume, *Articles*, and returns is weak and positive. Existing work in this field has found that media-attention has a negative correlation with returns. However, as with trade volume, the correlation is quite weak. A weak positive correlation exists between article volume and trade volume, which would be expected. As media-attention increases so do the volume of trades that occur. Similarly, a weak to moderate correlation exists between negative sentiment and article volume, which implies that as media-attention increases, the sentiment of that media is typically negative. This conforms to the concept of *no news is good news*.

In comparison to the correlation of the values found in social media data, presented in

Section 4.3, many of the relationships are equally as strong/weak and are the same sign. Negative sentiment found in tweets and negative sentiment found in formal media both have a negative relationship with returns, which is expected. The exception is that the volume of tweets is negatively correlated with returns, whereas article volume is positively correlated, however, both correlations are quite weak. Similarly, article volume is positively correlated with trade volume, but tweet volume is negatively correlated with trade volume. As mentioned above, intuitively, a positive correlation is expected in this instance.

Vector Autoregression

To compare the impact of sentiment in formal media versus tweets, the coefficients of the VAR models, M1 to M4 are computed again using the formal media time series. The output of the analysis is presented in Table 4.12, showing the coefficients of the sentiment variable, s_t, and the article volume variable, v_t. The significance for each coefficient is given at 0% (***), 1% (**), 5% (*) and 10% (.) levels and Newey-West adjusted standard errors are used to account for heteroskedasticity and autocorrelation of residuals. The adjusted-R^2 value is also reported ($\bar{R}^2$) for each model estimation to show the percentage of variance explained by the model.

The results show that first-order lag of returns, r_{t-1}, is significant at the 1% level, with an average impact of 925 basis points. These results are similar to the results presented in Table 4.8, the daily analysis of social media data, which is fit over a much shorter time period of only 6 months. In this, coefficients for r_{t-1} are significant at a 10% level, however, its basis points are twice as large. Similar to the social media models, sentiment is insignificant across all models, for all lags. Once again, media-volume, in this case, article volume is significant, which suggests that it acts as a proxy for investor sentiment. The first-order lag of article volume is significant in both models at the 5% level, with a positive impact of 0.376 basis points. What is interesting is that the sign of the coefficients for article volume are the opposite to the estimates for the coefficients for tweet volume. Although both models show alternating signs between lags, which suggests mean-reversion occurs. The fact that the two models present opposite signs is consistent with the correlation found between article volume and returns and tweet volume and returns, discussed above. The two correlation coefficients both had opposite signs, even though they would have been expected to both be negative. An analysis of the adjusted-r-squared value, $\bar{R}^2$, shows that the inclusion of sentiment decreases the explanatory power of the model, similar to the results found in the social media models. However, including article volume slightly increases the $\bar{R}^2$ value, when it is independent of sentiment.

Table 4.12: Coefficient estimates after vector autoregression on daily formal media data between April 2008 and April 2017 (n=2507). Coefficients are presented in basis points, one basis point equals a percentage point of 0.01%. The significance for each coefficient is given at 0% (***), 1% (**), 5% (*) and 10% (.) levels.

	M1		M2		M3		M4	
r_{t-1}	-921.58	**	-927.87	**	-931.90	**	-941.69	**
r_{t-2}	-130.27		-136.45		-114.16		-120.84	
r_{t-3}	-356.54		-348.55		-342.64		-333.30	
r_{t-4}	-169.08		-157.54		-153.91		-143.31	
r_{t-5}	155.15		166.26		159.22		166.96	
s_{t-1}			0.0601				-0.7078	
s_{t-2}			-0.9376				-1.1281	
s_{t-3}			3.3174				3.4102	
s_{t-4}			2.7307				3.0439	
s_{t-5}			0.2971				0.2787	
v_{t-1}					0.3763	*	0.3768	*
v_{t-2}					-0.0465		-0.0274	
v_{t-3}					0.0800		0.0646	
v_{t-4}					-0.0878		-0.1266	
v_{t-5}					-0.0615		-0.0732	
$Volume_{t-1}$	-7.78e-08		1.61e-07		-1.15e-06		-9.63e-07	
Intercept	2.8832		-7.5946		0.5338		-8.3353	
$\bar{R}^2$	0.77%		0.70%		0.79%		0.74%	

5 Discussion

The results in Chapter 4 are presented and discussed individually. This chapter discusses the strengths and limitations of the chosen method and further discusses the results previously presented drawing conclusions from the entire collection of results.

