2016-1

Identifying Market Indicators and Content Quality from a Financial Micro-Blog Platform

Siobhán McNamara
Technological University Dublin

Identifying Market Indicators and Content Quality from a Financial Micro-Blog Platform

Siobhán McNamara

D14128664

A dissertation submitted in partial fulfilment of the requirements of
Dublin Institute of Technology for the degree of
M.Sc. in Computing (Data Analytics)

January, 2016

DECLARATION

I certify that this dissertation which I now submit for examination for the award of MSc in Computing (Knowledge Management), is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the test of my work.

This dissertation was prepared according to the regulations for postgraduate study of the Dublin Institute of Technology and has not been submitted in whole or part for an award in any other Institute or University.

The work reported on in this dissertation conforms to the principles and requirements of the Institute's guidelines for ethics in research.

Signed: _____

Date: **02/01/2017**

ABSTRACT

Investment platforms and discussion platforms have come to change the face of finance. The stock market is open to both professional and non-professional investors via online financial channels. Information too comes via a shared domain as both professionals and non-professionals log onto online communication platforms to share, search and discuss market trends. Due to their growing role in finance, understanding online communities has become the focus of much stock market research. Determining who is influential in a network, how information spreads and what translates to buy or sell decision is potentially very lucrative.

In this research paper a dataset from Stocktwits, a finance microblog, is analysed in order to determine a mechanism for identifying trustworthy and informative content in relation to Apple (AAPL) stock. Text analysis, user reputation classification and social network analysis are performed to generate features to measure correlations between the network and market changes.

Key words: *stock prediction, social network analysis, LDA, topic classification, sentiment analysis, reputational classification*

ACKNOWLEDGEMENTS

I would like to express my sincere thanks to my supervisor Dr. Pierpaolo Dondio for his direction during this Msc thesis. In addition to that I would like to thank Máiréad McNamara for her contribution to editing and to all manners of support during the study period.

I would like to thank both colleagues Kevin and Mick Cooney who provided technical and subject matter expertise during the research design.

Finally I would like to thank Fernando Ayuso for the huge contribution in editing the final document and ensuring something comprehensive and legible was delivered.

TABLE OF CONTENTS

DECLARATION **2**

ABSTRACT **3**

ACKNOWLEDGEMENTS **4**

TABLE OF CONTENTS **5**

TABLE OF FIGURES **8**

1. INTRODUCTION **10**

1.1 Background 10

1.2 Research Questions 12

1.3 Aims & Objectives 12

1.4 Research Methods 13

1.5 Scope & Limitations 14

1.6 Paper Outline 15

2. LITERATURE REVIEW **16**

2.1 Introduction 16

2.2 Role of Social Networks in Finance 16

2.3 The Transition from Tradition 18

2.4 Trust & Reputation 19

2.5 Learning 21

2.6 Text Analysis 22

2.7 Unsupervised Topic Classification 25

2.8 Partially Supervised Topic Analysis 26

2.9 Social Network Analysis 27

2.10 Social Capital Theory, Game Theory & Trust 28

2.11 Measurement in Social Network Analysis 31

2.12 Conclusions 33

3. DESIGN/ METHODOLOGY **35**

3.1 Introduction 35

3.2 Data Requirements 36

3.3Network Data 36

3.3.1 Data Availability *36*
3.3.2 Selection Criteria *37*
3.3.3 Features *37*
3.4 Financial Data 39
3.4.1 Selection Criteria *39*
3.4.2 Features *40*
3.5 Timeframe under Investigation 41
3.6 Data Preparation 43
3.7 Social Network Data Manipulation 43
3.8 Social Network Feature Generation 45
3.8.1 Text Analysis *45*
3.8.2 Reputation *47*
3.9 Financial Variables 48
3.10 Daily Table 49
3.11 Social Network Analysis 50
3.12 Final Dataset 50
3.13 Data Modelling 55
3.14 Model Selection 56
3.15 Model Performance Measurement 57
4. IMPLEMENTATION & RESULTS 58
4.1 Tools 58
4.2 Experiment Implementation 59
4.3 Data Exploration 59
4.4 Model Preparation 61
4.5 Baseline Model 63
4.5.1 Baseline for predicting AAPL closing price *63*
4.5.2 Baseline model for predicting AAPL volatility *63*
4.5.3 Baseline model for predicting AAPL Volume *64*
4.6 Network Models 64
4.6.1 Price Prediction *64*
4.6.2 Same day Price and network correlation *67*
4.6.3 Volatility Prediction *67*
4.6.4 Same day volatility and network correlation *70*
4.6.4 Volume Prediction *71*

4.6.5 Same day correlation *73*
4.7 Prediction of network variables *74*
4.7.1 Proportion of technical tweets in the retweet network *74*
4.7.2 The proportion of technical tweets in the full network *75*
4.7.3 The proportion of retweets that are Rank 2 users *76*
4.7.4 The proportion of tweets in the full network posted by Rank 2 users *77*
4.7.5 Gini Index Score of the daily In-Degree Centralities *78*
4.7.6 Retweet Network Modularity *79*
4.7.6 Retweet Network Assortativity *79*
4.8 Summary of Results 80
5. EVALUATION / ANALYSIS 81
5.1 Evaluation of Results 81
5.1.1 Retweet Network 81
5.1.2 Text Analysis 82
5.1.3 User Rating 83
5.1.4 Social Network Analysis 83
5.2 Observations from the Results 84
5.3 Strengths of the Results 84
5.4 Limitations of the Results 85
6. CONCLUSIONS AND FUTURE WORK 87
6.1 Research Overview 87
6.2 Contribution & Impact 88
6.3 Future Work & Recommendations 89
BIBLIOGRAPHY 91

TABLE OF FIGURES

FIGURE 3-1 *DEPICTS THE PHASES TO EXECUTING THE EXPERIMENT FOR ACQUIRING THE DATA THROUGH TO GENERATING A FINAL MODEL INCLUDING A HOST OF NETWORK AND FINANCIAL FEATURES* 36

FIGURE 3-2*: APPLE STOCK PRICE 2012 – 2016* 42

FIGURE 3-3*: APPLE STOCK PRICE DURING TRADING DAYS IN THE ANALYSIS PERIOD 2015* 42

FIGURE 3-4*: S&P STOCK PRICE DURING TRADING DAYS OF THE ANALYSIS PERIOD 2015* 43

FIGURE 3-5*: HISTOGRAM DISPLAYING THE DISTRIBUTION OF RANKS OVER THE RETWEET DATASET* 48

FIGURE 3-6*: THREE DAILY NETOWRKS DISPLAY THE MOST CONNECTED USERS IN THE CENTRE SURROUNDED BY USERS WITH JUST ONE CONNECTION* 50

FIGURE 3-7*: SCATTERPLOTS DISPLAYING PAIRWISE CORRELATIONS OF NETOWRK VARIABLES WITH AAPL CLOSING PRICE* 56

FIGURE 4-8*: BASELINE TREE FOR AAPL CLOSING PRICE TOMORROW* 63

FIGURE 4-9*: BASELINE TREE FOR AAPL VOLATILITY TOMORROW* 64

FIGURE 4-10*: BASELINE TREE FOR AAPL VOLUME TRADED TOMORROW* 64

FIGURE 4-11*: NETWORK PREDICTION OF AAPL CLOSING PRICE* 65

FIGURE 4-12*: SAME DAY NETWORK AND PRICE CORRELATION* 67

FIGURE 4-13*: NEXT DAY VOLATILITY PREDICTION* 68

FIGURE 4-14 *: NEXT DAY VOLUME TRADING PREDICTION* 71

FIGURE 4-15 *: SAME DAY NETWORK AND VOLUME CORRELATION* 74

FIGURE 4-16*: CORRELATION WITH PROPORTION OF TECHNICAL RETWEETS* 75

FIGURE 4-17*: CORRELATION WITH PROPORTION OF TECHNICAL TWEETS* 76

FIGURE 4-18*: CORRELATION WITH THE PROPORTION OF RETWEETS THAT ARE RANK 2 USERS* 77

FIGURE 4-19*: CORRELATION WITH THE PROPORTION OF RETWEETS THAT ARE RANK 2 USERS* 77

FIGURE 4-20*: CORRELATION WITH THE PROPORTION OF TWEETS POSTED BY RANK 2 USERS* 78

FIGURE 4-21*: NETWORK FEATURES CORRELATION WITH THE GINI INDEX SCORE OF THE IN-DEGREE CENTRALITIES* 78

FIGURE 4-22*: NETWORK FEATURES IN CORRELATION WITH THE RETWEET NETWORK MODULARITY* 79

FIGURE 4-23*: NETWORK FEATURES IN CORRELATION WITH THE RETWEET NETWORK ASSORTATIVITY* 79

TABLE OF TABLES

TABLE 3-1: *ALL NETWORK VARIABLES INCLUDED IN THE ORIGINAL DATASET* 39
TABLE 3-2*: LIST OF FEATURES RETAINED FROM ORIGINAL STOCKTWITS NETWORK DATASET* 43
TABLE 3-3*: FINAL DATASET GENERATED FROM NETWORK VARIABLES, FEATURES GENERATED AND FINANCIAL VARIABLES* 55
TABLE 4-1: *PRICE PREDICTION V'S OBSERVED PRICE VALUES* 65
TABLE 4-2*: VOLATILITY PREDICTION V'S OBSERVED PRICE VALUES* 68
TABLE 4-3*: VOLUME PREDICTION V'S OBSERVED PRICE VALUES* 71

1. INTRODUCTION

In this section an introduction to the subject matter and the questions pursued by this experiment are detailed. This is followed by a brief review of the methods applied and the potential shortcomings there might be in robustly answering the research questions. To finish there is an outline of this the paper.

1.1 Background

Digital media has brought about a huge change in the way people search for and acquire information. No longer is it exclusively the domain of professional printing houses and universities to disseminate knowledge. Through the internet and its variety of social and information platforms these industries have been decentralized. Today the resources to learn all manner of subjects are at our fingertips. From one perspective this is levelling the playing field. There are not a few who control information spread and can impose their agenda onto it, it is democratic as more popular pages proliferate further through sharing and search criteria. From another perspective the decentralization is diluting information, to the detriment of everybody. Facts blend with opinions and agendas, and a noisy information economy is developing. Those with subject matter expertise may be equipped to tell the difference between opinion and a piece of technical analysis. Though, without a formal education in a subject area, it is not clear there is a capacity to learn from online information sharing platforms (Casarin, Casnici, Dondio, & Squazzoni, 2015).

Occurring in parallel to this decentralisation of information is the decentralisation of finance. Investment has moved from the exclusive domain of professionals working for wealthy financial institutions to an open market enabling any lay man to part-take. These two trends have culminated to change the approach to finance. Today with professional and non-professional investors influencing market trends interpretation of markets requires a greater shift from rational models with order and harmony (Shiller, 2003). Measuring the decision making of non-professional investors requires a fuzzier logic and the inclusion of features which might measure populist interpretations of

markets. The behavioural economist Richard Thaler (2015, p.21) recently described the difference "compared to this fictional world of econs, humans do a lot of misbehaving, and that means that economic models make a lot of bad predictions".

In search of market information and prediction, buddying investors are turning to the internet, to digital chat platforms among other things to inform their investment decisions. This in turn is producing new sources of data, in fact stifling volumes of data on the expectations of non-professional investors. So too are professional investors active on the same internet platforms to gather and share knowledge (Sassen, 2005; Preda, 2009a; Knorr Cetina, 2005). Stocktwits is one such platform devoted exclusively to content relevant to the stock market. Information is shared in 140 character messages and may include a link, an image, symbols / emoticons and a bullish or bearish tag. Information spreads through follower / following relationships. A posted message can be favourited, retweeted which is a message re-share or replied to. Along with follower/following connections, retweets and replies enable the tracking of information flow. They indicate that a user has read a tweet and rates the messages as important enough to incentivise a re-share with their followers or a response (Zaman, Herbrich, Gael & Stern, 2010).

With the simple premise that expectations influence price, these platforms could offer an insight to how prices will behave over the near term. The lucrative gains from such a finding have incentivised many researchers to mine data from online platforms and to classify various features in search of indicators of future change in the stock price, volatility or volume traded. To date the results have been mixed. Correlations of network features with financial variables have been reported. Prediction however remains elusive, at least in the published literature.

In practical terms very little is known about the users of these platforms and the catalysts for their activity. Personal details are sparse and un-verified and the motivations driving their communication are opaque. Users' identity remains anonymous and they may write whatever they like about themselves. Their interested in a platform might be information search, to express opinions, to spread ideas, to gain approval or to spread news stories. There is no means to gain real insight into their expertise and their intentions. In their survey of an online financial platform Casarin,

Casnici, Dondio & Squazzoni (2015, p.51) found that non-professional users posted more during times of volatility and that the content of their messages changed, with spam and opinion based messages increasing during periods of higher uncertainty relative to more technical analysis during calm market periods. Simply put, they found volatility generates noise in the network. Other research, while attempting price prediction rather than a network analysis have offered insight to the mechanisms of online platforms. For example, Bollen, May & Zeng (2011 p.2) found predictive value in the incidence of the world bull or bullish on Twitter and stock price increase. In both of these examples the application of text analysis has has measured a relationship between network behaviour and market trends. In addition to that Casarin et al. (2015, p.51) classified users and measured a structural change in the network as more non-professional users were more active during higher volatility and a functional shift towards a greater proportion of information search.

1.2 Research Questions

The primary research question under investigation here is whether a filter can be generated that will robustly distinguish good quality information that gives an insight to market trends versus lesser informative pieces? A number of secondary questions are posed through the generation of features to classify and measure network characteristics.

- Do connections between users change under varying market circumstances?
- Do a retweet network condense more important information than the one contained in the full network?
- Does the text content of messages posted contain information that is indicative of investors intentions?

1.3 Aims & Objectives

The aim of this research is to build on the gains from previous efforts of text analysis, user classification and social network analysis to classify information and distinguish good quality tweets from noisy tweets that do not inform users about market trends. In

contrast to Casarin, Casnici, Dondio & Squazzoni (2015, p.51) who surveyed users, reputational features here are dependent solely on network behaviour, no additional information on users is available. Using this reputational feature alongside topic and sentiment features and social network analysis, correlations and a potential predictive capacity will be explored with respect to apple stock price, volatility and volume. The interactions between these features and their internal network relationships will also be explored.

1.4 Research Methods

The methodologies to generate the features and the final model are quantitative and informed from other research in the area. Features reported to contain predictive power or to correlate with market trends are engineered in this project in a manner that reflects the social network under investigation, while a greater emphasis is placed on trust and reputation metrics.

The network data came from Stocktwits, a platform that has drawn little attention in earlier research but which has a large and growing user base. In appearance and functions Stocktwits mirrors Twitter, perhaps the most popular network for informal financial chat today and the subject of a bulk of the research in the area. In contrast to Twitter Stocktwits is focused exclusively on the stock market.

In developing a reputation metric for this experiment two methods were applied. First a ‘retweet network’ was generated which contained retweets, replies and directed tweets. The users in this network are those replied to / retweeted. They had generated interested in their content and were contributing to the spreading of information. All additional features are generated for this retweet network and for the full network of tweets.

The second reputation metric was a user ranking system based on the number of followers, following and tweet count of users. Those in the top 20% of an aggregated score of these three features were ranked higher as rank 2 while the rest were rank 1.

A number of sentiment and topic classifiers were generated to test the premise that tweets can be bundled into broad classifications that have meaning in finance. A Latent Dirichlet Allocation model and a Supervised Bayes Classifier performed topic classifications splitting tweets into two bundles; a) opinion & spam, b) technical analysis & news. For sentiment classifiers three approaches were tested; the McDonald and Loughran dictionary of positive and negative words in finance was used to count positive and negative word instance in messages (McDonald and Loughran, 2011), the incidence of the words 'bull', 'bullish', 'bear' and 'bearish' in tweets were counted and the use of the inbuilt 'bullish' and 'bearish' tags were counted.

Finally, social network features were generated in order to test the assertion that the retweet network embodies network wide trends that correlate with changes in stock features. These ideas are borne out of social network theory and graph theory which has popularly been applied to the exploration of online social networks in an effort to map their user trends. Included in this experiment were assortativity, modularity and the Gini score of the in-degree centrality calculated using the Gini index.

1.5 Scope & Limitations

This study covers one year of data. It is sufficient to enable conclusions to be drawn and it is comparable to periods studied in other relevant research, some spanning a few weeks to a few months. However, others have covered a longer period and adding years to the analysis would open up more opportunities for exploration and would add weight to any of the findings reported.

In addition to the time limitation this experiment looks solely at the behaviour on one network associated with one stock. The entire network is omitted from this analysis, including only tweets that referenced AAPL. It was deemed that taking the entire network into account would enhance the noise and lead to the requirement for many more financial features to control for observed but spurious relationships captured. The inclusion of more stocks and more tweets referencing those stocks would be preferable however only AAPL was referenced frequently enough to enable robust analytics. This is likely associated with the non-professional user base. High capital stocks resonate

more with the public and in particular AAPL is a stock people are very familiar with due to the company's prominence in media.

This could be interpreted as a limited keyhole view of a market with a variety of push and pull factors at play. It could also be argued that such a keyhole view is required in order to carry out a more comprehensive study that isolates a few features of interest. Economics in particular is fraught by the inability to isolate and control for confounding factors. Comparability of features within the experiment and with features generated in other similar research is applied in the design of this methodology in order to better place it among the current knowledge and limit exaggerations or misguided conclusions.

1.6 Paper Outline

- In the following section, Chapter 2, is a review of the research already conducted in this space. In this section a full view of the findings to date and the limitation of those are presented. This is an interdisciplinary experiment and therefore literature is drawn from economics, computer science, phycology and sociology.

- Following on from that Chapter 3 contains the methodology. Here the financial and network datasets, the data exploration, feature engineering, and the final model are described in detail.

- Chapter 4 presents the results to the process described in Chapter 3.

- In Chapter 5 is an evaluation of the methods and the model to discuss how robust they are at measuring the underlying phenomena and the extent to which the final results generalizable beyond the data here alone.

- Finally, Chapter 6 is a conclusion and a review of the contribution of this experiment to the literature and where the best gains might be achieved next.

2. LITERATURE REVIEW

2.1 Introduction

Since their inception, online financial communities have grown and diversified. Among those most popular currently in the west are Twitter, Bloomberg Chat, Yahoo Finance and Linked-in. In each of these, a range of professional and non-professional investors seek knowledge on the market and share information. These forms have now attracted the interest of researchers and investors as a potential source of leading market indicators. The particular data under scrutiny and the methods used to unearth any patterns have varied. To date results have been mixed. What is clear is that there is no consensus on a pattern of network variables which are predictive of stock price changes.

In this chapter a review of the research to date will be presented. It stretches beyond strictly finance related subject matter to explore thoroughly the social, psychological, economic and modelling components of the experiment. Initially there is a review of the research to date exploring the relationship between online social networks and the stock market. Following on from that are sections detailing the most impactful research into feature generation from social networks, including trust, learning, sentiment and topic. Finally, social network theory and its role in feature generation is detailed. Entwined in each of these sections is a review of the relevant theory from economics, sociology and psychology to provide a deeper understanding of the drivers of market and network behaviours.