5.1 System Strengths and Limitations

The results presented in Chapter 4 highlight some of the strengths and weaknesses of the methods employed. The data retrieval component of the system leverages the Twitter APIs to collect as many tweets as possible. A strength of this is that the *Streaming* API fetches tweets in real time, which facilitates high-frequency sentiment analysis. Unfortunately, Rocksteady, the sentiment analysis component, does not facilitate programmatic analysis of dynamic text. It requires a corpus to be exported in a specific format and imported into its system where it performs sentiment analysis. While this is sufficient for research purposes, it would not be suitable for application in high-frequency trading. The rest of the statistical analysis is performed in *R*, a programming language for statistical computing, after importing the results of sentiment analysis and the financial time series. *R* provides very efficient methods for aggregating both time series by the same time period and performing statistical analysis such as correlation, autocorrelation and vector autoregression. *R* is very flexible and allows for the analysis of complex linear models such as those discussed above.

One limitation of the approach taken is that the sentiment time series is aligned with the financial time series by date and time. At high-frequency this becomes an issue; a large portion of the sentiment ends up being ignored because the NASDAQ does not open until 2.30pm GMT, which discards several hours of potentially useful sentiment published in the morning. Similarly, the frequency of tweets decreases in the evenings in GMT, as little news is released after working hours, yet this makes up the bulk of the daily high-frequency sentiment that is analysed. A better approach could be to incorporate the morning's sentiment into the first trading interval of each day, as it is not uncommon for stock prices to decrease dramatically on opening due to bad news announced outside market hours.

5.2 Results

The value of social media data, as discussed in Chapter 2 is that it is high-frequency compared to formal media. The results show that from morning to close of markets the average tweet period (about Ryanair) is 26.16 seconds. This level of data-flow is not available with formal media. This allows sentiment analysis to be performed on a 5-minute frequency and 1-minute frequency, to allow for intra-day trading as opposed to simply analysing the results at a daily level. However, despite allowing for high-frequency analysis, the results of the VAR models examined show that daily analysis is still much stronger than high-frequency analysis. The adjusted-r-squared values, $\bar{R}^2$, of daily analysis are significantly higher than that of 1-minute or 5-minute analysis. Furthermore, the models which included sentiment and tweet volume at a daily frequency resulted in a negative $\bar{R}^2$ value, suggesting that the inclusion of those independent variables reduced the explanatory power of the models. In contrast, the $\bar{R}^2$ values at a high-frequency did decrease as sentiment and tweet volume were added, but were not negative. In all cases except 1-minute frequency, sentiment was insignificant. At 1-minute frequency, 3-minute lag of negative sentiment was significant at a 10% level, whereas all other sentiment variables were insignificant. The lack of significance might be due to the ineffectiveness of *Rocksteady* in dealing with short, casual pieces of text such as tweets. Rocksteady was designed to extract affect words from formal media such as news articles and newswires, which are typically much longer than 140/280 characters. The short, casual nature of tweets, with common use of slang, abbreviations and (intentional) misspellings, makes it difficult to extract sentiment using a lexicon based approach without doing significant work in creating slang dictionaries. Thus, simply observing tweet volume might be a more effective proxy for investor sentiment.

Another explanation as to why the daily models had higher explanatory power than the high-frequency models could be the change in volatility across the six month period. In particular, the month of December saw large spikes in volatility with respect to the volatility across the 6-month period as Ryanair's pilots called for strike action and caused the share price to drop from \$115 to \$103 over the course of 3 days (Figure 4.4). This makes it difficult to fit a model that deals with periods of both low volatility and high volatility. The week-by-week analysis suggests that sentiment, particularly the first-order lag of sentiment, had a significant effect on returns during certain weeks. Analysis of volatility during these weeks shows that the effect of sentiment was most significant during weeks of consistent volatility. The month of December presented examples of consistent high volatility weeks and consist low volatility weeks, and in both situations sentiment was significant at the 10% level or lower. Furthermore, on a daily basis, returns are calculated off end of day closing prices, and sentiment is accumulated and aggregated throughout the day. This means that the phenomenon of changes in sentiment occurring before, simultaneously with or after changes in returns is masked as sentiment figures for the day are aggregated causing variations on a high-frequency scale to be hidden. Thus it is easier to correlate sentiment with returns, whereas on a high-frequency level, it is clear that in some situations sentiment proceeds the change in returns by a few minutes or a few seconds, or sometimes the change in sentiment succeeds the change in returns, sometimes quickly and sometimes slowly throughout a whole day. This makes it very difficult to model the returns based on changes in sentiment at such a volatile frequency. So while it is easier to model at a daily level it doesn't allow for high-frequency analysis, and consequently, trading.

A stylised fact, as described by Taylor, is a general property that is expected to be present in any set of returns[43]. The analysis of the returns at different frequencies confirms that the same stylised facts do not apply to high-frequency data. Taylor states that there is almost no correlation between returns on different days. This holds true for the daily-analysis performed. Table 4.1 presents the summary statistics of RYAAY since its initial public offering (IPO) in 1997. An incredibly weak negative correlation of -0.0104 exists between returns and previous day returns, which is consistent with the stylised fact. The VAR analysis at daily-frequency does contain significant autocorrelation of returns, but only at a 10% significance level. In comparison, the 1-minute analysis shows that previous minute returns are significant at the 0% level, which suggests a very strong correlation even after Newey-West adjusted standard errors are used to account for heteroskedasticity and autocorrelation of residuals. This might be explained by one of Taylor's intraday stylised fact; *"Intraday returns from traded assets are almost uncorrelated with any important dependence usually restricted to a negative correlation between consecutive returns"*. The first-order autocorrelation of returns at 1-minute frequency has a value of -0.083, which is definitely greater than that found at a daily level, but otherwise still weak, as explained by the stylised fact.

Rocksteady, as mentioned, uses a lexicon based approach to extracting sentiment from a corpus of text. It is no surprise that it proved difficult to extract sentiment from the corpus of tweets, due to their short nature. In any given tweet the number of tokens that are identified as negative can be very low, and often zero. Very few tweets are identified as highly negative, and very often they are simply passengers complaining about their own poor experience which would have little or no effect on the companies stock price. It is only when a large cohort of customers complain about their own experiences in unison, such as during the potential pilot strikes in December 2017, a significant relationship is seen between negative sentiment in tweets and returns. This suggests that (negative) sentiment models the uncertainty in the analysis of an asset and works particularly well during times of large uncertainty. However, there might be better approaches to extracting sentiment from tweets. Similar to how a custom POS tagger for Twitter was developed by Gimpel et al. a custom sentiment analysis system could be extended from Rocksteady with a focus on detecting sentiment in casual texts, by detecting common abbreviations, slang, sentiment in emoji and more.

To perform high-frequency analysis, a high-frequency source of sentiment is required. At a firm-specific level, this can be challenging. In this case, the period between tweets during the hours of analysis was on average 26.16 seconds, with the time between the top 25% of sequential tweets being 5 seconds or less. This is certainly acceptable for high-frequency sentiment analysis assuming the tweets do indeed contain sentiment about a brand, product or firm. If this is not true, as is often the case[8] then very little sentiment will be found, particularly when it is already difficult to extract sentiment from the medium as discussed above. Thus, it is incredibly important that the firm in question has sufficient social media attention for high-frequency sentiment analysis to be considered appropriate, as smaller firms or perhaps Business-to-Business (B2B) firms with fewer customers tweeting about them might prove even more challenging to model due to data sparsity.

5.3 Conclusion

This thesis aimed to investigate the effect of sentiment in high-frequency financial markets. Chapter 1 provided an introduction to the thesis and the theory which motivated the investigation discussing the growth in high-frequency trading and the role played by sentiment analysis. It also explained why Twitter presents itself as a good source of high-frequency sentiment. Chapter 2 presented the existing work in the field of sentiment analysis and Twitter text analysis techniques which shaped the methods and objectives of this research. Chapter 3 described the methods employed in reaching the research objectives, by outlining and discussing the four distinct phases involved. This included the implementation and development of a system for generating a sentiment time series from social media and traditional news articles and the incorporation of those time series into an autoregressive statistical model with financial data. The statistical analysis presented four vector autoregressive models which independently assess the impact of exogenous variables such as negative sentiment and media volume on financial returns. Chapter 4 presented a case study of Ryanair, an Irish low-cost airline, which was used to evaluate the system and the effect of sentiment on its returns. Negative sentiment was extracted from approximately 360,000 tweets about Ryanair published between October 16th, 2017 and April 2018 and included in different models of financial returns at different frequencies, paying particular attention to high-frequency analysis. The results of the analysis are included in this chapter where they are discussed individually. Finally, this is followed by a more general discussion of the results where the context and implications of all of the results, at different frequencies, are considered.