2.2 Role of Social Networks in Finance

Research into stock market prices originally focused on Random Walk Theory, part of Efficient Market Hypothesis (EMH) (Famma, 1991; Cootner, 1964; Famma, Firsher, Jensen and Roll, 1969). According to EMH stock price change is driven by information generation e.g. news, rather than present or past prices. As news is

unpredictable stock price fluctuation is also unpredictable, it is a 'random walk' and cannot be determined with any greater accuracy than 50%. However, the field has evolved vastly since then.

More recently the internet enabled anyone to access investment platforms so both professional and non-professionals share the space. It has also changed the way investors gather and share knowledge (Sassen, 2005; Preda, 2009a; Knorr Cetina, 2005). In addition to prices and statistics and research reports, investors can now look to online communities for ideas and information about real-time market trends (Tetlock, 2007). Analysis messages posted on Yahoo! Finance and Raging Bull, Frank & Antweiler (2004, p.1259) they found a correlation between message numbers and volatility and between the level of measured disagreement between posted messages and higher trading volumes. In 'The information content of stock microblogs, European Financial Management' the authors distinguished a quarter of a million tweets that were related to stock. With the application of a sentiment analysis of the text they found a correlation between mood and price changes and between disagreement and price volatility (Sprenger, Tumasjan, Sandner, Welpe 2014). Further they were able to identify trust in the community, higher retweet and follower numbers were indicative of users who had consistently in the past given high quality information.

In an effort to distinguish factors influencing contribution to online forums and the nature of those contributions. (Racca, Casarin, Squazzoni, & Dondio, 2016) mined 11 million messages from 2005 to 2012 from the Italian forum finanzaonline.com. They developed a ranking of forum users to distinguish expert investors from non-experts. Aptly this period covered the inception and development of the recent financial crisis. Applying supervised and unsupervised text classification and measures of market volatility they compared the content of messages pre-crisis, during the outbreak and finally the crisis progression. They found that expert investors on the form reacted to the crisis inception but less so during its progression. By contract volatility and uncertainty increased activity among non expert users particularly in their sharing of news items.

Regarding the predictive power of online financial communities, (Dondio, 2012) tried to correlate traffic on online communities with traffic returns, finding a positive

economic return but not statistically significant. In a follow-up paper - (Dondio, 2013) - the authors showed how the predictive power of online users is not the same for all the users, and provided a method to identify users with higher predictive potential based on past performance analysis.

2.3 The Transition from Tradition

Traditionally there are two categories for the analysis of investment decision making; (i) technical analysis and (ii) fundamental analysis (Sankar, Vidyarajb, Satheesh, 2014). Technical analysis looks at the price movement of security and uses this data to predict future price movements. Fundamental analysis, on the other hand, looks at economic factors, known as fundamentals (Jansen, Langager and Murphy, 2016).

Mining data from online financial communities' fits into fundamental analysis, using indicators of expectations and behaviours of investors from their online activity. However, the machine learning techniques for analysis do not fall within either field. Improvements in computational power have been enabling larger data inputs and methods developed in a computer science laboratory to find their way into financial analysis. Moving away from the traditional methodologies mixed methods employing fuzzy expert systems and artificial neural networks are more popularly employed to analyse stock prices and measure the attractiveness of stocks (Sankar, Vidyarajb, Satheesh, 2014).

These models enable fast processing of huge dataset, in complex models with a longer list of inputs. At the same time digital activity is providing larger and more diverse datasets to justify these complex models. Despite the promises some of the factors generating imperfect markets in traditional economic theory remain challenging in this new data and model rich environment. Information asymmetry refers to decision making in which one party has more or better information than another.

Sankar et al. (2014) point out the difficulty in reliably establishing trust in an online financial community. They point out the lack of information regarding other members and particularly a lack of clarity on the motivations of those investors. Noise or lies

could be intentionally generated. While these challenges are understood the solution to generating trust from online social networks is not clear. Classifying the type of knowledge different investors have is a problem and trustworthy experts for an investment decision are hard to identify (Sankar et al., 2014).

A detailed network structure is difficult to account for in the context of online forums. Little is known about the identity of the users and therefore classification based on trust, motivations, expertise or any other identifying features has not been possible. Disambiguating the noise on these networks requires the classification of 'useful' information with a degree of lag before the predicted market response.

2.4 Trust & Reputation

The dynamics and the dissemination of information on online platforms has been studied in a number of fields. In an analysis of users' credibility and influence on Twitter Abu-Salih, Wongthongtham, Beheshti & Zhu (2015, p.460) found a retweet scored higher for trust than a favourite. (Dondio, Barrett, Weber, & Seigneur, 2006) and (Dondio & Longo, 2011) and (Longo, Dondio, & Barrett, 2007) proposed a computational model of trust for online communities that he applied to Wikipedia, eBay and online financial communities such as Finanzaonline.com. The model computes trust as a reasoning process using the available evidence collected from the application, including users longevity, level of activity, persistency and profile information.

They concluded a retweet, an expression of trust, had greater influence on the network. Luo, Liu and David (2002) designed a decision support system to predict buy or sell decisions applying fundamental analysis with technical indicator systems. Reputation or trustworthiness here functioned as the classifier to distinguish information that is useful from information that is not useful for predicting stock price. The authors explain that labelling data as such enables a reduction of noise from the use of fuzzy expert systems.

Users influence on social networks has also been the focus of analysts in a number of other domains to generate personalised recommendation systems and expertise

retrieval (Salih, Wongthongtham, Beheshti & Zhu, 2015). Matchbox is a Bayesian inference model that makes use of user and item meta data and binary feedback to recommend future user preferences (Stern, Herbrich & Graepel, 2009). It was originally designed as a movie recommendation system. Zaman et al. (2010) applied it to predict retweets. They trained the model on tweeter features, retweeter features and tweet content and found that relevant features for retweet prediction were the tweeter and retweeter. Examining network attributes of stock related message behaviour on Twitter Sprenger & Welpe (2010, P.89) found that retweets contained 'above average investment advice', though this relationship corresponded to the users who tweeted and retweeted rather than the message content.

In an effort to gain greater insight into users of online financial communities (Casarin, Casnici, Dondio, & Squazzoni, 2015) surveyed members of a popular Italian online forum. This was designed to investigate the motivations, risk propensity, education and online experience of investors on the platform. While acknowledging a potential bias from self-selected survey responders they reported that knowledge sharing and learning in virtual communities did not facilitate better investment decisions for non-professional investors, while professionals were able to distinguish useful information from noise. They found that online exposure increased participants' propensity for risk. This corresponds with the positive bias or bullishness reported in studies correlating stock prices and online forum messages.

Sankar, Vidyarajb and Kumarb (2015, p.299) used a social network approach to analyse the activity of a portfolio of stocks in 'Trust Based Stock Recommendation System'. From that they developed a recommendation model for amateur investors tailored to their preferences. Their goal was to determine a trusted social network, in an attempt to reduce the noise that is encountered when analysing the data from online communities. To this end they defined a collection of 'trusted mutual friends' and their portfolio of mutual funds based on the ratings of their stock by an independent rating agency. They believe their model is an improvement on those using data from online platforms as experts or trustworthy members are difficult to distinguish and pursuing an investment decision of a trustworthy mutual fund is less risky than following advice from an expert individual. It is intended that this improves the quality of information

for amateur investors and also that it saves a lot of time for them by filtering information to the leading stocks and including the investment price range.

In the research into finanzaonline.com (Casarin et. al, 2015) the authors were attempting something close to a sociological experiment to understand what the investor chatting online is like and what the implications of their online exposure are. The expertise of those involved were a mix of sociologists and computer scientists. The theory driving the survey was informed predominantly by sociological models. The survey provided an insight into the demographics of online communities and the role the platform plays in their investment decisions, that could not otherwise be inferred. This tool stands out among the research in this area as it goes beyond building automatic classifiers to infer the content and user characteristics of a network, but asks them directly and in turn uses that information to extract more information from the technical tool box.

Similarly, in 'Trust Based Stock Recommendation System', such insight to investors risk propensity and to their investment decisions was achieved but by contrast this was a purely quantitative study. The authors applied social network analysis methods to analyse financial markets – stock performance and price. In both cases, the goal was to establish a robust understanding of trustworthy or reputable investment decision making. Casarin et al. (2015, p.51) concluded that only formal financial education and trading experience promote good portfolio performance, and help investors to keep risk under control. 'Trust Based Stock Recommendation System' led with this conclusion. They selected an experienced and successful set of investors with strong performing stock portfolios as their reputable base. They designed a low risk investment recommendation model based on this small reputable population.

2.5 Learning

Learning has been associated with information shared among online communities (e.g., Anderson, 2004). However, financial research has typically been focused on the activities of financial institutions and professional investors. Little has been done to understand how this decentralized community learns (Casarin, Casnici, Dondio,

Squazoni, 2015). In the context of stock investment learning, indicators of market activity and market expectations could be gleaned from online forums. Here information regarding events taking place and the opinions of investors culminate to potentially provide insight for investment decisions e.g. if the market expects prices to increase they buy and the prices are driven up, if they expect them to decrease they sell and prices are driven down. However, there is no information regarding the demographic and motivation of the investor on an online forum. Perhaps they work for a particular financial institution whose interests they are pursuing, or in the case of non-professional investors there could be doubted over how accurate / trustworthy their perspectives of market events are. Casarin et al, (2015, p.51) found that expert investors were able to extract useful information for the Italian forum whereas non-professionals did not have the same ability to discriminate and neither did time contribute to their ability. Put another way, they did not learn after months and years of exposure to the forum. Those with a formal education had the ability to gain knowledge whereas the plethora of noise distracted the rest. Education and news via internet platforms has proliferated. A means to distinguish good quality from bad, useful from useless and to learn is at the heart of its value. However, it is not clear that that extension exists. With unorganized and unfiltered information there may be little value without the additional ability to identify *learning* material. There is no evidence that gaining information from internet based communication platforms affords an understanding of market sentiments and trends.

2.6 Text Analysis

In investigating what is useful, with little information available on the user, research to date has predominantly focused on the content of messages shared and their respective sentiment (Tetlock, 2007; Das & Chen 2007; Oh & Sheng 2011; Bollen, Mao & Zeng, 2011; Pang and Lee 2008; Mao, Counts & Bollen, 2015). However, results of topic and sentiment analysis have been mixed. In particular, when it comes to stock price correlation analysis a positive bias has been common in the research e.g. when the price is going up a positive sentiment on online forums is correlated with its persistent inflation.

A sentiment analysis measures whether a message is positive or negative or in this scenario whether it indicates a buy or sell. The premise behind a sentiment analysis is simple, if market actors are optimistic of price rises they will buy stock thus pushing the price up and visa versa if they are pessimistic about the market they will sell or hold their position and prices may drop. Mood has also been investigated more broadly for its role in investment. We know from psychological research that emotions, in addition to information, play a significant role in human decision making (Kahnehman and Tversky, 1979). Additional work in behavioural finance has identified a role for emotion in finance related decisions (Nofsinger, 2005). A sentiment analysis therefore attempts to classify market actors' moods and / or their expectations based on positive and negative language and infers a price rise or drop from that trend.

Studies have varied predominantly on the dictionary applied to define the sentiment of messages. Tetlock (2007) measured the frequency of words in news items and classified them based on the Harvard negative word list (Harvard-IV-4-TagNeg) to develop a pessimism indicator. However, the Harvard dictionary was developed for use in psychology and sociology. Loughran and McDonald (2011) found that vocabulary classified as negative by the Harvard dictionary was not negative in finance. They developed a financial negative word list of 2337 words and reported their dictionary outperforming the Harvard dictionary in measuring financial sentiment. Mao, Counts & Bollen, 2015 p.4 pointed out the difficulty in classifying the 'variegated contexts and subtleties of human language' while performing a sentiment analysis of Twitter. Text analysis of tweets has met with more difficulties than text in longer paragraphs or news articles due to their limited word count (140 characters) and their syntax often a mix of shorthand, emoticons and hashtag summaries. For that reason, a variety of methodologies have been formed to tailor to these syntactical challenges. It the following two examples language lexicons were omitted and a very crude classifier achieved a more powerful prediction that has been reported elsewhere.

In 'Quantifying the effects of online bullishness on international financial markets' Mao, Counts & Bollen, 2015 p.4 tried to reduce the complexity of a sentiment classifier to two distinguishing features. They classified mood from Google searches and Twitter messages to be positive or negative based on a mood tracking tool. In this

case the market mood is measured from Twitter posts (tweets) using two mood tracking tools, OpinionFinder and Google Profile of Mood States GOPMS. OpinionFinder divides mood as measured from the tweets into positive and negative groupings. The GOPMS by contrast has six levels 'Calm, Alert, Sure, Vital, Kind, and Happy'. The authors undertook this approach as they believed that mood would be an indicator of emotions and emotions with regard to stock investment decisions would be correlated to prices e.g. a positive mood would be indicative of buying and rising stock prices and negative of selling and dropping stock prices. They found that 'calm' and 'happiness' as per GOPMS was correlated with stock price changes.

Bollen, May & Zeng, (2011 p.2) in 'Twitter mood predicts the stock market' made headlines for their finding that they had successfully predicted the stock market with Twitter. This case is purely a sentiment analysis. They tested both Google searches and tweets for the presence of finance related text. Within this text the counted the presence of the words 'bullish' or 'bearish'. The volume of these sentiment indicators was a correlated with the Dow Jones Industrial Average (DJIA) for the United States, the FTSE 100 for the United Kingdom, the S&P/TSX Composite Index (GSPTSE) for Canada, and the SSE Composite Index (SSE) for China.

A positive bias is common in efforts to predict stock prices from text analysis. That means when the price is going up text classifiers can predict with some power that it will continue to appreciate but a change from that trend or a price retraction cannot be predicted. The authors of 'Twitter mood predicts the stock market' reported that their 'bullish' / 'bearish' analysis could predict the stock market. However, it was the case that only the volume of the word 'bullish' could predict an increasing price during an upward trend. The reverse was not the case while a price retracted, neither was it able to predict a point of change or a measure of volatility.

In 'Back to Basics! The Educational Gap of Online Investors and the Conundrum of Virtual Communities', the authors found from a survey of social network users that those who are active on social networks report a bias towards bullish investment, (Casarin et. al, 2015). However, aside from sentiment analysis a positive bias has been found in other methodologies including the case study 'Trust Based Stock Recommendation System' which was designed to eliminate the untrustworthy nature of sentiment analysis

(Sankar et al. 2014). The author designed an investment recommender system dependent exclusively on other investors past success and current preferences. In this case the positive bias existed as the success of the model was in the buy recommendations but not on the sell side.

Looking deeper than the 'positive-negative' division it is unclear if emotions or expectations can be robustly determined from sentiment analysis. In addition to that it is unclear if such emotions could be captured in the 140 characters per message on Twitter and Stocktwits.

2.7 Unsupervised Topic Classification

The above methods applied to sentiment analysis refer to supervised topic classifiers. The classifications are pre defined, e.g. from lexicons of positive and negative words and text is allocated based on the count and importance of its positive and negative words present. Unsupervised topic modelling algorithms, such as Latent Dirichlet Allocation (LDA) (Blei, Ng & Jordan, 2003) and related methods (Blei, 2012) have grown in popularity.

With LDA each document is represented as a probability distribution over topics, where each topic is modelled by a probability distribution over words in a fixed vocabulary (Nguyen, Billingsley, Du & Johnson, 2015). The number of topics are pre determined by the researcher and the model determines the probabilities of each word in each document belonging to each topic.

Topic modelling techniques such as LDA have most successfully been applied to corpora composed of long documents with regular vocabulary and grammatical structure, such as news and scientific articles (Alverez-Melis & Saveski, 2016). Efforts to apply them to Microblog text such as tweets which are often short and noisy, topics have been uninformative and hard to interpret (Alverez-Melis & Saveski, 2016). For this reason, other techniques to pool tweets and make a more homogenous set prior to applying LDA have gained some traction. This is seen as a means to reduce the noise and enable a higher quality division of topics within a topic grouping.

Alverez-Melis & Saveski (2016, p.519) compared both LDA and Author Topic Model (ATM) for identifying topics among tweets that were un-pooled, pooled by user, pooled by hashtag and pooled by conversation. In order to evaluate the efficacy of each of the eight samples they manually labelled a number of tweets for comparison with the model classification. They found that homogeneity from pooling did improve classification over un-pooled. The LDA on pooled hashtag performance was highest followed by conversations.

Among the most successful techniques of applying LDA to tweets has been merging all of the tweets by each user into a single document and defining the user's topic distribution (Hong and Davison 2010) or merging all those containing the same hashtag (Mehrotra et al. 2013). Both of these have drawbacks and it is the nature of the text and the research dictate which it most suitable. For example, clustering by author classifies the authors' writing rather than the documents themselves and time relevant features are lost. In the case of hashtag pooling homogeneity is not always the outcome, as the interpretation of any given hashtag can vary vastly between users. In addition, hashtag pooling has resulted in tweet duplication and longer training times (Zhao et al. 2011).

2.8 Partially Supervised Topic Analysis

Methods for performing partially supervised Topic classification have met with mixed success. For example, Sahami and Heilman (2006, p.377) and Phan, Nguyen, le, Neguyen, Horiguchi & Ha (2011, p.961) employed external sources of information. Sahami and Heilman (2006) used web search data to give greater context to their text classification and Phan et al. (2011) used models trained on larger corpus such as Wikipedia to classify microblog text on a related topic. In this later example homogeneity of the external and the microblog topic is crucial. A deviance from the external source and the the topics will become noisy and impractical.

Choosing the topic number has been highlighted as a weakness of all LDA models, irrespective of any alterations to the raw unsupervised model, as it is unknown how

many topics a text may contain. LDA has been reported to perform poorly with few, while the computational power for more is very expensive. Due to the size, diversity and computation power required, modeling content on Twitter requires techniques that can readily adapt to the data at hand and require little supervision (Ramage, Dumain & Liebling, 2010).

In order to mix sentiment and topic classification Si, Mukherjee, Liu, Li, Li & Deng (2013, p24) applied a Continuous Dirichlet Process Mixture (cDPM) model to learn the daily topic set of Twitter messages in order to predict stock price fluctuations. A sentiment time series was built based on these topics. Accuracy was poor, however it does not discredit the model. Their sentiment topic dual model omitted any given stock as a classifier. Distinguishing topic, sentiment and the particular stock in question can add to the power of predictive models (Nguyen & Shirai, 2016). In contrast the ambition for a global market predictor is very ambitious.

Topic Sentiment Latent Dirichlet Allocation (TSLDA) estimates opinion word distributions for individual sentiment categories for each topic (Nguyen & Shirai, 2016). An extension of LDA, TSLDA infers topics and sentiments simultaneously. Both are interpreted for each sentence with sentiment linked to adjectives and adverbs and topics linked to nouns.

In this paper the topic was already refined to Apple (AAPL) stock, while two topic classifiers and two sentiment analysis methodologies were applied to extract value from the text. This was intended to perform in a similar manner to a topic – opinion – sentiment model, though without requiring one model to distinguish each nuance. Tweet text is so short and littered with abbreviations, slang, URL links, emoticons and turns of phrase that make it very difficult for a model to accurately classify. It seems from this review of text analysis methods that in fact those cruder models to generate simple positive negative groupings have performed best, particularly when financial correlation is sought.