This research successfully leveraged computation linguistics to automatically construct a domain-specific affect dictionary in the domain of consumer air travel. This was done by identifying highly frequent and highly *weird* common nouns. These were then used during sentiment analysis to override any negative words that are neutral in the domain context. The automation of this process fundamentally improves the efficiency of sentiment analysis as it allows for the analysis of any domain with an existing corpus.

The analysis found that sentiment found in social media, specifically Twitter, plays a small role in explaining returns of an asset. It found that the explanatory power of sentiment varies at different frequencies. In fact, its explanatory power is highest at a daily frequency due to the aggregation of sentiment throughout the day which masks the subtle changes used in high-frequency trading. However, it was found that high-frequency sentiment was significant during different periods across the 6-months analysed. A week-by-week analysis found that it was most significant during periods of consistent volatility, be it consistently high such as during a threat of strike action or consistently low such as the period leading up to Christmas.

A comparison of sentiment extracted from formal media versus sentiment extracted from social media found that it is easier to extract sentiment from formal media due to their formal language and larger volume of text. Rocksteady was designed to extract sentiment from such texts and thus provided better results than the analysis of tweets. However, negative sentiment was not significant in the vector autoregressive models examined but a first-order lag of article volume was, at the 5% level. This suggests that article volume is a proxy for investor sentiment.

As previously mentioned, high-frequency negative sentiment extracted from tweets was significant across many weeks at different levels, but the conditions in which it is significant are less than clear, besides the initial observations that it has the most significant impact during periods of consistent volatility. Future research will be carried out into the investigation of when and why high-frequency sentiment is significant, by further analysing its effect on shorter time periods and by varying the conditions. If sentiment is to be used as part of a high-frequency trading strategy it would need to be possible to determine when to take sentiment into account and how much influence it should have. Furthermore, this work focused mostly on a firm-specific analysis in the commercial airline industry. It would be interesting to determine whether the results are consistent for other firms in the commercial airline industry, and how much they might differ in another industry altogether.

In conclusion, the negative sentiment found in tweets is very appropriate for use in high-frequency trading strategies, as there is a very significant relationship between sentiment and returns of a financial asset but the circumstances under which it is significant need to be thoroughly examined before it can be fully exploited.

Bibliography

[1] Nuno Oliveira, Paulo Cortez, and Nelson Areal. On the predictability of stock market behavior using stocktwits sentiment and posting volume. In Luís Correia, Luís Paulo Reis, and José Cascalho, editors, *Progress in Artificial Intelligence*, pages 355–365, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-40669-0.

[2] J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2011. URL http://scholar.google.de/scholar.bib?q=info:7jxZLbM1mzwJ:scholar.google.com/&output=citation&hl=de&as_sdt=0&ct=citation&cd=54.

[3] Timm O. Sprenger, Andranik Tumasjan, Philipp G. Sandner, and Isabell M. Welpe. Tweets and trades: the information content of stock microblogs. *European Financial Management*, 20(5):926–957. doi: 10.1111/j.1468-036X.2013.12007.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-036X.2013.12007.x.

[4] Malcolm Baker and Jeffrey Wurgler. Investor sentiment in the stock market. Working Paper 13189, National Bureau of Economic Research, June 2007. URL http://www.nber.org/papers/w13189.

[5] Deb Dutta Das, Sharan Sharma, Shubham Natani, Neelu Khare, and Brijendra Singh. Sentimental analysis for airline twitter data. *IOP Conference Series: Materials Science and Engineering*, 263(4):042067, 2017. URL http://stacks.iop.org/1757-899X/263/i=4/a=042067.

[6] Andrew Haldane. Patience and finance. 01 2010.

[7] Alexander Hogenboom, Daniella Bal, Flavius Frasincar, Malissa Bal, Franciska de Jong, and Uzay Kaymak. Exploiting emoticons in sentiment analysis. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, SAC '13, pages 703–710, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-1656-9. doi: 10.1145/2480362.2480498. URL http://doi.acm.org/10.1145/2480362.2480498.

[8] Bernard J. Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury. Twitter power: Tweets as electronic word of mouth. *J. Am. Soc. Inf. Sci. Technol.*, 60(11):2169–2188, November 2009. ISSN 1532-2882. doi: 10.1002/asi.v60:11. URL http://dx.doi.org/10.1002/asi.v60:11.

[9] Gary D Eppen and Eugene Fama. Cash balance and simple dynamic portfolio problems with proportional costs. *International Economic Review*, 10(2):119–33, 1969. URL https://EconPapers.repec.org/RePEc:ier:iecrev:v:10:y:1969:i:2:p:119-33.

[10] Bo Qian and Khaled Rasheed. Stock market prediction with multiple classifiers. *Applied Intelligence*, 26(1):25–33, Feb 2007. ISSN 1573-7497. doi: 10.1007/s10489-006-0001-7. URL https://doi.org/10.1007/s10489-006-0001-7.

[11] Khurshid Ahmad, JingGuang Han, Elaine Hutson, Colm Kearney, and Sha Liu. Media-expressed negative tone and firm-level stock returns. *Journal of Corporate Finance*, 37:152 – 172, 2016. ISSN 0929-1199. doi: https://doi.org/10.1016/j.jcorpfin.2015.12.014. URL http://www.sciencedirect.com/science/article/pii/S0929119915001637.

[12] K. Adhmad. Notes on sentiment analysis: I. the power of affect and language. *Trinity College Dublin, Dublin, Ireland & Copenhagen Business School Copenhagen, Denamrk.*, Jan 2014.

[13] Werner Antweiler and Murray Z. Frank. Is all that talk just noise? the information content of internet stock message boards. *The Journal of Finance*, 59(3):1259–1294, 2004. ISSN 00221082, 15406261. URL http://www.jstor.org/stable/3694736.

[14] Francesco Corea. Can twitter proxy the investors' sentiment? the case for the technology sector. *Big Data Research*, 4:70 – 74, 2016. ISSN 2214-5796. doi: https://doi.org/10.1016/j.bdr.2016.05.001. URL http://www.sciencedirect.com/science/article/pii/S2214579615300174.

[15] Merve Alanyali, Helen Susannah Moat, and Tobias Preis. Quantifying the relationship between financial news and the stock market. *Scientific Reports*, 3:3578 EP –, 12 2013. URL http://dx.doi.org/10.1038/srep03578.

[16] Chong Oh and Olivia Sheng. Investigating predictive power of stock micro blog sentiment in forecasting future stock price directional movement. In Dennis F. Galletta and Ting-Peng Liang, editors, *ICIS*. Association for Information Systems, 2011. ISBN 978-0-615-55907-0. URL http://dblp.uni-trier.de/db/conf/icis/icis2011.html#OhS11.

[17] Johan Bollen, Huina Mao, and Alberto Pepe. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena, 2011. URL https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2826.

[18] Hong Kee Sul, Alan R. Dennis, and Lingyao (Ivy) Yuan. Trading on twitter: Using social media sentiment to predict stock returns. *Decision Sciences*, 48(3):454–488, 2017. ISSN 1540-5915. doi: 10.1111/deci.12229. URL http://dx.doi.org/10.1111/deci.12229.

[19] Jasmina Smailović, Miha Grčar, Nada Lavrač, and Martin Žnidaršič. Stream-based active learning for sentiment analysis in the financial domain. *Information Sciences*, 285:181 – 203, 2014. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2014.04.034. URL http://www.sciencedirect.com/science/article/pii/S0020025514004885. Processing and Mining Complex Data Streams.

[20] A. Rao, N. Spasojevic, Zhisheng Li, and T. Dsouza. Klout score: measuring influence across multiple social networks. pages 2282 – 9, Piscataway, NJ, USA, 2015//. URL http://dx.doi.org/10.1109/BigData.2015.7364017. Klout Score;influence measurement;social networks;influence scoring system;hierarchical framework;.

[21] Kylie jenner tweet hammers snapchat shares, 2018. URL http://www.bbc.com/news/business-43163544.

[22] Jordan Valinsky. Snapchat stock loses €643.00 million after rihanna responds to offensive ad, Mar 2018. URL http://money.cnn.com/2018/03/16/technology/snapchat-stock-rihanna/index.html.

[23] G. Preethi, P. V. Krishna, M. S. Obaidat, V. Saritha, and S. Yenduri. Application of deep learning to sentiment analysis for recommender system on cloud. In *2017 International Conference on Computer, Information and Telecommunication Systems (CITS)*, pages 93–97, July 2017. doi: 10.1109/CITS.2017.8035341.