2.9 Social Network Analysis

Social Network Analysis (SNA) is a means of measuring and mapping relationships in a similar manner to mapping physical networks. It applies network and graph theories to measure social structures, (Otte, Rousseau, 2002). Borne from sociology, social network analysis has developed into an inter-disciplinary endeavour and has been applied across a range of specialisations; biology, economics, anthropology, history, organisational studies, political science, development studies and computer science, (Wasserman and Faust, 1994).

A simple analogy of SNA application is a network of train routes, the map of the London Tube. Connections on routes and distances are inferred in a manner that is simple and intuitive. In reality it does not reflect the physical structure of the underground system but it is an approximation that is far easier for a reader to interpret and it appropriately represents the relationships between the routes. Similarly networks of people can be mapped and aspects of it measured. In the context of online social networks, people are connected by their 'friendship' of 'following' status and also by messaging. Networks contain their own intricacies and the goal of any analysis will vary, particularly across disciplines. However, no matter the application, challenging all social network analysis is the huge bulk of noise within networks and the classification 'valuable' information. Noise refers to content that is not reliable and does contribute to learning, knowledge gain and informed decision making. Due to the gravity of this challenge no singular tool nor application of SNA has come to define it.

2.10 Social Capital Theory, Game Theory & Trust

Social capital theory defines networks as collections of common people with key players at the centre. Norms are defined by those at the centre and spread through the network bringing about a convergence of in-groups (Garson, 2006). Social capital are the resources inherent in social relations which facilitate collective action (Garson, 2006) . Those include trust, norms and networks between participants that share some common goal or purpose.

With respect to digital networks social ties can be conceived as channels for the flow of data or information. The network itself defines the goal of common purpose of the

participants, it gives context to the actions and the resources that define the context of connections. For example, on Linked-In, a platform to facilitate professional networking, the connections and flow of data are intended to be based on career direction.

Identifying the prominent players, those who are trusted or have large reputations and those who's actions predicate the changes under analysis is the goal of SNA. In markets there are some huge financial institutions, central banks and some key commentators who hold great sway in financial networks. Similarly, in politics there are those who can swing public opinions, who can catalyse great change by spreading ideas through networks of trusted connections. These are a few examples of the catalysts of trends. If the same is true of all networks, if they are all dictated by a set of social norms and adherence to the moves of few key figures then identifying those in digital social networks could be an effective filter for noise.

Compromise and trust have often been highlighted as crucial to forms of resilient social capital. In analysing the diamond trade in New York city Chung, Piraveenan, Levula & Uddin (2013, p.1993) found trust was established by proximity. Vendors in the New York diamond market relied on neighbouring merchants to inspect and verify the authenticity of their collection. No third party oversight or escrow played a role in dissuading theft of underhanded behaviour. The process worked extremely well. Diamond trade in New York was considered an efficient and lucrative market for vendors with trust, borne out of merchant proximity, ensuring the sustained integrity of this simple system (Cloeman, 1998).

Such sustaining economic relationships could also be described from the perspective of game theory, a classical economic theory that explores conflict and cooperation among intelligent rational decision-makers (Myerson, 1991). Game theory is mainly used in economics, political science, and psychology to describe when cooperative and selfish decision prevail. Originally as a mathematical problem it addressed zero-sum games, in which one participant gains with an implicit consequence for another. Since then the field has grown, as part of the Nash Equilibria a cooperative game generates what is know as 'the prisoners' dilemma' e.g. two prisoners in two cells are faced with the same option, snitch on their counterpart and receive a smaller sentence, if their

counterpart snitches first they receive a life sentence, alternatively neither snitch (Nash, 1951).

Elaborated upon in experimental economics and psychology, behavioural game theory was developed to describe how social cooperation alters the factors influencing a rational decision maker's choice. It has been found that proximity and transparency similar that experienced by the diamond sellers above changes the optimal decision for players. If their decision is anonymous they will decide upon the selfish outcome with no regard for the other. If on the other hand they are facing their fellow player while they make the decision the people by and large opt for the pareto-efficient choice, the one that makes everybody better off and does not attribute gains to one with implicit consequences to another (Gintis, 2009). In addition to transparency norms, dominant players and social punishment for deviating from social cooperation have been demonstrated by the economist and behaviour scientist Herbert Gintis in a series of economic experiments (Gintis, 2009). Game theory highlights the complication in establishing a trust / reputation metric for users of online social platforms who remain anonymous.

In an effort to generate trust methods similar to those that upheld trust in the New York diamond market are mirrored in number of popular online platforms. For example, on the popular AirBnB website good behaviour, honesty and safety are maintained by reviews creating transparency of who is a fair player and who is not in the market. Similarly, digital restaurant locators include rating systems voted for by the pubic and reviews again generating transparency of good and bad service. In the context of online financial platforms, user anonymity creates a huge challenge in defining trust.

A thorough SNA analysis that reveals the motivations of members and the contagion of behaviours through connections is so complex due to it's horizontal network wide approach. This is in stark contrast to the focus on the rational decisions of a single unit as in Game Theory. Actors in SNA are interpreted as interdependent rather than independent, links between members represent a flow or transfer of resources. The network both constrains and provides opportunities for members with unifying features. The network models mirror robust relationships between individuals

conceptualizing a definite structure to behavioural patterns (Wasserman & Faust, 1994)

2.11 Measurement in Social Network Analysis

Graph Theory is at the core of social network analysis. It enables the identification of a number of features of a network and their measurement. Social networks can further be visualized with points representing network members and lines the connections between them. Nodes may be structural representing actors themselves of compositional representing their characteristics in an effort to understand connection between shared traits rather than individuals. Wasserman and Faust (1994, p.71) define a graph for SNA as "a single non-directional dichotomous, relation, the node represent actors, and the lines represent the ties that exist between pairs of actors on the relation". A position of importance of a node in a network can be defined by their centrality, that is a measure of the number of connections they have to other nodes in that network. A more connected or central node can be interpreted to have a greater degree of influence over the network or to be of greater importance or prominence to the other members, (Bourdieu, 1986).

Social network graphs depict and measure two network features; the number or set of nodes N = {n1, n2, n3. . . } and the set of lines L = {l1, l2, l3} between pairs of nodes where N = Nodes and L = Lines. Links may or may not be present. If absent it is a network of undirected dichotomous relations, if present they indicate connection and may be directed or undirected (e.g. display sender and receiver). The strength of a connection can be denoted by the frequency and / or duration of a connection. In this paper duration over time was not a metric but frequency of connections within a given day were measured to define daily centrality of nodes.

If the network is directed (meaning that ties have direction), then we usually define two separate measures of degree centrality, namely in-degree and out- degree. In this instance the person being referenced is the 'in' node. In-degree is a count of the number of ties directed to the node.

Degree centrality has been applied to the measurement of financial networks to identify stocks influential in aggregate market price fluctuations following the collapse of Lehman Brother in 2008 (Roy & Sarkar, 2011). While channels of financial contagion have long been investigated through trade and monetary links, with the application of SNA metrics this research highlighted European based stocks as the more influential in the global market.

Zheng Chen and Xiaoqing Du (2013) found a correlation between the volume and price of stocks traded on Shanghai/Shenzhen stock exchange and the online Chinese stock forum Guba.com.cn by measuring the average degree centrality and degree centrality of the network. As a learning tool for new investors Koochakzadeh et al (2012) created a social network of financial experts categorized according to their respective interests and behavioural trends. Degree centralities were measured to define investor risk appetites relative to one another. This was then used to generate investment suggestions based on the decisions of investors with similar interests and risk appetites. Badham (2013, p213) added a Gini calculation to the in-degree centrality measures of all nodes in order to generate a normalized metric of in-degree centrality. The Gini index is commonly used in economics in order to measure relative income equalities in a region. Values range from [0, 1] with 0 indicating complete equality and 1 complete inequality. A higher value in the Gini-index of the in-degree centrality indicates greater network inequality, or more messages are targeting fewer network users. This was later applied by (Casnici, Dondio, Casarin, & Squazzoni, 2015) when they used a normalised measure of the in-degree centralities of the Italian network finanzaonline.com in order to measure relative user influence under varying market conditions.

Assortativity another common metric in SNA, measures the degree to which similar users connect.. Similarity may be defined as a range of characteristics, however often in the context of digital social networks it refers to the number of messages one sends and receives. It could be considered as a measure of networks within networks. In visualization this is depicted as clustered hubs within the larger connected network of sparse connections. For example, in an experiment of the mood associated with connection on Twitters Bollen, Mao & Zeng (2011, p.7) used the ‘Subject-well-being index’ to measure users happiness. They found that users who were measured as

‘happy’ tended to connect with other happy users whereas unhappy users predominantly connected with other unhappy users.
Modularity measures the number of edges or links in a cluster minus the expected number of edges. It is as a measure of the strength of the division of a network into modules or into clusters. Borrowed from biology where systems can be defined in terms of their regional regions such as the brain. Beyond biological systems alone many more networks of interest in the sciences, including a variety of social networks, are found to divide naturally into clusters or modules (Newman, 2006).

Social Network Analysis is a hugely powerful tool to see inside communities of mass movement. However, despite their defined structures, the connections within and the influence of central nodes, networks are not organized and rational structures. They evolve and shift and influence develops and mitigates along with a host of external factors.

2.12 Conclusions

Data from online financial forums are drawing increasing attention due to the scale, range of detail from news to personal perspectives, the frequency of update and the completeness of the participating community. Despite research efforts it’s predictive capacity remains in question, largely due to challenges in reducing noise from the data, identifying trustworthy members or information, a lack of information about the members posting and a frequent positive bias in the results.

These research pursuits into digital social networks and their relationship with finance hope to understand whether there is a clear indicator of investment decision making shared online and if that is reflective of wider market sentiment. Increased data availability and the improvement of computational power are driving intrigue in more complex classification models and deep learning / fuzzy expert systems to understand and predict market trends.

What has been learned from research to date is that there is an abundance of non-professional investors using online social networks for indicators of smart investment

choices. Those tend to have a relatively high risk propensity and a positive bias with regard to stock investment decisions. Expert investors gain useful market knowledge from online communication platforms. However information shared on the network does not seem to facilitate learning among non-professional investors. There is a lot of noise and an inability for non-professional investors to discriminate meaningful or helpful information. Among the noise however correlations of behaviour have been found with price. The most lucrative question remains unanswered, whether there are indicators or future price changes buried in these digital social engagements.

3. DESIGN/ METHODOLOGY

3.1 Introduction

In this chapter the methods applied to execute the experiment are detailed:

1) Initially the data collection process is described. This includes the consideration of suitable datasets, the criteria required and the constraints. Following that is a description of the initial phase of data exploration to refine the datasets to those features already present which were of relevance to this experiment
2) Following on from that are three sections which reflect the three pillars of the social network measurement: a review of the four models applied to text classification in order to extract value from the unstructured tweet content; the social network variables measured in order to measure the collective behavior of a network on nodes and edges within the retweet network; finally the generation of a reputation variable in order to ascertain whether reputations do develop in anonymous networks and if so to measure their relationship with stock market variables
3) Financial features were generated as a compliment to the original market open and close prices, these included the volatility, historical volatility and the performance of apple relative to the S&P500.
4) Finally, the generation of a model is described in which each of the above social network features were included as independent variables while the financial data features were the dependent variables.

Experiment Execution

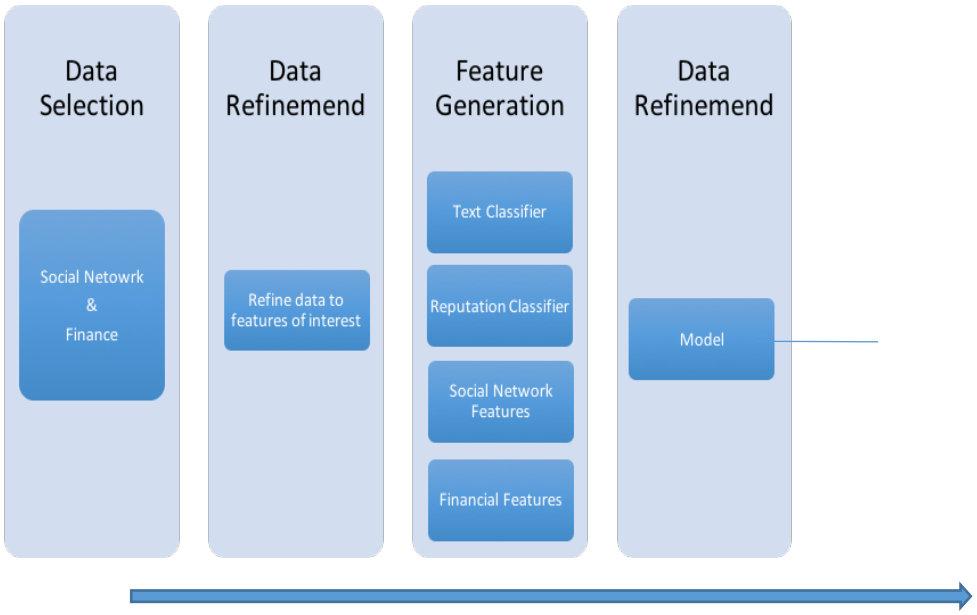

Figure 3-3-1 ***depicts the phases to executing the experiment for acquiring the data through to generating a final model including a host of network and financial features***

3.2 Data Requirements

Two dataset were required for this experiment. One from a social network comprising online conversations about the stock market and preferably a large body about particular stocks. In addition, that financial data reflecting the same stocks under discussion was required. The financial data at a minimum would include daily opening and closing prices of particular stocks and the daily opening and closing of the S&P500.

3.3Network Data

3.3.1 Data Availability

Acquiring a dataset adequately large and granular to enable the desired network and market analysis carried out for this project was one of the greatest challenges. Data of such nature is often considered very valuable, as the premise of this research has outlined and therefore is scarcely available.

In order to perform a thorough analysis of a network that might reflect market behaviour a representative network had to be analysed. That is a platform that enjoys a large amount of traffic and a large user base (relative to other online platforms).

A Stocktwits dataset was selected as the online financial platform to analyse for network effects. Stocktwits, a large and growing financial platform, was founded in 2008 and has grown in size and usage to accommodate close to 300,000 members today. In this instance the experiment met with great fortune as a colleague pursuing doctoral research acquired the data and shared it for the purpose of this experiment. The dataset for 2015 comprised of 14,638,930 messages.

3.3.2 Selection Criteria

Due it's huge popularity and the shrinking and converging of other networks Twitter was the first considered and sought after network. In addition, Twitter has been the topic of much research generally exploring facets of online platforms and specifically searching for financial market indicators. One large enough was not possible for the intended experiment as Twitter limit scraping and charge a high cost for even small datasets. Stocktwits however mirrors Twitter in it's structure, function and appearance. The user base is adequately large and the with tweet content limited to the stock market it lended itself to a more focused financial experiment than a more general network such as Twitter might.

In order to make a contribution to the most recent literature focused on online networks and market correlations something almost identically structurally and functionally to Twitter was deemed an excellent resource.

3.3.3 Features

The original Stocktwits dataset spanned 6 years from 2010 to 2015 inclusive. It was broken into monthly files of json data, comprising a row for each Tweet posted on the

network and 26 columns of metadata detailing the timestamp associated with the posting, the message text itself, a number of ‘Tweeter’ descriptors and a number of ‘Tweet’ descriptors.

Column	Data Type	Description
Id	Varchar	Unique id associated with each tweet
body	Varchar	Tweet text
id	Varchar	Unique id associated with each user
objectType	Varchar	Indicator whether it is a person or an automated firehose posting
displayName	Varchar	Tweeter name
preferredUsername	Varchar	Tweeter chosen Username
followersCount	Int	Number of followers
followingCount	Int	Number of following
statusesCount	Int	Number of tweets to date
summary	Varchar	Self composed profile
links	Varchar	Any links a tweeter associates with their profile
image	Varchar	Link to a chosen profile picture
tradingStrategy	Varchar	Self composed description of trading habits
approach	Varchar	Self composed description of trading decision mechanism
experience	Varchar	Self-proclaimed professional or amateur
classification	Varchar	Self-proclaimed professional or amateur

id	Varchar	Id associated with tweet type
objectType	Varchar	Text of link
postedTime	Timestamp	Time of tweet posting
updatedTime	Timestamp	Time of tweet update
summary	Varchar	A brief self composed profile
link	Varchar	Any additional links added to summary
symbold	Varchar	Any symbols added to profile
sentiment	Varchar	A sentiment flag for each tweet
Chart	Varchar	Link to a chart added to tweet
Video	Varchar	Link to video attached to tweet

Table 3-1: ***all network variables included in the original dataset***

3.4 Financial Data

3.4.1 Selection Criteria

Any stock under consideration would have to be listed on the New York Stock Exchange (NYSE) and to be listed among the S&P500. Originally 31 high capital stocks were shortlisted (Apple, Facebook, Twitter, Google, IBM, Microsoft, Dell, Eyegate Pharamaceuticals, Ernst & Young, KPMG, PriceWaterHouse Cooper, Tata Consultancy, Infosys, Accenture, Cognizant Technology Solutions, Nke, Addidas, Tesla, Netflix, Amazon, JP Morgan, Barclays, BNP Paribas, Deutsche Bank, Starbucks, Wallmart, Toyota, BMW). Discussion on online financial platforms tend to focus more on high cap stocks. This was also the case with Stocktwits users. Further the S&P500 was used as a reference for market price change, volatility and volume

against which to evaluate the performance of any stock under investigation in this experiment. It is a useful reference as volatility among the largest companies will often be lower than that experienced by smaller companies. Further the low volume of stocks of small cap companies in the market can skew the analysis and in particular do not play a large role in non professional chat forums (Casarin et. al., 2015). One drawback of this could be that those in the S&P500 change. In this case it did not matter, such stocks were not under consideration.

It was found that in fact only Apple stock was discussed in enough of the Stocktwits dataset to enable this particular experiment. The experiment sought close to 80,000 tweets annually and 89,000 were found in relation to Apple for 2015. These criteria were outlined in order to ensure a detailed enough analysis to enable power in any correlation or prediction measure without the concern for overfitting and further to enable comparison to other similar studies. Messages referencing Apple stock were selected based on the presence word ‘Apple’ itself or the ticker ‘$AAPL’ in any given Tweet. Tickers were introduced initially by Stocktwits to enable stocks to be traced in much the same way that a hashtag can trend and be traced. The ticker is a link and leads to all messages posted containing that ticker on the network in order of recency. This feature has since been added to Twitter.

3.4.2 Features

For the time frame covered apple stock data detailed a row for each trading day of the year. Features included the opening price, the closing price, the volume at close, and four measures of volatility.

Daily volatility and historical volatility were generated using the Garman-Klass Yang-Zhang (GKYZ) estimator (Yang & Zhang, 2000). This measure combines the previous close, and the current open-high-low-close each day which includes the opening, closing, high and low price of the day. The advantage of the GKYZ is that in volatile days of trading measures that take into account close to close or open to close prices exclusively, do not measure that intraday change. They would close reasonably unchanged whereas the GKYZ will measure far higher for volatility. For this

experiment volatility throughout the trading day was very important as the network is active all day and volatility if it does impact network behaviour would be expected to impact it in real time rather that in a fashion that summarises the end of day close.

The four volatility measures included were daily volatility as described above and then an index of historical volatility generated from the same measure over 10 days prior to each trading day, 20 days, 60 days, 120 days and 252 days. The historical volatilities were only applied in the initial exploratory part of the analysis though for the final measurement of correlation it was real time volatility, price and a number of financial features generated that were adopted as dependent variables.

3.5 Timeframe under Investigation

While network data covered a six-year period only the most recent year 2015 was included in this experiment. The potential for an expansion of the analysis over the full six years is there at a later time, however for this paper sufficient financial data was not available. It was deemed more than adequate to limit the experiment to 2015 due to its recency and it is in itself a comparable timeframe to many experiments detailed in the literature preceding it.

Many earlier experiments have found a correlation between network features and stock price while stock price is on a positive trend. In contrast 2015 offered an opportunity to explore network behaviour during a stable climbing price, a reasonably constant price, and a moderate depreciating price. However the range was not large, max price came to $133 while the minimum was $103.12. This max was the highest value the stock has achieved to date, a factor that could have brought with it some diverse network effects.

2015 was also a suitable first year for analysis as there were no confounding factors to influence price or market effects. For example in 2014 Apple (AAPL) stock underwent a split, dividing shares by 7 and equally dividing the price of each stock by 7.

AAPL $, 2012 - 2015

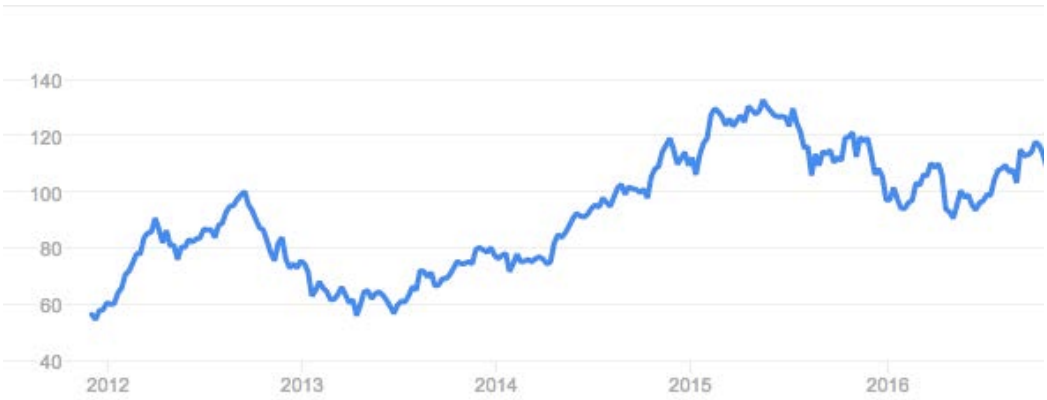

Figure 3-2*: Apple stock price 2012 – 2016*

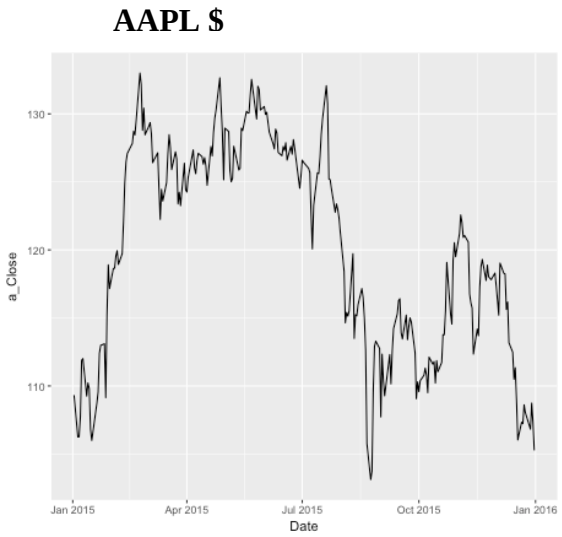

Figure 3-3*: Apple stock price during trading days in the analysis period 2015*

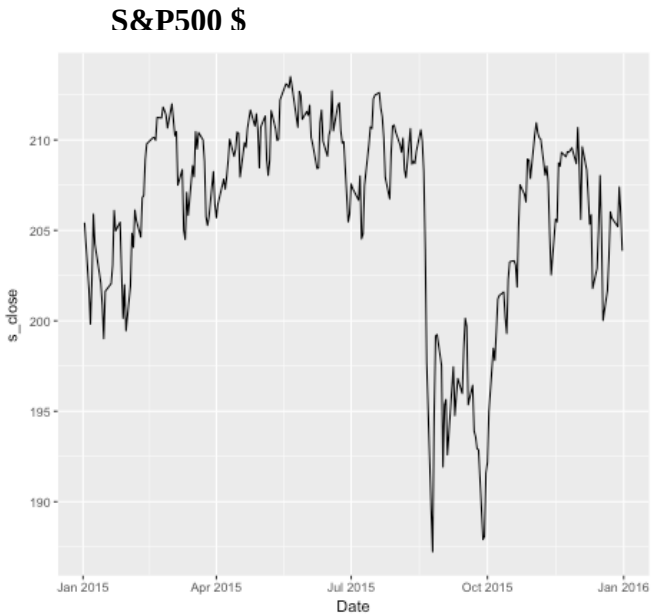

Figure 3-4*: S&P stock price during trading days of the analysis period 2015*

3.6 Data Preparation

Initially the network data was explored to identify features of relevance to the experiment and to generate a table from which data would be easily manipulated. The json was converted to csv format and all months were concatenated to enable observation of year long trends. The features extracted from the 27 features of the original dataset are listed below. There were no missing values among these.

Column
Id
Body
Username
TimePosted
inReplyTo
Sentiment
Followers
Following
statusesCount

Table 3-2*: list of features retained from original Stocktwits network dataset*

3.7 Social Network Data Manipulation

A ‘retweet’ network was generated in order to refine the dataset of tweets to tweets that were referencing another user. The majority or tweets on Stocktwits are posts that do not have a connection to another user, 82% in 2015. The hypothesis behind a retweet network was that any tweet referencing another user was implicit in information spreading. A message that references another user might be a retweet, that is a re-posting of another user’s message, a reply to a tweet or it might simply be a message to another user. While data is shared on Stocktwits to no person in particular it is difficult to measure information spread. However, in the retweet network, made up

exclusively of tweets that reference another user there is the direct goal of spreading information. It was further hypothesis that this might be an indication of interest or trust in another user. While online chat platforms and microblogs have been highlighted in the literature as very dense with noise these ideas were intended to refine the original dataset to a more useful and less noisy one.

In order to generate this all tweets that were either a retweet, in reply to another tweet or a message to another user were extracted. These could be identified by the presence of '*@name'* in the body of a tweet. In the case when it was a reply it also had the id of the original tweet it was in reply to in another in the 'inreplyto' column. If it was a direct message this field would be empty. A retweet was identified by 'RT *@name'* in the body of the tweet. For these the responder or retweeter in question, the username associated with that tweet, was called 'Retweeter' while the person referenced was entitled the 'Tweeter'.

The number of followers, following and tweets of the Retweeter were retained. The same fields for the tweeter had to identified through a search and a selection of that metadata on the same day or on the nearest date preceding that day. This search was carried out on the full dataset rather than the retweet dataset in order to ensure that the tweeter metadata for was retrieved or as many tweeters as possible. If that tweeter never referenced another uses themselves in 2015 their metadata wouldn't be in the retweet dataset. For some Tweeters the metadata was not retrieved. It is possible they never tweeter themselves in 2015 and it is also possible that some of the retweets were generated outside of the network. For example, if a user were to link their Stocktwits and their Twitter account, when they retweet a user on twitter that same message will appear on Stocktwits but the user referenced might not be a member of Stocktwits. 9% of tweeters or 8350 were not found and therefore their metadata remained as Null.

A second dataset with every tweet from the same year was also retained. The same features were generated for both and included in the final model. This enables comparisons later between behaviour on a retweet network and behaviour on the entire network with the intention of giving insights to the mechanisms underlying connectivity on online platforms.

For both Stocktwits datasets a unique id was generated for each tweet. The id from the network was retained to enable replies to be matched with the original tweet, however this new unique id was generated as the same tweet might be duplicated, if it referenced more than one user for example it was included twice alongside the metadata of each of the two tweeters referenced.

Tweet sentiment was also extracted from the original dataset. This is a label that tweeters may choose to add to their own post, either ‘bullish’ or ‘bearish’ to indicate the sentiment of the tweet. Finally, the date was retained from the original dataset. This enabled features to be calculated on a per day basis in order to facilitate comparison with the daily financial data.

Once condensed to this smaller manageable dataset of ten features and a new id, the body of the tweets was analysed to count the volume of tweets at each of the original thirty-one stocks under consideration. Of these it was only Apple that was referenced often enough in the retweet network to enable the experiment. Therefore, all tweets without a reference to ‘Apple’ or the ticker ‘$AAPL’ were excluded from further analysis. This brought the retweet network to close to 88,000 tweets and the entire network to 502,167.

3.8 Social Network Feature Generation

Once the retweet network of tweet nodes and edges had been determined further feature generation was necessary to enable a more thorough understanding of network intricacies. These were initially calculated at the tweet level and then an aggregate was generated to add to a ‘daily’ table alongside the financial data.

3.8.1 Text Analysis

The body of the tweet is the text. It is limited at 140 character and may also contain a link to an image or another website. This text can be considered unstructured data, there are a number of techniques commonly used in the literature that were applied to develop features to classify the content of the body.

While a sentiment feature was already included a second similar sentiment analysis was conducted counting all instances of the words ‘bull’, ‘bullish’, ‘bear’, ‘bearish’ in addition to the features bullish bearish labels. This was intended to enhance the inbuilt sentiment classifier and to mimic the work of (Bollen, May & Zeng, 2011) who performed the same sentiment classifier with a twitter dataset.

A final sentiment analysis was performed using the Loughran and McDonald Sentiment Word List (Loughran & McDonald, 2011). A count of the positive and negative words as they are labelled in this dictionary was recorded for each tweet. A further probability of the sentiment of these tweets was not performed due to the low number of matches. Often there was not a match, in fact only 27% of tweets in the retweet network had at least one matching word. It was very rare that more than one word would match or that words from both the negative and positive list would match in the same tweet. A classifier or probability estimation to determine the sentiment beyond that seemed futile and so these raw numbers were recorded as an indicator of positivity and negativity.

Following the sentiment analysis, a supervised topic classifier was generated to mimic the efforts of Racca, Casarin, Squazzoni & Dondio (2016) in generating ‘Technical’, ‘News’, ‘Spam’ and ‘Opinion’ classifications. To this end 1000 tweets were manually labeled from 1 – 4 to indicate whether it was a technical analysis, an opinion, news or spam respectively. From those categories word features were generated dependent on their importance / frequency in their respective dictionary. The remaining tweets were labelled by comparing their words with these dictionaries and by applying a naieve-bayes classifier to estimate the log probability of a tweet being in one topic and not the others. The topic with the highest log probability score was retained.

A second topic classifier was conducted using a Latent Dirilchlet Allocation (LDA) algorithm. In this instance the underlying mechanisms are very similar and once again the topic with the highest probability is taken. However, LDA is unsupervised and thus the topics were not initially manually generated. A corpus was generated of word instance in each tweet. From there the tweets were clustered into topics depending on how many topics were initially selected and the frequency of co-occurring words

throughout the dataset. Again a Bayesian probability statistic is computed labelling the probability of each tweet being in one of the clustered tweet ‘topics. These were tested against the 1000 manually labelled tweets and a misclassification rate was recorded. It was found that a low number of 2 or four topics performed very poorly. With 10 topics of more performance levelled off. Categorising each of the four topics performed poorly with LDA, with an accuracy rate close to 50% for each of the 10, 15 and 20 topic options. It was in distinguishing spam from opinion and news from technical analysis that caused the poor performance. Indeed these are very similar and with the short text of a tweet it is not straightforward to classify them manually. A ten topic LDA was selected, bundling 9 of the topics together into the opinion and spam categories and one topic into the technical analysis and news category. Once reduced to a two topic classification the performance rate of the LDA with two topics reached 80% accuracy (20% misclassification rate) as compared with the 1000 tweets been manually labelled.

3.8.2 Reputation

In generating the retweeter dataset reputation was already a factor that is was intended was being enhanced on average across tweeter. In addition to that Tweeters were classified as having a high or a lower reputation based on the number of follower, following and tweets to date they had. In order to generate this ranking system initially the maximum for each of these counts was calculated per day, e.g. the maximum followers of anyone tweeting per day, the maximum following and the maximum tweets to date. Once this dataset had been generated the followers, following and tweet count associated with each tweet was divided by the max that day and the three were summed. This number was divided by three to normalize the rank (a tweeter could at most score a 3 in the sum of the max followers, following and tweets if they had the maximum number of each on the day of their tweet).

(Following /Max Following + Followers / Max Followers + Tweet Count / Max TweeetCount) / 3

This generated a highly skewed dataset, with the vast bulk falling into the lowest rank.

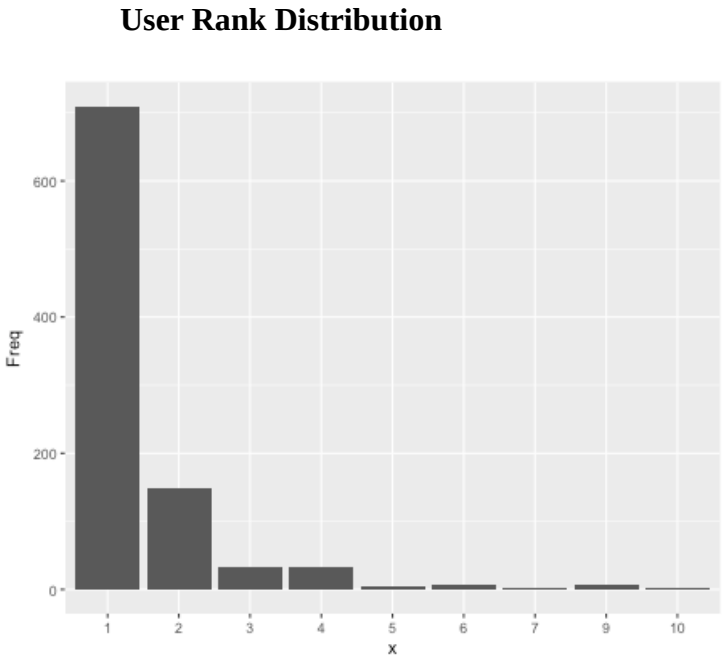

Figure 3-5*: histogram displaying the distribution of ranks over the retweet dataset*

80% of tweeter at the time of their tweet fell into the range of score between .1 and .2 while only 20% were above that. This was deemed reasonable due to the high reported noise on these platforms. A harsh ranking feature was of interest as a matter of experimentation. Scores that fell into the [.1 : .2] range were allotted 'Rank1' classification while the top 20% scoring between .4 and .10 were classified as 'Rank2'.

3.9 Financial Variables

The opening, closing, volume traded, volatility and historical volatility were included in the initial financial dataset. In addition to that the following financial variables were calculated:

1) the difference between Apple and the S&P500 a_open – s-close
2) the absolute difference between Apple and the S&P500 [a_open – s-close]
3) the intraday change in Apple price a_open – a-close
4) the absolute intraday change [a_open – a-close]
5) the proportional value of the change a_open – a-close / a_open
6) absolute proportional value of the change [a_open – a-close / a_open]

3.10 Daily Table

In order to generate a final model the financial data and the network data were added to the same daily table.

Initially the total number of users referenced in the retweet dataset, ‘Tweeters’ was aggregated to a daily sum. The total number of users in the full dataset was summed to a daily value. The total number of retweets in the retweet dataset were summed and the total number of tweets in the full dataset were summed.

For each of the sentiment, the intrinsic sentiment feature and the second one counting all instances of ‘bull(ish)’ and ‘bear(ish)’ use in the text, they were aggregated to a daily value by summing their respective occurrences e.g. the number of bullish tags or ‘bull(ish)’ word occurrences that day. Similarly for the McDonald and McLoughlan sentiment word classifier the positive words and negative words per day were aggregated.

Following from that proportions of tweets that were positive or bullish and negative or bearish for each measure were calculated for both the full dataset and the retweet dataset.

The same was done for the topic classifiers, the instance of tweets falling into each topic area was summed per day and added to the daily table under their respective topic headings. Their respective proportions were calculated by dividing each by the total number of retweets that day for the retweet dataset and total tweets per day for the full dataset.

With respect to the reputation table those falling into either rank was summed. Following a unique count of users referenced, ‘Tweeters’ in each rank (e.g. if a user of Rank1 posted 10 times in 1 day that user was only counted once) was divided by the number of ‘Tweeters’ that day. For those ranked in the full dataset, the number per rank was divided by the total number of users that day.

3.11 Social Network Analysis

Three features were generated at the daily level in order to measure the retweet network behavior cohesively in terms of social network theory. The in-degree centrality was measured, the assortativity and the modularity using the network-x library in python. The in-degree centrality measures the inward links, in this case the tweeters in the retweet network per node. The assortativity is the degree to which nodes in the network connect to other nodes who tweet with similar in-degree centrality. The network modularity is a measure of the strength of the division of the network into smaller modules. With respect to the in-degree centrality the Gini-coefficient of in-degree centralities per day was measured to generate a normalized index between 0 and 1 every day of the relative connectivity of users. Below are examples of daily networks of in-degree centrality, displaying the variety in shape and the relative convergence on very densely connected nodes some days and a more distributed network others.

Daily Social Networks

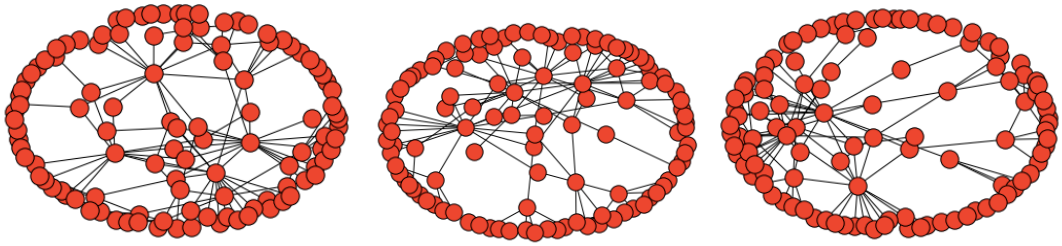

Figure 3-6*: Three daily netowrks display the most connected users in the centre surrounded by users with just one connection*

3.12 Final Dataset

In summary, combining the two network datasets and the financial data, a final daily dataset was generated. This contained 56 features including 17 financial features, 22 text classification features, 8 reputation features, 6 network features and three social

network features, the date and a unique id covering both the retweet network and the full dataset of 2015 tweets.