[24] M.S. Neethu and R. Rajasree. Sentiment analysis in twitter using machine learning techniques. pages 5 pp. –, Piscataway, NJ, USA, 2013//. URL http://dx.doi.org/10.1109/ICCCNT.2013.6726818.

[25] Peter D. Turney. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. ACL '02, pages 417–424, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073153. URL https://doi.org/10.3115/1073083.1073153.

[26] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press, Cambridge, MA, 1998. ISBN 978-0-262-06197-1.

[27] Jaap Kamps, Maarten Marx, Robert J. Mokken, and Maarten de Rijke. Using WordNet to measure semantic orientation of adjectives. In *LREC 2004*, volume 4, pages 1115–1118, 2004. URL http://citeseer.ist.psu.edu/kamps04using.html.

[28] Yelena Mejova. Sentiment analysis: An overview. University of Iowa, Computer Science Department, 2009, 2009.

[29] Pedro Domingos and Michael Pazzani. On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning*, 29(2):103–130, Nov 1997. ISSN 1573-0565. doi: 10.1023/A:1007413511361. URL https://doi.org/10.1023/A:1007413511361.

[30] Z. Niu, Z. Yin, and X. Kong. Sentiment classification for microblog by machine learning. In *2012 Fourth International Conference on Computational and Information Sciences*, pages 286–289, Aug 2012. doi: 10.1109/ICCIS.2012.276.

[31] Luciano Barbosa and Junlan Feng. Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 36–44, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1944566.1944571.

[32] Y. Wu and F. Ren. Learning sentimental influence in twitter. In *2011 International Conference on Future Computer Sciences and Application*, pages 119–122, June 2011. doi: 10.1109/ICFCSA.2011.34.

[33] Atro Voutilainen. Part-of-speech tagging. *The Oxford handbook of computational linguistics*, pages 219–232, 2003.

[34] Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 42–47, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-932432-88-6. URL http://dl.acm.org/citation.cfm?id=2002736.2002747.

[35] Stephen Kelly. News, sentiment, and financial markets: A computational system to evaluate the influence of text sentiment on financial assets. *Trinity College Dublin*, 2016.

[36] Marco Lui and Timothy Baldwin. langid.py - stand-alone language identification system. https://github.com/saffsd/langid.py, 2016.

[37] Khurshid Ahmad, Nicholas Daly, and Vanessa Liston. What is new? news media, general elections, sentiment, and named entities. In *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011)*, pages 80–88. Asian Federation of Natural Language Processing, 2011. URL http://www.aclweb.org/anthology/W11-3712.

[38] Paul C. Tetlock. Does public financial news resolve asymmetric information? *The Review of Financial Studies*, 23(9):3520–3557, 2010. doi: 10.1093/rfs/hhq052. URL http://dx.doi.org/10.1093/rfs/hhq052.

[39] Tim Loughran and Bill McDonald. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65, 2011. ISSN 1540-6261. doi: 10.1111/j.1540-6261.2010.01625.x. URL http://dx.doi.org/10.1111/j.1540-6261.2010.01625.x.

[40] L. C. G. Rogers, S. E. Satchell, and Y. Yoon. Estimating the volatility of stock prices: a comparison of methods that use high and low prices. *Applied Financial Economics*, 4 (3):241–247, 1994. doi: 10.1080/758526905. URL https://doi.org/10.1080/758526905.

[41] Minitab Blog Editor. Multiple regression analysis: Use adjusted r-squared and predicted r-squared to include the correct number of variables, Jun 2013. URL http://blog.minitab.com/blog/adventures-in-statistics-2/multiple-regession-analysis-use-adjusted-r-squared-and-predicted-r-squared-to-in

[42] Barry O'Halloran. Ryanair carries more international passengers than any other airline, 2016. URL https://www.irishtimes.com/business/transport-and-tourism/ryanair-carries-more-international-passengers-than-any-other-airline-1.2768447.

[43] Stephen J. Taylor. *Asset Price Dynamics, Volatility, and Prediction*. Princeton University Press, stu - student edition edition, 2005. ISBN 9780691134796. URL http://www.jstor.org/stable/j.ctt7t66m.

[44] RTÉ News. Ryanair offers to meet with union before strike action, Dec 2017. URL https://www.rte.ie/news/2017/1216/927837-ryanair-impact/.