Feature	Group	Description
ID		Unique id per day
Date		Date
A_open	Finance	Apple opening price
A_close	Finance	Apple closing price
A_change	Finance	Change in Apple Price
A_volume	Finance	Volume of Apple traded
A_histvol	Finance	Measure of Apple 10 day historical volatility
A_realvol	Finance	Intraday volatility
A_histvol020	Finance	Measure of Apple 20 day historical volatility
A_histvol060	Finance	Measure of Apple 60 day historical volatility
A_histvol120	Finance	Measure of Apple 120 day historical volatility
A_histvol252	Finance	Measure of Apple 252 day historical volatility
Abs_a_change	Finance	Absolute change in apple price between open and close
Abs_a_changeP	Finance	Absolute change in apple proportional change between open and close
A_changeP	Finance	Proportional change in apple price between open and close
S_open	Finance	S&P500 opening price
S_close	Finance	S&P500 closing price
asdiff	Finance	Difference between apple

		and S&P closing price
Abs_asdiff	Finance	Absolute difference between apple and S&P closing price
Bullish	Text Classification	Count of bullish tags in retweet network
Bearish	Text Classification	Count of bearish tags in retweet network
Bull_RT	Text Classification	Proportion of retweets with bullish tag
Bear_RT	Text Classification	Proportion of retweets with bearish tag
BullRT	Text Classification	Count of all 'bull(ish)' words in text and tags in retweet dataset
BearRT	Text Classification	Count of all 'bear(ish)' words in text and tags in retweet dataset
BullT	Text Classification	Count of all bullish tags in full dataset
BearT	Text Classification	Count of all bearish tags in full dataset
Bull_T	Text Classification	Proportion of tweets with Bullish tag in full dataset
Bear_T	Text Classification	Proportion of tweets with Bearish tag in full dataset
McLpos	Text Classification	Number of positive words from the McDonand and Loughran dictionary in retweets per day
McLneg	Text Classification	Number of negative words from the McDonand and

		Loughran dictionary in retweets per day
McLposT	Text Classification	Number of positive words from the McDonand and Loughran dictionary in all tweets per day
McLnegT	Text Classification	Number of negative words from the McDonand and Loughran dictionary in all tweets per day
Opinion	Text Classification	Number of opinion & spam classified retweets per day
Technical	Text Classification	Number of technical & news classified retweets per day
OpinionT	Text Classification	Number of opinion & spam classified tweets per day
TechnicalT	Text Classification	Number of technical & news classified tweets per day
Opintion_RT	Text Classification	Proportion of opinion & spam classified retweets per day
Technical_RT	Text Classification	Proportion of technical & news classified retweets per day
Opinion_T	Text Classification	Proportion of opinion & spam classified tweets per day
Technical_T	Text Classification	Proportion of technical & news classified tweets

		per day
Users	Network Activity	Count of all unique users posting per day in full dataset
RtUsers	Network Activity	Count of all unique tweeters in retweet network per day
RtUsers_Users	Network Activity	Proportion of unique users retweeted per day
Tweets	Network Activity	Total Tweets per day
Retweets	Network Activity	Total Retweets per day
RTP	Network Activity	Proportion of tweets that are retweets
NewRank1	Reputation	Number of tweeters in the retweeted network in rank 1 per day
NewRank2	Reputation	Number of tweeters in the retweeted network in rank 2 per day
NewRankT1	Reputation	Number of tweeters in the full dataset in rank 1 per day
NewRankT2	Reputation	Number of tweeters in the full dataset in rank 2 per day
NewRank1_RT	Reputation	Proportion of tweeters in the retweeted network in rank 1 per day
NewRank2_RT	Reputation	Proportion of tweeters in the retweeted network in rank 2 per day
NewRankT1_T	Reputation	Proportion of tweeters in the full dataset network

		in rank 1 per day
NewRankT2_T	Reputation	Proportion of tweeters in the full dataset network in rank 2 per day
InDegreeCentrality	Social Network Analysis	Gini coefficient of the daily in-degree centrality
Assortativity	Social Network Analysis	
Modularity	Social Network Analysis	

Table 3-3*: Final dataset generated from network variables, features generated and financial variables*

3.13 Data Modelling

Once the final daily dataset was compiled a model appropriate to measure for a correlation between the network features and the financial features was required.

Once aggregated to daily values both the financial and network variables were continuous. A multivariate regression and regression tree were considered appropriate options. As detailed later in the results the data did not meet the assumptions of linear regression. Both logistic regression and a regression tree were considered. On testing a regression tree fit the baseline models better, it achieved a lower error rate when compared to test data.

A regression tree uses recursive partitioning to create a tree where each node represents a partition. The mean squared prediction error is the criteria for choosing the partitions in the model. The nodes at the top contribute most to explaining the variance in the model. The leaf nodes, the final nodes are the ones beyond which splitting the data does not explain enough of the variance to be relevant in describing Y. For each leaf node and training sample the model for a regression tree is the following.

$$\hat{y} = \frac{1}{c}\sum_{c=1}^{c} = Y1$$

In this case many decision boundaries are distinguished to determine the relationship between the network and financial variables, whereas a logistic regression often works better if there is just one. A regression tree is simpler to interpret. With the goal of this experiment to distinguish features with robust information and of value for learning rather than a prediction, the interpretation of a regression tree is the most fitting choice.

Before a model was set up a lag was required between the financial and network data. The most optimistic goal of the regression was to test whether features of the network today could inform prediction of the stock price tomorrow. For that reason, the financial data was altered to shift back one day for the analysis.

3.14 Model Selection

Initially a pairwise correlation using the Pearson Correlation was carried out in order to explore the data and look for a relationship between the dependent financial variables and the independent network variables. Furthermore, the variables of particular interest were plotted in scatter plots with a line of best fit to visualize the relationship. Line plots were also generated to view financial and network trends over the year.

Scatterplots of Pairwise Correlations

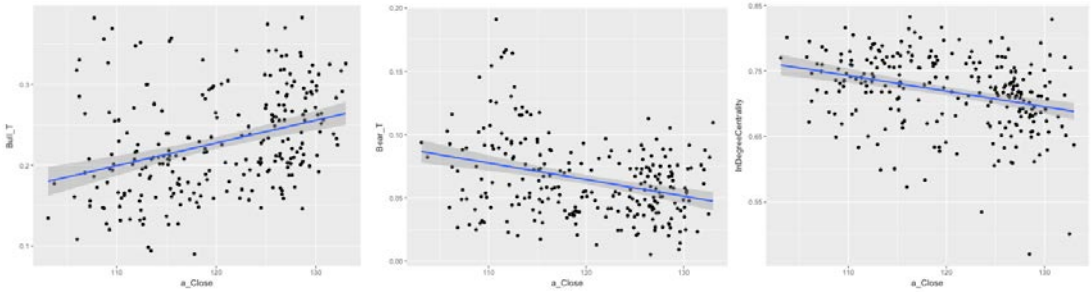

Figure 3-3-7*: scatterplots displaying pairwise correlations of netowrk variables with AAPL closing price*

Following that regression trees were generated. Baseline models and six experimental models were created to explore the relationship between the network data and the three financial variables; AAPL price, volume traded and daily volatility. For each, there

was a tree to predict tomorrow's market and one to correlate within day features. In addition to that four network only trees were modelled to explore the network behavior with respect to the generated features Rank and text classification into technical and opinion based tweets.

3.15 Model Performance Measurement

Initially the data is split into training and test sets. A model is built on the training set and it's accuracy and generalizability are tested by comparing it's predictions to the test set.

The root mean squared error (RMSE) is the main measure applied to model evaluation. This measures gives a greater weight to larger residuals than smaller ones, e.g. it penalizes large deviations in predictions from observed values. The values are in the same units as the target value making error interpretations simple.

$$RMSE = \sqrt{\sum_{i=1}^{n} (ti\text{-}M(di))\text{^}2 \,/n}$$

The mean squared error is also applied as a secondary metric of model performance. The MAE gives and equal weight to all residuals. Put simply it measures the magnitude of the error from model predictions in comparison with observed values.

$$MAE = \sum_{i=1}^{n} (ti\text{-}M(di))/n$$

Both fall into the range [0, ∞] and smaller values indicate better model performance.

In addition to that the RSE and MAE are compared to those of the baseline models to add further depth to the understanding of the correlation strength of the network and financial variables.

4. IMPLEMENTATION & RESULTS

In this chapter the results to the above described experiment will be presented along with an evaluation of the methodology.

- The tools and practical steps to approaching the data and modelling are initially outlined
- Following on from that is a section on data exploration and relevant visualisations
- Regression trees for the 3 baseline models predicting next day AAPL price, volatility and volume are depicted and their respective implications detailed
- Regression trees for the 6 network predictor trees of the same financial variables are presented
- Finally, are 7 network exploitative trees

4.1 Tools

In order to carry out the piece of analysis described the software required was all open source and computationally light. It was possible to run everything from a standard laptop with the additional storage of an external hard drive.

Once the initial raw network dataset had been refined to the few columns that were included in this experiment that data was transferred to a MySQL database. There, a set of tables were designed in a relational database to capture all the data from the retweet network, the full Stocktwits dataset and the financial data. Features generated thereafter were added to their respective table.

Data manipulation and feature generation was performed using python and the Pandas library. The final model to test the correlation and the subsequent model evaluation were carried out in R.

4.2 Experiment Implementation

- Only the daily table was used for modelling once all features were generated and added to it from both networks datasets and the financial data
- Data exploration was carried out generating a base table of max and min values, range and standard deviation of all features. Following that pariwse scatterplots were generated to explore relationships between the network and financial features. These are depicted in the methods section. Line plots over the year were also generated and are depicted below.
- The data was split into training and test sets
- Regression trees were trained on the training dataset and their performance tested on the test-set. Three baseline trees were generated to compare the best prediction models or AAPL next day price, volatility and volume to later network models. 6 regression trees were trained to generate network prediction and same day correlation models with the referenced financial variables as the dependent variables.
- 7 More network only trees were generated to explore in network relationships. These were carried out with retweet and the full network dependent features of rank, topic classification and SNA metrics.
- Trees were pruned to generate the smallest tree with the maximal performance
- Models were tested against the test set and the root mean squared error and mean squared error were calculated to rate performance against the baseline models.

4.3 Data Exploration

AAPL did not meat the criteria for a parametric linear regression as the data is not normally distributed. The closing price and volatility have a bimodal shape and volume is positively skewed. In addition to that residuals are not normally distributed and outliers are prominent in data. Outliers could not be omitted as deviations were considered valuable for analysing relationships.

Histograms of AAPL Financial Variables

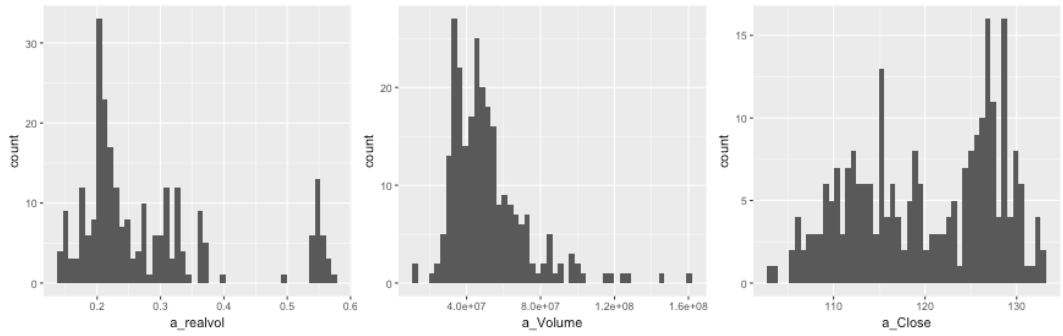

Figure 4-1 ***distribution plots of the three financiall variables under review***

The data was plotted to explore potential trends initially. Pairwise scatterplots were constructed to display correlations between any two variables. In addition to that line plots were generated to display trends over the experimental period and highlight any potential relationships between outliers.

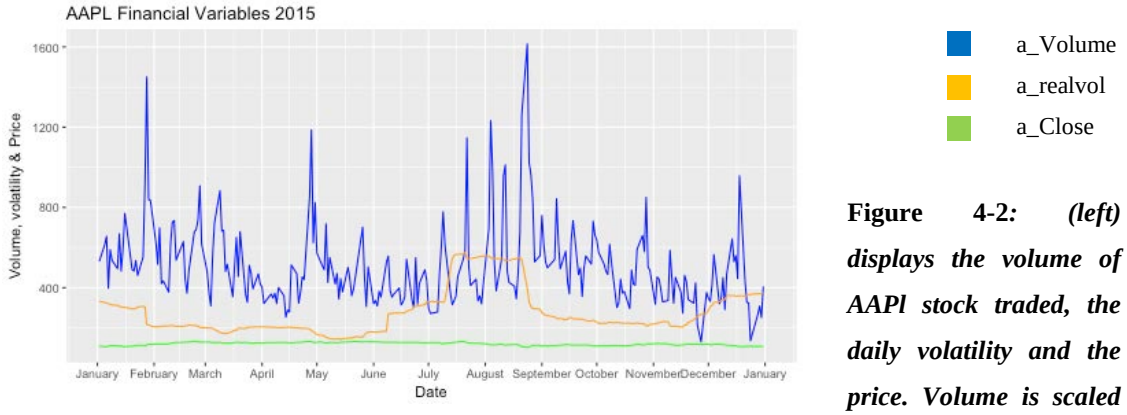

Figure 4-2*: (left) displays the volume of AAPl stock traded, the daily volatility and the price. Volume is scaled down (*1/100000) and volatility scaled up (*100) to display relative trends on one plot.*

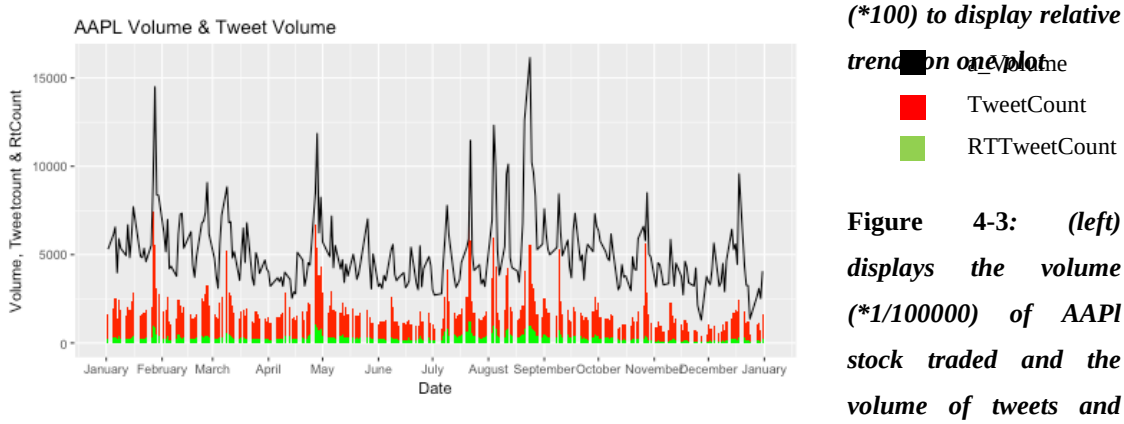

Figure 4-3*: (left) displays the volume (*1/100000) of AAPl stock traded and the volume of tweets and retweets posted.*

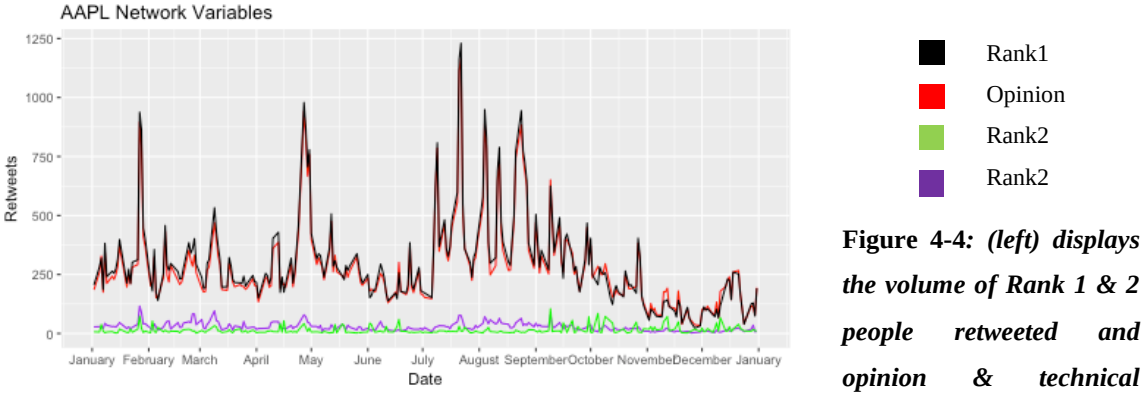

Figure 4-4*: (left) displays the volume of Rank 1 & 2 people retweeted and opinion & technical retweets.*

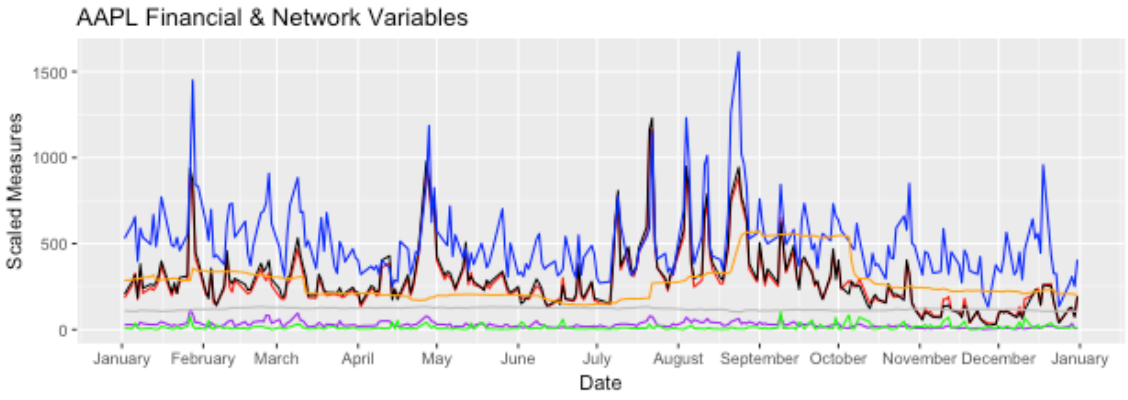

Figure4-5*: (above) displays the retweet network variables; volume of Rank 1 users retweeted, Rank 2 users retweeted, opinion retweets & technical retweets along with the three financial variables; volume (* 1/100000), volatility (*100) & price.*

4.4 Model Preparation

The data was split into training and test sets at a random 80:20 split. There are 252 trading days in the dataset. The markets are shut on weekends and holidays. A larger

test set would have been desirable however due to the scarcity of trading days a larger portion, 80% of the data was allocated to the training set. There is a risk of overfitting with a learning model such a regression tree, a 20% test set is intended to enable a measure of the models strength and generalisability. For reproducibility the seed is set and the test set is generated.

Regression trees were modelled to distinguish the network features with the strongest relationship with the AAPL financial features; price, daily volatility and volume.

Baseline models were constructed including all financial and network variables. These were considered the best potential prediction models using all available information and used as a comparison for the utility of the network.

Experimental models omitted financial features from the analysis. These were dominant in prediction and a more thorough network analysis was pursued. Trees were modelled with same day and next day AAPL price, volume and volatility as the root nodes. Same day correlations were not performed for prediction, network would remain active after market closing, instead these are intended as a retroactive exploration of correlation on the day. For next day models the predictions against the test set are included in tables. Networks features with the strongest correlations with market trends are considered trustworthy and robust information identifiers.

In addition to the models correlating the network and the market, regression trees are modelled with the network variables as root nodes; proportion of Rank 2 users retweeted, proportion of Rank 2 users tweeting, proportion of retweets that are technical and the proportion of all tweets that are technical. This was carried out to better identify network dynamics. Not all network variables were included in each model, depending on their particular relevance and due to concerns for overfitting.
All trees are pruned after modelling to prevent overfitting and to reduce the tree branch number without a loss of accuracy as measured against the test set. A smaller tree with fewer splits requires fewer decision rules and a tree that is easier to interpret.

4.5 Baseline Model

Baseline prediction models were constructed by including all financial variables and network variables into the models and establishing the tree with the lowest root mean squared error (RMSE) and mean absolute error (MAE) as compared to the training set. Next day AAPL price, volatility and volume are the root nodes for the three baseline models.

4.5.1 Baseline for predicting AAPL closing price

Only AAPL's (AAPL) closing price yesterday is included in the pruned for predicting closing price tomorrow, *(RMSE: 2.656, MAE: 2.146).*

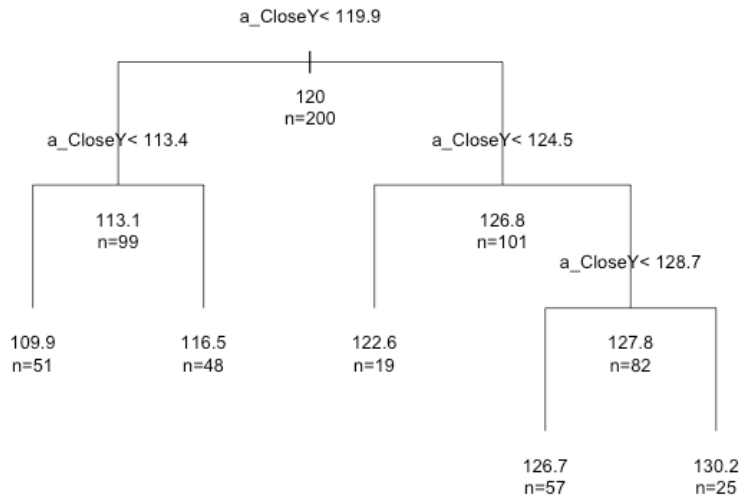

Figure 4-3-8*: Baseline tree for AAPL closing price tomorrow*

4.5.2 Baseline model for predicting AAPL volatility

Similarly the best performing tree for predicting volatility splits the data solely on volatility the prior day *(RMSE: .0358, MAE: .0216).*

.

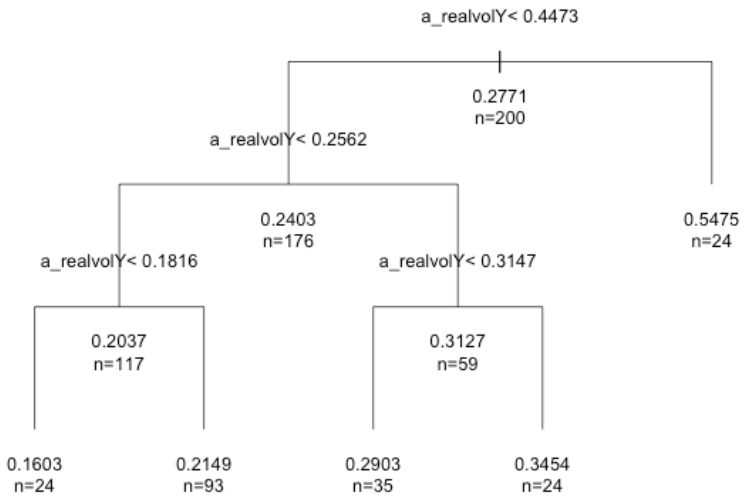

Figure 4-3-9*: Baseline tree for AAPL volatility tomorrow*

4.5.3 Baseline model for predicting AAPL Volume

In contrast to volatility and price, the only variable in the final tree predicting the volume of stock traded tomorrow is a network variable, the number of Rank 1 users in the full network (*RMSE: 22411491, MAE: 14370463*).

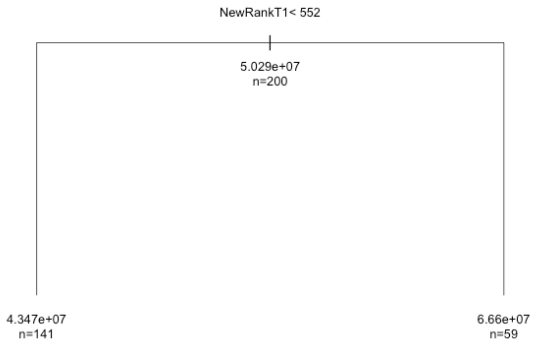

Figure 4-3-10*: Baseline tree for AAPL volume traded tomorrow*

4.6 Network Models

4.6.1 Price Prediction

A regression tree was constructed initially to test which features in the network on any given day have the greatest relevance in predicting price at close tomorrow. The first

model converged on the 12 features. A second pruned tree reduced the nodes to four *(RMSE:6.8, MAE: 5.29).* On average the predicted price is 5.29 off the observed price.

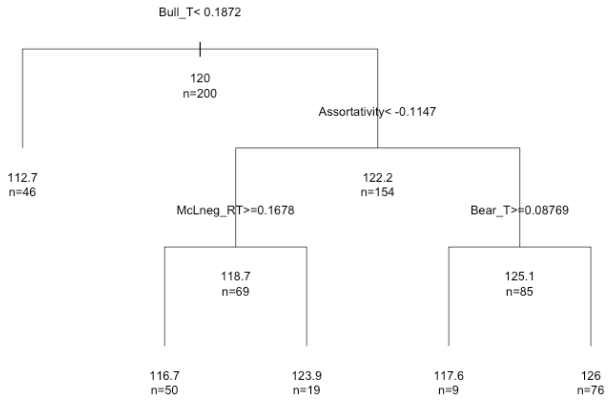

Figure 4-3-11*: Network Prediction of AAPL closing price*

- The proportion of tweets with a bullish tag in the full network below .18 predict a price of 122.7.
- Above that and if the assortativity is below 1.1147 the data is split again on the proportion of negative words in the retweet network as measured by the McDonald and Loughran lexicon.
- If those are above .168 and the other criteria are met the price will be 116.7, if they are below it will be 123.9.
- If assortativity is higher than -.1147 and the proportion of bearish tweets in the full network are greater than .088 the price is predicted to be 117.6
- If the bearish tweet proportion fall below the price is predicted to be it's highest in this model 126.

Below is a table of the model's predicted prices and the observed prices in the test set.

Day	Prediction	AAPL
2	116.7068	106.25
5	123.8632	106.26
7	116.7068	111.89
21	116.7068	112.4
26	116.7068	109.14

27	116.7068	115.31
33	116.7068	118.65
35	123.8632	119.94
36	125.9836	118.93
42	125.9836	126.46
44	125.9836	127.83
49	116.7068	128.45
55	117.5633	128.79
58	125.9836	129.09
65	125.9836	127.14
69	125.9836	122.24
70	125.9836	124.45
72	125.9836	124.95
91	116.7068	125.32
97	123.8632	125.6
100	125.9836	126.85
138	125.9836	130.07
139	125.9836	130.06
154	125.9836	129.36
155	125.9836	128.65
163	117.5633	126.92
175	125.9836	127.5
187	116.7068	125.69
188	116.7068	122.57
191	125.9836	125.66
196	123.8632	128.51
197	123.8632	129.62
201	125.9836	130.75
212	112.7061	118.44
216	112.7061	115.4
233	112.7061	103.12
258	116.7068	116.41
264	125.9836	113.4
265	112.7061	114.32
272	112.7061	110.3
292	125.9836	113.77
294	125.9836	115.5
300	116.7068	119.27
313	123.8632	116.77
317	112.7061	114.18
323	116.7068	119.3
327	112.7061	118.88
341	123.8632	118.23
342	112.7061	115.62
344	116.7068	113.18
363	125.9836	107.32

Table 4-1 ***price prediction v's observed values***

4.6.2 Same day Price and network correlation

The first tree of 12 nodes (RMSE: 6.69, MAE: 5.37) was pruned to two nodes with a marginal improvement in error (RMSE: 6.56, MAE: 5.27).

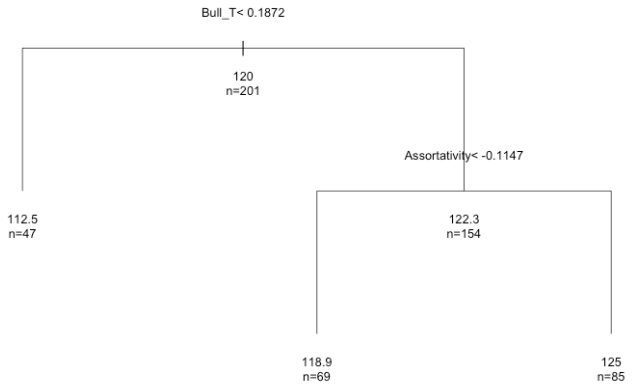

Figure 4-3-12*: same day network and price correlation*

The same two features, Bullish tags in the full network and the assortativity of the retweet network, have the highest value in the prediction model as we see here. The RMSE and MAE are almost equal with only a marginal improvement with same day correlation.

4.6.3 Volatility Prediction

A second pruned tree performed better for volatility prediction (*RMSE: .092, MAE: .064).* For reference volatility has a range of [.14, .57] during the experiment trading period. The volatility predications deviate on average by .57 from the observed values.

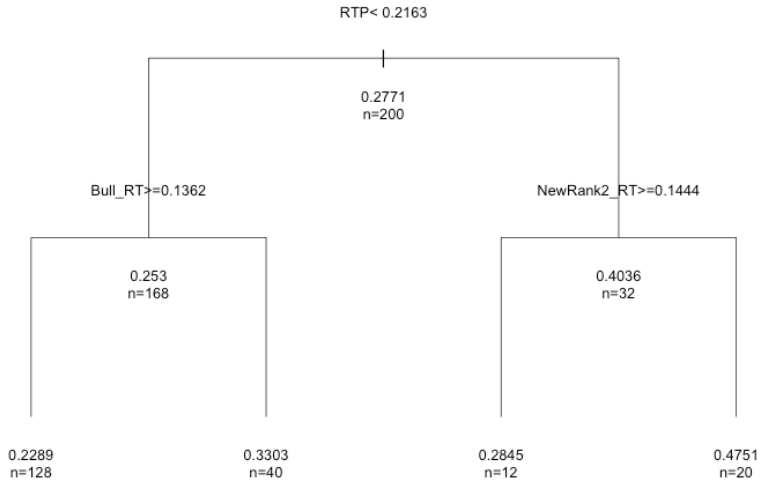

Figure 4-3-13*: next day volatility prediction*

The retweets played an important role in the volatility prediction. The proportion of retweets in the retweet network are the most important feature.

- If the proportion of all tweets that are retweets is above .216 the data is split on the proportion of higher ranked (rank2) tweeters in the retweet network.
- If the proportion of rank 2 tweeters in the retweet network is above .14, volatility is measured to be .28. If there are fewer it is estimated to be higher at .47
- If the proportion of retweets in the network is less than .216 the data is split again on the proportion of bullish tweets in the retweet network.
- If those are above .136 the volatility is estimated to be .229. and if they are below volatility is estimated to be higher at .33.

Day	Prediction	Volatility
2	0.2288962	0.326988
5	0.2288962	0.322654
7	0.2288962	0.317044
21	0.2288962	0.296538
26	0.2288962	0.305291
27	0.2288962	0.21656
33	0.2288962	0.206466

35	0.2288962	0.207809
36	0.2288962	0.209127
42	0.2288962	0.208794
44	0.2288962	0.207286
49	0.2288962	0.210996
55	0.2288962	0.208522
58	0.2288962	0.199534
65	0.2288962	0.177109
69	0.2288962	0.173836
70	0.2288962	0.170439
72	0.2288962	0.183402
91	0.2288962	0.204556
97	0.2288962	0.202889
100	0.2288962	0.201783
138	0.2288962	0.150521
139	0.2288962	0.149084
154	0.2288962	0.178755
155	0.2288962	0.183188
163	0.2288962	0.273667
175	0.4750942	0.308186
187	0.2288962	0.329913
188	0.4750942	0.329385
191	0.2288962	0.496357
196	0.2288962	0.565953
197	0.4750942	0.566441
201	0.4750942	0.575302
212	0.4750942	0.555246
216	0.3303081	0.544288
233	0.3303081	0.39816
258	0.2845203	0.252344
264	0.2845203	0.242062
265	0.3303081	0.243466

272	0.2288962	0.234063
292	0.2288962	0.220449
294	0.2288962	0.224891
300	0.2288962	0.208727
313	0.2288962	0.210662
317	0.3303081	0.204045
323	0.2288962	0.226547
327	0.3303081	0.249543
341	0.3303081	0.337958
342	0.3303081	0.337236
344	0.3303081	0.362029
363	0.3303081	0.371624

Table 4-2*: volatility prediction versus observed values*

4.6.4 Same day volatility and network correlation

The First unpruned tree performed best for same day correlation *(RMSE: 0.09, MAE: 0.06).* RMSE and MAE were only marginally better than the prediction model

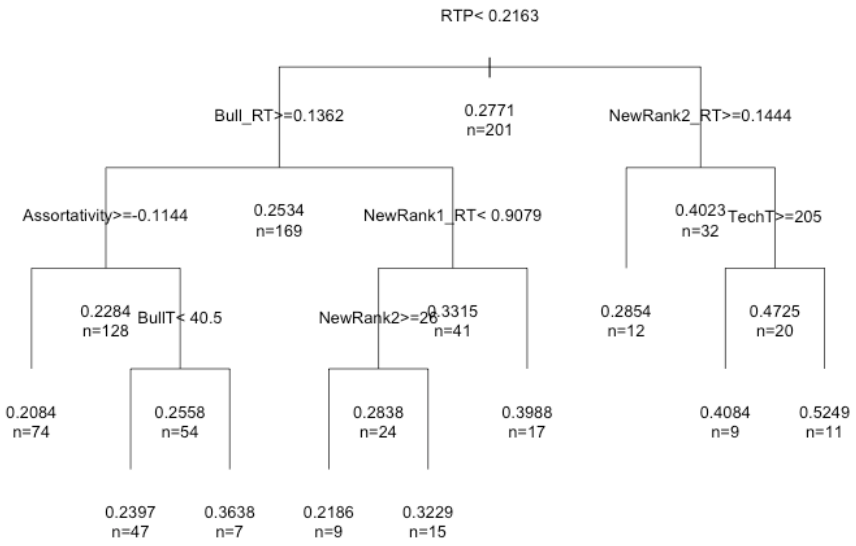

Figure 4-13*: same day network and volatility correlation*

Here again the retweet network proved to be an important feature. The same three features have the highest correlation with volatility, namely the proportion of retweets, the proportion of retweeters in rank two and the proportion of bullish tags in the retweet network. Fewer technical tweets, more Rank 1 users retweeted, lower assortativity and more bullish tags in the full network correlate with higher volatility as measured the same day.

4.6.4 Volume Prediction

In a prediction model for the volume of AAPL the second pruned tree of just one node was the most powerful, *(RMSE: 22411491, MAE: 14370463).* This is the same tree as in the baseline model, indicating the correlation between more Rank 1 users in the full network and volume of stock traded the following day. For reference the range of volume traded over the experimental period is [161454200, 13023700].

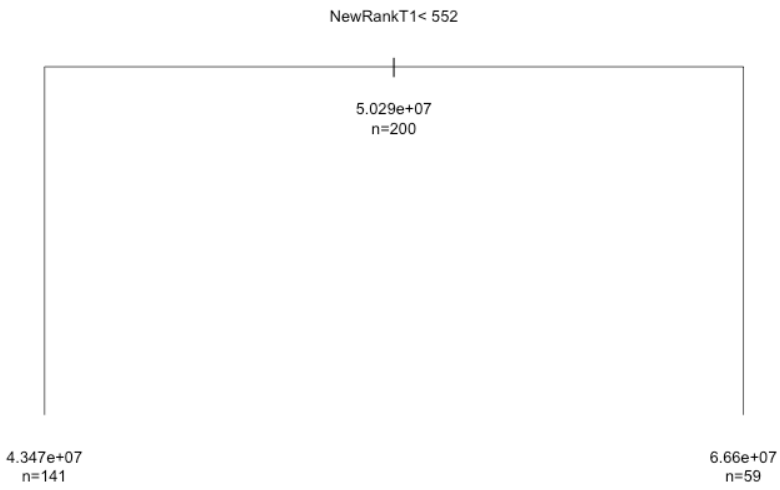

Figure 4-3-14 ***: next day volume trading prediction***

Day	Prediction	Volume
2	43470627	61366300
5	43470627	65618800
7	43470627	58946800

21	43470627	53570700
26	66603915	91929200
27	66603915	145000000
33	43470627	51640300
35	66603915	42072200
36	43470627	43400000
42	66603915	73418200
44	66603915	62972700
49	43470627	37241100
55	66603915	73855500
58	66603915	47949900
65	66603915	88347500
69	66603915	68582700
70	66603915	48145700
72	43470627	35507400
91	43470627	32120700
97	43470627	37193100
100	66603915	35964400
138	66603915	44351200
139	43470627	36000000
154	43470627	38229300
155	43470627	35314200
163	43470627	39842600
175	66603915	31816700
187	43470627	46716100
188	43470627	60490200
191	66603915	37237800
196	43470627	35866800
197	43470627	45693300
201	66603915	58898800

212	66603915	69500000
216	66603915	99153800
233	66603915	161454200
258	43470627	36900000
264	43470627	49800000
265	43470627	35645700
272	66603915	66100000
292	43470627	48778800
294	43470627	41272700
300	66603915	85023300
313	43470627	58635100
317	43470627	37700000
323	43470627	34103500
327	43470627	42426900
341	43470627	34254500
342	43470627	45017700
344	43470627	46640500
363	43470627	25110600

Table 4-3*: Volume predicted versus observed values*

4.6.5 Same day correlation

For a correlation of the variables on the same day the unpruned tree performed best *(RMSE: 15709314, MAE: 10103517).*

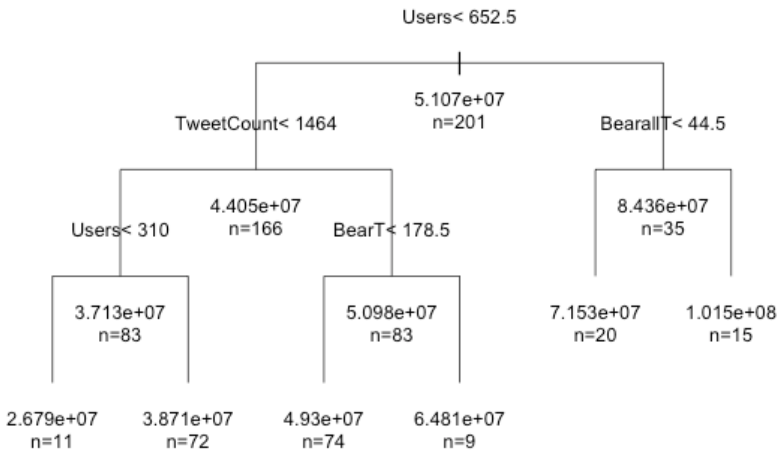

Figure 4-3-15 ***: same day network and volume correlation***

In the same day tree there are entirely new features correlated with volume traded relative to the prediction model. A greater number of unique users in the full network, a greater count of bear(ish) words tweeted are correlated with the highest level of volume. More tweets in the full network, more bearish tags and again more users split at a lower level are correlated with higher volume.

4.7 Prediction of network variables

Models for the prediction of some few key network variables are generated here in order to give a better insight into the hypotheses proposed regarding high ranked users and technical content.

4.7.1 Proportion of technical tweets in the retweet network

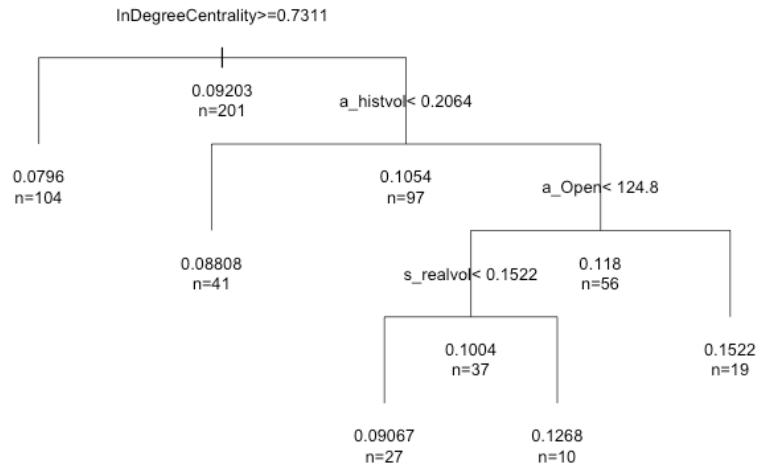

Figure 4-3-16*: correlation with proportion of technical retweets*

A lower measure of the network dispersion (Gini index of the in-degree centrality) and higher volatility over a 10 day period were the two most important variables respectively. A higher stock price and higher in-day volatility of the S&P market are also included *(RMSE: 0.04, MAE 0.03).*

4.7.2 The proportion of technical tweets in the full network

The number of tweets dominated the model to predict daily technical tweets. They were inversely related, fewer tweets increase the proportion of technical tweets or put another way when the tweet number goes up it is more opinion based tweets while the number of technical tweets remains relatively constant (*RMSE: 0.027, MAE: 0.022*).

In a second model omitting tweet count RMSE and MAE are almost unchanged. Lower ranked users are negatively correlated with technical tweets. Fewer bullish and bearish tags and a lower closing value on the S&P are also associated with a greater proportion of technical tweets (*RMSE: 0.028, MAE: 0.021*).

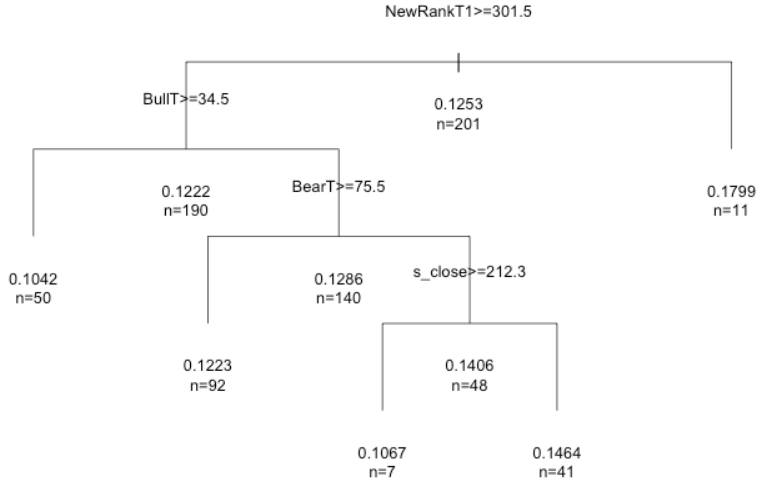

Figure 4-3-17*: correlation with proportion of technical tweets*

4.7.3 The proportion of retweets that are Rank 2 users

The number of users in the full network is the most important variable in the smallest tree for predicting the proportion of higher ranked users retweeted. Fewer users are correlated with with more rank 2 users retweeted (*RMSE: 0.1, MAE: 0.08*).

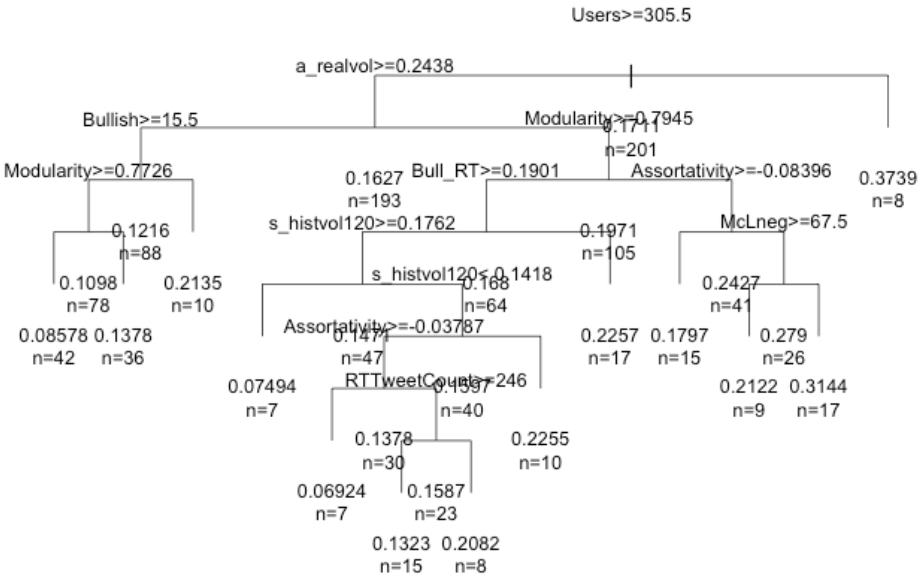

Figure 4-3-18*: correlation with the proportion of retweets that are Rank 2 users*

Omitting user number from a second model the other network features become apparent. Lower volatility of AAPL, lower assortativity in the retweet network, fewer users retweeted and a higher proportion of technical tweets are the strongest predictors a greater proportion of higher ranked users retweeted *(RMSE: 0.09, MAE: 0.06).*

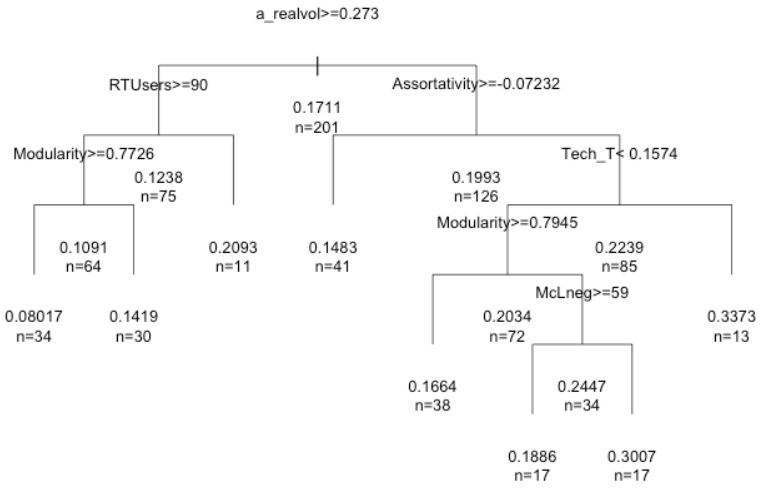

Figure 4-3-19*: correlation with the proportion of retweets that are Rank 2 users*

4.7.4 The proportion of tweets in the full network posted by Rank 2 users

Fewer tweets, fewer retweets and lower S&P volatility are correlated with a higer proportion of rank two users tweeting in the whole network *(RMSE: 0.01, MAE:0.007).*

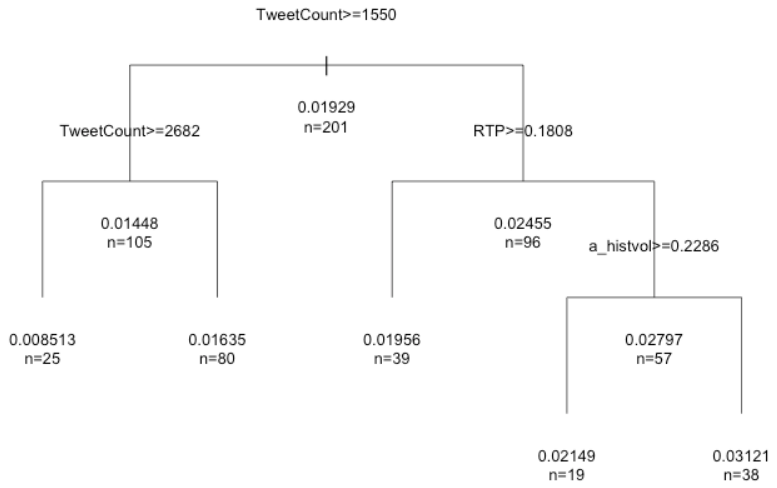

Figure 4-3-20*: correlation with the proportion of tweets posted by Rank 2 users*

4.7.5 Gini Index Score of the daily In-Degree Centralities

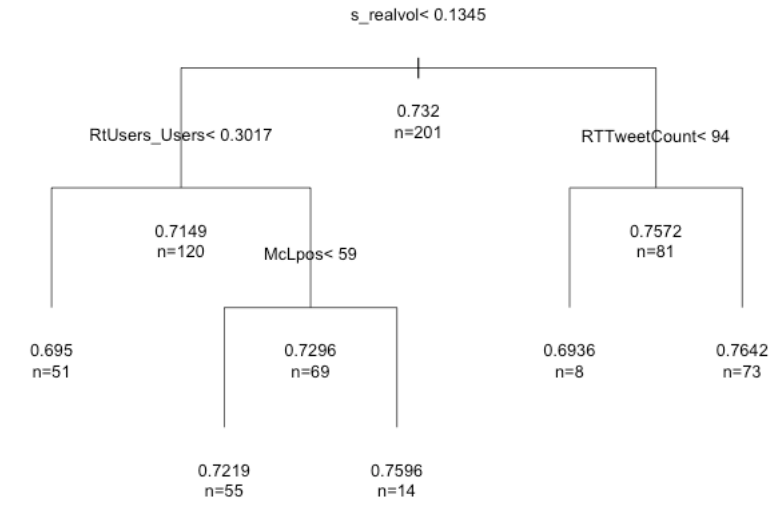

Figure 4-3-21*: Network features correlation with the Gini index score of the in-degree centralities*

S&P500 daily volatility is split first in the tree to predict the daily Gini index score for the in-degree centralities. Higher S&P500 volatility, a higher retweet count, retweet users and positive words from the McDonald and Loughran dictionary are all correlated with a higher score or a less equal network of in-degree centralities (*RMSE: 0.04, MAE*:0.03).

4.7.6 Retweet Network Modularity

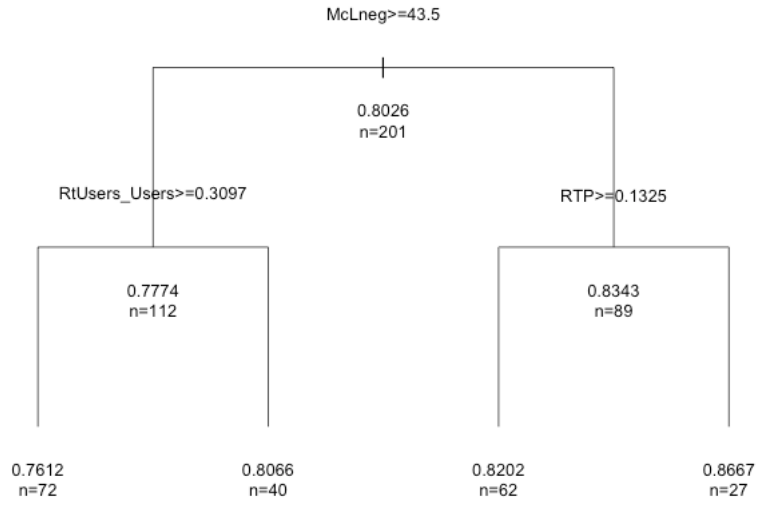

Figure 4-22*: network features in correlation with the retweet network modularity*

Fewer negative words from the McDonald and Loughran dictionary, and a smaller proportion of retweets in the network are correlated with higher modularity *(RMSE:0.04 ,MAE:0.03).*

4.7.6 Retweet Network Assortativity

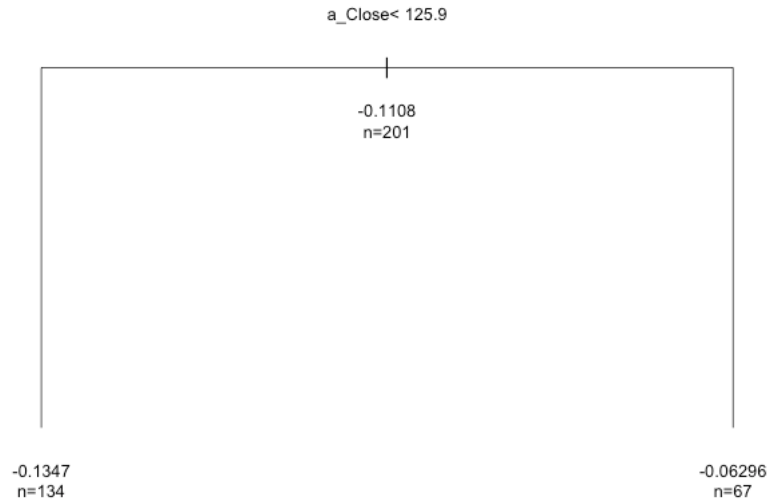

Figure 4-23*: network features in correlation with the retweet network Assortativity*

AAPL closing price is the only feature split in the regression tree with dependent variable assortativity of the retweet network. A higher price is associated with users connecting with others with similar in-degree centrality as themselves.

4.8 Summary of Results

- The baseline models for next day closing price and volatility are best predicted by todays' respective values
- The baseline model for next day volume traded is best predicted by the network variables number Rank 1 users tweeting.
- For the network model predicting next day closing price the bullish sentiment tag in the full network performed best, followed by assortativity in the retweet network. Three sentiment features were included in the model.
- Network next day volatility prediction is attributed to retweet behaviour
- Next day volume prediction mirrored the baseline model
- Same day correlation trees offered little improvement in strength of relationships

5. EVALUATION / ANALYSIS

This section is a review of the strength research conducted and the results.

- Results are evaluated and their implications are elaborated upon. This is done from an view over the entire experiment and followed by amore focused evaluation of the features under scrutiny.
- The significance of the results are then outlined with respect to the literature
- The final two section focus upon the strengths of the experiment and the findings and the weaknesses therein

5.1 Evaluation of Results

The primary strength of the results is the selection of features in the network which most closely reflect market changes. With respect to the prediction models initially, the network proved the more prominent than financial variables in predicting next day volume traded. The number of Rank 1 users in the full network was the only feature included in the baseline model.

The baseline models performed far better with lower RMSE and MAE for price and volatility prediction, in both cases yesterday's respective value was best for predicting tomorrows. We can see for the line plots that volume varies hugely over the experimental period whereas price and volatility are reasonably consistent. With little variation there is little to search for externally. Volume offered a greater opportunity for network correlation investigation.

With respect to a network exploration all three models have offered an insight to network dynamics and how information is spread on a social platform.

5.1.1 Retweet Network

The retweet network featured most prominently in the volatility models where features measuring retweet behaviour alone were selected for the prediction model. The model

is not strong in prediction relative to yesterdays volatility alone but it displays the propensity to retweet in correlation with next day volatility. The proportion of tweets that are retweets has the strongest relationship with next day volatility. In addition to that a lower proportion of rank two users retweeted and fewer bullish tags are predictive of higher volatility.

5.1.2 Text Analysis

The bearish and bullish tags had the strongest relationship with the market. The proportion of bullish and bearish tags were correlated with next price rise and retraction respectively. Though, the predictive capacity of this model was weak relative to the baseline.

The McDonald and Loughran negative word count had a modest negative correlation with same day price. Only the bullish tags featured in the same day correlation model. The proportion of bullish tags in the retweet network were also negatively correlated with next day volatility and bearish tags and the bear(ish) word count correlated with same day volume.

While the LDA classification of the text into opinion and technical bins had a low misclassification rate, the propensity to tweet technical content did not vary. Technical tweets only featured in the same day volatility model in which the technical tweets in the full network was one leaf node, negatively correlated with market volatility.

The number technical tweets in the full network is the first node split in the Rank 1 model. Technical tweets are negatively correlated with the proportion of rank 1 users in the full network. This was followed by a negative correlation with bullish and bearish tags. Rank 1 users are negatively correlated with tags and technical content.

In addition to that the proportion of technical tweets is correlated with the proportion of rank 2 users in the full network. In the full network therefore there is some evidence for a correlation between rank 2 users and technical content. This was not the case in the retweet network.

5.1.3 User Rating

From the initial exploration we know rank 1 users vary whereas rank 2 users remain relatively constant in both the networks. The number Rank 1 users in the full network has the strongest relationship with next day volume of stock traded in the baseline model.

The proportion of rank two users retweeted is negatively correlated with volatility. The proportion of retweets in the network is the first node to split volatility in the prediction model, from this we understand that there is more retweeting but that it is Rank 1 users.

In both of these it is more lower ranked users in the network which is correlated with higher greater market activity.

The number of users in the full network has the strongest relationship with the proportion of rank 2 users retweeted. This is intuitive from the line plots, we know that changes in activity are largely attributable to rank 1 users.

Assortativity and modularity and more technical tweets are also correlated with the proportion of rank 2 users retweeted. This implies fewer clusters form within the network and users are connecting less with similar users. There is also a greater proportion of technical content.

5.1.4 Social Network Analysis

Assortativity featured most frequently of the social network variables, though the models in question did not have a strong relationship with the dependant variable relative to the baseline models. Assortativity was the second feature split in both the price prediction and same-day correlation models. As price goes up people retweet and reply to users similar to themselves. It also featured in the same day volatility, lesser assortativity is associated with higher volatility.

The Gini index of the in-degree centrality was the first node to split the proportion of technical tweets in the retweet network. This implies more equality between nodes while there is more technical content. Lower assortativity and modularity are associated with the proportion of rank 2 users in the retweet network.

5.2 Observations from the Results

The premise of this study is not to elaborate on model metrics and improve prediction but instead to identify network features that provide trustworthy and high quality information. In order to that correlation between content and market variables was investigated.

It appears that the volume of lower ranked users and opinion based tweets have a correlation with market variables however the propensity for higher ranked users and technical remain constant. Higher ranked users and technical content do correlate however those are not associated with retweet behaviour nor clusters of users or fewer central users as measured by SNA metrics. Put another way higher ranked users and technical content as measured here does not draw more attention during certain market circumstances. The propensity to retweet does, the tweetcount does but that increase in the volume of activity is not considered good quality content according to the parameters derived in this experiment. Of the text classification methods, the bullish and bearish tags users can add to their tweet had the most insightful correlation with market circumstances particularly price changes.

5.3 Strengths of the Results

Three factors stood out for their unique contribution to the literature; the increased activity of rank 1 users in correlation with next day volume of stock traded, the propensity to retweet and reply in correlation with next day volatility, the relationship between the bullish and bearish tags and market variables in particular price.

A second strength of the results is they mirror other findings, namely the increased activity of lower ranked users in correlation with greater market activity. The inherent voting mechanisms of the network are not functioning here to spread better quality content. In fact, network correlation with market activity it is not technical content or the best connected users but lesser connected ones and opinion based pieces that are posted and retweeted and replied to. When the network forms greater modules or the in-degree centrality is higher and a few central users are gaining more attention that is lower ranked people and opinion based content again.

A final strength is that the models are simple and computationally light. They could be altered and extended to accommodate more features and larger datasets in order to build on what has been achieved here.

5.4 Limitations of the Results

The first issue with these models is that of overfitting. Overfitting could be due to a spurious split or to the accumulation of small errors from many splits. It has been addressed by pruning the trees and testing them on a smaller dataset withheld from the model training phase. However, in these models it could remain an issue. If prediction was sought this step of the modelling could be very beneficial to selecting features which are higher in the tree and filtering out spurious features lower down which measure similar phenomena. For example, including both the proportion of rank 2 users and rank 1 users retweeted, as is the case in the model for same day volume correlation here, would likely cause overfitting.

The next limitation of these models is that of model development. The potential for generating more features which describe network dynamics is huge. In a network considered to be so noisy the value is likely to be found in the nuance. It is with the generation of more diverse features to measure more discrete changes that may identify *'useful'* information. In addition to that these models depends largely on the propensity to tweet. What that means is that a number of features are generated and their volume was correlated with market fluctuations. There is no reason to believe that

volume of one genre of tweets versus another will be tweeted more often under any circumstance. In fact, as Casarin et. al, (2015) found from the Italian social network, it was lower ranked members and opinion based content that correlated with greater market activity.

An associated limitation of these models is that of parameter tuning. This experiment took a horizontal approach, generating features to measure a range of network phenomena. The time for further model development and parameter tuning was not available. In fact, each category of features, the social network, text and ranking justify an experiment of their own in order to establish the best set of measures that most closely define underlying network dynamics. For example, the ranking system is based upon activity and a voting mechanism; tweet count, number of followers and following. From the results here we know the network has a greater propensity towards noisy content during greater market activity. Therefore, a user ranking system that is not dependent on network voting might perform better, though it would almost certainly be more complex to generate.

In addition to the model and parameter improvements it could be that the model of best fit and the parameters of greatest relevance vary depending on the nature of the market itself. They have not been tested during isolated events such as a period of high volatility, during a shock, during market expansion or retraction or to identify a particular change. It is evident from the models that price, volatility and volume are associated with unique categories of network activity. It is plausible therefore that events and trends bear with them specific feature relationships that would lead to more powerful models which do not generalize over all market circumstances.

Another limitation of this research is its unique focus on one stock during one year only. The range of volatility and price fluctuation is narrow relative to riskier stocks. It is a limitation of such a social network where discussion is focused on stocks that resonate with the public. However, this is their nature and unless a wider interest is developed with respect to more diverse stocks then this is where their utility might be restricted to.

6. CONCLUSIONS AND FUTURE WORK

In this final chapter is a review of the experiment that was performed, it's contribution to the field and recommended next steps to elaborate on the findings here.

6.1 Research Overview

Three categories of social network features were generated to measure a relationship between a social network and a the AAPL stock price, daily volatility and volume. Two networks were tested, a full Stocktwits network and a Stocktwits retweet network containing all replies and retweets to highlight the most popular content. The three categories of features, reputation, text analytics and SNA were generated for both networks. The research question was whether reputational ranking, text classification and SNA metrics could distinguish between content that enables learning about market trends and noisy content.

The challenge in verifying information from open platforms has come under the spotlight recently. What is deemed *'useful'* or *'good quality'* information is case specific. In developing a filter to search out that which is relevant and helpful for understanding AAPL stock, theory and models from finance, economics, psychology, sociology and social network analysis came to play. Both investment decision making and social network indicators were required to develop a robust method for feature selection.

Here higher quality information was sought by seeking correlations between AAPL stock changes and changes in network behaviours associated with AAPL stock. In addition to that correlation between network features were investigated in order to go beyond measurements of propensity to tweet in correlation with financial changes and to better understand which network behaviours occur is parallel.

6.2 Contribution & Impact

This experiment adds three new pieces of information to the current understanding or social networks and their relationship with financial markets.

- The first in relation to price is the correlation of the bullish and bearish tags with a next day increasing and decreasing price respectively. These tags are not built into other social media platforms and therefore offer a new tool for correlation analysis. They have out-performed three other models for text analysis and market correlation in this experiment.

- The second in relation to market volatility is the higher propensity to retweet and reply to other users in correlation with next day volatility. The role of replies and retweets with respect to the market has not been widely reported on.

- The final contribution and in fact the one with predictive power above financial features available was the correlation of rank one users in the full network with a next day volume increase.

In addition to these new insights this experiment adds weight to findings reported previously by other researchers while investigating similar phenomena using datasets from other social networks and the stock market. Bollen, May & Zeng, (2011, p2) found that the incidence of the word ‘bull’ or ‘bullish’ in tweets could predict a rising price. The same was tested here and it did not apply. The ‘bullish’ and ‘bearish’ are not embedded in Twitter and that might account for the difference. Casarin, Casnici, Dondio & Squazzoni (2015, p.51) found that non-professional users posted more during times of volatility and that the content of their messages changed, with spam and opinion based messages increasing during periods of higher uncertainty relative to more technical analysis during calm market periods. The same was found here with respect to retweets. More specifically the additional retweets correlated with volatility were of rank 1 users and opinion based topics. A positive bias was also apparent here as in many of the predicating studies. The ‘bullish’ tag had a stronger relationship with market variables than it’s ‘bearish’ counterpart.

6.3 Future Work & Recommendations

The future work from this experiment lends from the limitations. The easiest gains in improving the finding here would likely be in lengthening the period of analysis and expanding financial variables beyond AAPL stock alone. In addition to that concurrent data from another social network would add depth to the research and offer a comparison set of network behaviours.

Feature tuning warrants a number of intricate experiments. There might be alterations to the features as they are measured here that better represent the underlying phenomena under investigation. Establishing the most powerful classification of reputation, topic analysis and SNA features are each complex and have been little explored with respect to Stocktwits.

Perhaps rather than reputation a 'truth' measure could be generated. This involves filtering trustworthy users using a history of user interaction and the measurement of things such as, subject matter expertise / past predictions / recommendations / opinions / accuracy ratings / recency of accurate assertions. With a different ranking system other features could be added, such as the propensity for higher and lower ranked people to use bullish and bearish tags and their accuracy.

The topic analysis metrics have room for improvement, sentiment alone is a crude measure and the strongest here was from the tags not the text. The work by (Longo, Dondio, & Barrett, 2010) and (Dondio & Longo, 2011) and (Dondio et al., 2006) proposes a trust metrics in the context of online search engines that could be applied to our dataset. An improved measurement of text analytics would be a required input to the above-proposed trust metric. Another area of potential future works is the introduction of a more fine-graned text analysis. In this context, the area of argumentation mining, such the work by (Dondio, 2014) could be applied to the Stocktwits dataset. The text analysis here would require a more refined model to establish correct assertions. Experimentation with a variety of alternative pooling mechanisms might enable a more powerful classification model.

There is also potential to develop features which have not been touched upon here. An imagination and working understanding for network behaviour could inform the development of features which have not yet been conceptualized. For example, the role of time could have significance, e.g. there might be a time component such that tweets with a bullish tag within a certain interval or a bearish tag have a greater correlation with next day closing price, or perhaps the time within which a retweet or reply takes place is of significance. Stern et. al, (2008, p.14) included only retweets within the hour in their model to predict future retweets. Retweets outside of that horizon were no longer deemed relevant to the events to which they were originally referring. All retweets were included in this experiment. It is possible time plays a role in many of the features measured and nudges their importance one way or another. While rank 2 and technical tweets remained at a constant rate throughout 2015 it might have a time component, perhaps higher quality users and content is posted during a particular time frame.

BIBLIOGRAPHY

Abu-Salih, Bilal, Pornpit Wongthongtham, Seyed-Mehdi-Reza Beheshti, and Dengya Zhu. "A Preliminary Approach to Domain-Based Evaluation of Users' Trustworthiness in Online Social Networks," 460–66. IEEE, 2015. doi:10.1109/BigDataCongress.2015.74.

Alvarez-Melis, David, and Martin Saveski. "Topic Modeling in Twitter: Aggregating Tweets by Conversations." In *Tenth International AAAI Conference on Web and Social Media*, 2016. http://socialmachines.media.mit.edu/wp-content/uploads/sites/27/2014/08/topic-modeling-twitter.pdf.

Antweiler, W., & Frank, M. (2004). Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards. *The Journal of Finance, 59*(3), 1259-1294.

Badham, Jennifer M. "Commentary: Measuring the Shape of Degree Distributions." *Network Science* 1, no. 02 (2013): 213–25.

Bandari, Roja, Sitaram Asur, and Bernardo A. Huberman. "The Pulse of News in Social Media: Forecasting Popularity." *arXiv Preprint arXiv:1202.0332*, 2012. http://arxiv.org/abs/1202.0332.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3, no. Jan (2003): 993–1022.

Bollen, Johan, Bruno Gonçalves, Guangchen Ruan, and Huina Mao. "Happiness Is Assortative in Online Social Networks." *Artificial Life* 17, no. 3 (2011): 237–51.

Bollen, Johan, Huina Mao, and Xiaojun Zeng. "Twitter Mood Predicts the Stock Market." *Journal of Computational Science* 2, no. 1 (March 2011): 1–8. doi:10.1016/j.jocs.2010.12.007.

Bourdieu, Pierre. "The Forms of capital.(1986)." *Cultural Theory: An Anthology*, 2011, 81–93.

Braess, D. "Über ein Paradoxon aus der Verkehrsplanung." *Unternehmensforschung Operations Research - Recherche Opérationnelle* 12, no. 1 (December 1968): 258–68. doi:10.1007/BF01918335.

Casarin, R., Casnici, N., Dondio, P., & Squazzoni, F. (2015). Back to Basics! The Educational Gap of Online Investors and the Conundrum of Virtual Communities. Journal of Financial Management, Markets and Institutions, 3(1), 51–69.

Casnici, N., Dondio, P., Casarin, R., & Squazzoni, F. (2015). Decrypting financial markets through e-joint attention efforts: on-line adaptive networks of investors in periods of market uncertainty. PloS One, 10(8), e0133712.

Chung, Kon Shing Kenneth, Mahendra Piraveenan, Andrew Vakarau Levula, and Shahadat Uddin. "Assessing Online Community-Building through Assortativity, Density and Centralization in Social Networks," 1993–2002. IEEE, 2013. doi:10.1109/HICSS.2013.104.

Coleman, James S. "Social Capital in the Creation of Human Capital." *American Journal of Sociology* 94 (January 1988): S95–120. doi:10.1086/228943.

Computer Society, and Association for Computing Machinery, eds. *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012): Istanbul, Turkey, 26 - 29 August 2012 ;* Piscataway, NJ: IEEE, 2012.

Das, Sanjiv R., and Mike Y. Chen. "Yahoo! For Amazon: Sentiment Extraction from Small Talk on the Web." *Management Science* 53, no. 9 (September 2007): 1375–88. doi:10.1287/mnsc.1070.0704.

Dondio, P. (2012). Predicting Stock Market Using Online Communities Raw Web Traffic. In Web Intelligence and Intelligent Agent Technology (WI-IAT), 2012 IEEE/WIC/ACM International Conferences on (Vol. 1, pp. 230–237). IEEE.

Dondio, P. (2013). Stock market prediction without sentiment analysis: using a web-traffic based classifier and user-level analysis. In System Sciences (HICSS), 2013 46th Hawaii International Conference on (pp. 3137–3146). IEEE.

Dondio, P. (2014). Toward a computational analysis of probabilistic argumentation frameworks. Cybernetics and Systems, 45(3), 254–278.

Dondio, P., Barrett, S., Weber, S., & Seigneur, J. M. (2006). Extracting trust from domain analysis: A case study on the wikipedia project. In International Conference on Autonomic and Trusted Computing (pp. 362–373). Springer.

Dondio, P., & Longo, L. (2011). Trust-based techniques for collective intelligence in social search systems. In Next Generation Data Technologies for Collective Computational Intelligence (pp. 113–135). Springer.

Gintis, Herbert. *The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences*. Princeton, N.J: Princeton University Press, 2009.

Huang, Dongxu, Jing Zhou, Dejun Mu, and Feisheng Yang. "Retweet Behavior Prediction in Twitter," 30–33. IEEE, 2014. doi:10.1109/ISCID.2014.187.

Kahneman, Daniel, and Amos Tversky. “Prospect Theory: An Analysis of Decision under Risk.” *Econometrica: Journal of the Econometric Society*, 1979, 263–91.

Liu, Gang, Chuan Shi, Qing Chen, Bin Wu, and Jiayin Qi. “A Two-Phase Model for Retweet Number Prediction.” In *Web-Age Information Management*, edited by Feifei Li, Guoliang Li, Seung-won Hwang, Bin Yao, and Zhenjie Zhang, 8485:781–92. Cham: Springer International Publishing, 2014. http://link.springer.com/10.1007/978-3-319-08010-9_84.

Longo, L., Dondio, P., & Barrett, S. (2007). Temporal Factors to evaluate trustworthiness of virtual identities. In Security and Privacy in Communications Networks and the Workshops, 2007. SecureComm 2007. Third International Conference on (pp. 11–19). IEEE.

Longo, L., Dondio, P., & Barrett, S. (2010). Enhancing social search: A computational collective intelligence model of behavioural traits, trust and time. In Transactions on computational collective intelligence II (pp. 46–69). Springer.

Racca, P., Casarin, R., Squazzoni, F., & Dondio, P. (2016). Resilience of an online financial community to market uncertainty shocks during the recent financial crisis. Journal of Computational Science, 16, 190–199.

Mao, Huina, Scott Counts, Johan Bollen, and others. “Quantifying the Effects of Online Bullishness on International Financial Markets.” In *ECB Workshop on Using Big Data for Forecasting and Statistics, Frankfurt, Germany*, 2014. http://www.busman.qmul.ac.uk/newsandevents/events/eventdownloads/bfwgconference2013acceptedpapers/114925.pdf.

Myerson, Roger B. *Game Theory: Analysis of Conflict*. 6. print. Cambridge, Mass.: Harvard Univ. Nash, John. “Non-Cooperative Games.” *The Annals of Mathematics* 54, no. 2 (September 1951): 286. doi:10.2307/1969529.

Press, 2004.

Nguyen, Dat Quoc, Richard Billingsley, Lan Du, and Mark Johnson. “Improving Topic Models with Latent Feature Word Representations.” *Transactions of the Association for Computational Linguistics* 3 (2015): 299–313.

Nguyen, Thien Hai, and Kiyoaki Shirai. “Topic Modeling Based Sentiment Analysis on Social Media for Stock Market Prediction.” In *Proceedings of the 53rd Annural Meeting of the Association for Computational Linguistics*, 2015. http://www.aclweb.org/old_anthology/P/P15/P15-1131.pdf.

Oh, Chong, and Olivia Sheng. “Investigating Predictive Power of Stock Micro Blog Sentiment in Forecasting Future Stock Price Directional Movement,” 2011. http://aisel.aisnet.org/icis2011/proceedings/knowledge/17/.

Otte, E., and R. Rousseau. “Social Network Analysis: A Powerful Strategy, Also for the Information Sciences.” *Journal of Information Science* 28, no. 6 (December 1, 2002): 441–53. doi:10.1177/016555150202800601.

Pang, Bo, and Lillian Lee. “Opinion Mining and Sentiment Analysis.” *Foundations and Trends® in Information Retrieval* 2, no. 1–2 (2008): 1–135. doi:10.1561/1500000011.

Peng, Huan-Kai, Jiang Zhu, Dongzhen Piao, Rong Yan, and Ying Zhang. “Retweet Modeling Using Conditional Random Fields,” 336–43. IEEE, 2011. doi:10.1109/ICDMW.2011.146.

Phan, Xuan-Hieu, Cam-Tu Nguyen, Dieu-Thu Le, Le-Minh Nguyen, Susumu Horiguchi, and Quang-Thuy Ha. “A Hidden Topic-Based Framework toward Building Applications with Short Web Documents.” *IEEE Transactions on Knowledge and Data Engineering* 23, no. 7 (July 2011): 961–76. doi:10.1109/TKDE.2010.27.

Ramage, Daniel, Susan T. Dumais, and Daniel J. Liebling. “Characterizing Microblogs with Topic Models.” *ICWSM* 10 (2010): 1–1.

Sankar, C. Prem, R. Vidyaraj, and K. Satheesh Kumar. “Trust Based Stock Recommendation System – A Social Network Analysis Approach.” *Procedia Computer Science* 46 (2015): 299–305. doi:10.1016/j.procs.2015.02.024.

Shiller, Robert J. *The New Financial Order: Risk in the 21st Century*. Princeton, N.J.: Princeton University Press, 2003. http://www.books24x7.com/marc.asp?bookid=30603.

Sprague, Ralph H., Shidler College of Business, and Computer Society, eds. *2013 46th Hawaii International Conference on System Sciences (HICSS 2013): Wailea, [Maui], Hawaii, USA, 7 - 10 January 2013 ; [proceedings]*. Piscataway, NJ: IEEE, 2013.

Sprenger, Timm O., Andranik Tumasjan, Philipp G. Sandner, and Isabell M. Welpe. “Tweets and Trades: The Information Content of Stock Microblogs: Tweets and Trades.” *European Financial Management* 20, no. 5 (November 2014): 926–57. doi:10.1111/j.1468-036X.2013.12007.x.

Stern, David H., Ralf Herbrich, and Thore Graepel. “Matchbox: Large Scale Online Bayesian Recommendations.” In *Proceedings of the 18th International Conference on World Wide Web*, 111–20. ACM, 2009. http://dl.acm.org/citation.cfm?id=1526725.

Tetlock, Paul C. “Giving Content to Investor Sentiment: The Role of Media in the Stock Market.” *The Journal of Finance* 62, no. 3 (2007): 1139–68.

Thaler, Richard H. *Misbehaving: How Economics Became Behavioural*. London: Lane, 2015.

Wasserman, Stanley, and Katherine Faust. *Social Network Analysis: Methods and Applications*. Structural Analysis in the Social Sciences 8. Cambridge ; New York: Cambridge University Press, 1994.

Yadati, Narahari, and Ramasuri Narayanam. “Game Theoretic Models for Social Network Analysis.” In *Proceedings of the 20th International Conference Companion on World Wide Web*, 291–92. ACM, 2011. http://dl.acm.org/citation.cfm?id=1963316.

Yang, Dennis, and Qiang Zhang. “Drift-Independent Volatility Estimation Based on High, Low, Open, and Close Prices*.” *The Journal of Business* 73, no. 3 (2000): 477–92.

Zaman, Tauhid R., Ralf Herbrich, Jurgen Van Gael, and David Stern. “Predicting Information Spreading in Twitter.” In *Workshop on Computational Social Science and the Wisdom of Crowds, Nips*, 104:17599–601. Citeseer, 2010. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.208.5687&rep=rep1&type=pdf.

Zhao, Huidong, Gang Liu, Chuan Shi, and Bin Wu. “A Retweet Number Prediction Model Based on Followers’ Retweet Intention and Influence,” 952–59. IEEE, 2014. doi:10.1109/ICDMW.2014.152.