**THE RELATIONSHIP BETWEEN TWITTER SENTIMENT AND STOCK PERFORMANCE: A DECISION TREE APPROACH**

|  |  |  |
| --- | --- | --- |
| Yichao Dai | Rongjuan Chen | Ruoqian Dong |
| *yichaod@kean.edu* | *ronchen@kean.edu* | *ruoqiand@kean.edu* |
| Wenzhou-Kean University | | |
| Jilei Lin | | |
| *jileil2@illinois.edu* | | |
| University of Illinois at Urbana-Champaign | | |

**Abstract**

In the era of social media and big data technologies, investors nowadays find it easy and, thus, may prefer to use social media to make informed decisions. Social media has become not only a communication tool, but also valuable database for researchers and practitioners to gather information, share knowledge, as well as express opinions about stock performance. The sentiment based on social media content can be further analyzed, in order to predict stock performance. Although numerous past studies have attempted to predict stock price movement based on social media sentiment, some emerging analytical tools, like existing lexicons, may require further testing and validation in a financial decision making context. And, these behavioral factors are definitely worth consideration in the view of Finance and Economics scholars. In this study, we develop and test predictive models for stock price and trend forecasting. By using a large-scale sample of tweets collected from Twitter, related to four companies, Apple, Google, Microsoft, and Netflix, we propose a novel decision tree approach to stock performance prediction. Based on our findings, we finally provide theoretical and practical implications as well as directions for future work.

Keywords: Social media, Stock market, Sentiment analysis, Decision tree

1. Introduction

With the rapid development of social media, investors nowadays tend to use information gathered from the Internet to make their decision to buy or sell stocks. Although buying or selling stocks seems to be more like a practical, and sometimes intuitive, decision made by investors in the real world, using online information to predict stock price has become increasingly popular among finance and economics researchers. Quantitative approaches have been widely and thoroughly employed to answer the question – whether a stock in one’s consideration will be bullish or bearish in both short- and long-term [1]. As a result, the movement of a stock’s price will heavily influence an investor’s decision to manage the portfolio.

For decades, stock price prediction has captured the attention of a large number of researchers and practitioners in various disciplines, such as computer science, statistics, finance, and economics [2]. We have noticed the rising trend of using sentiment analysis, based on the information diffused through social media, to forecast the future movement of stocks [3, 4]. According to past studies, investment decision can be greatly impacted by online opinions and social media content-embedded sentiment [4, 5, 6, 7].

In this study, we aim to explore the behavioral finance factors, particularly the sentiment extracted from social media content, and examine their influence on stock performance. We employ existing lexicons to assign sentiment scores to stock-related tweets and use these scores as predictive variables in multiple linear regression and decision tree model for comparison. The results of both models are reported and the significant findings, by taking a decision tree approach, are discussed in detail. Finally, practical implementations and future research directions are addressed, regarding the theoretical contribution of our findings as well as the recommendations for developing next generation social media-based applications and tools for better decision making of investors.

1. Research background
   1. Twitter and social media

Social media has become a popular communication channel for people to exchange ideas and express opinions. It is easy and convenient for one to create and share timely information through social media. Past research has defined social media as Internet-based applications that allow creation and exchange of user-generated content [8].

Particularly, Twitter is a microblogging social media platform that allows users to post and repost a short message of no more than 140 characters, called “tweet”. Past studies have claimed that Twitter is one of the most popular social media channels in the world [9]. In this study, we aim to analyze twitter data to further explore the predictive value of social media information for the stock price movement.

* 1. Twitter sentiment and financial market

Since Twitter is a widely used social media platform for timely information exchange, it is very likely that the information spread on Twitter can influence how people perceive information and use the information to make a decision. A past survey conducted by Thomson Reuters has found that a vast number of financial professionals prefer to make informed decisions in their professional practices, by using social media information [10].

Previous studies have revealed the correlation between microblogging posts and stock market events [11]. A study has proposed that public moods can predict the stock market [12]. In addition, researchers have demonstrated that information available on social media can predict future stock returns [13].

To extract online opinions for better decision making, researchers have attempted to analyze social media information to understand its impact on equity value, stock price, and attitude toward public policy and commercial products [14, 15, 16]. The relationship between online content and stock performance has been addressed in numerous early studies. However, given the rapid growth of social media technologies and big data applications, some emerging text analytical tools need to be further implemented and evaluated. In this work, we are attempting to test and compare several lexicons for text analysis using data collected from Twitter, in order to better understand the public opinions, regarding financial decision making, of social media users.

1. Sentiment analysis of Twitter posts

Since social media-based prediction has received considerable attention by researchers and practitioners in recent research, sentiment analysis based on social media content has been widely adopted by scholars and investors to better understand the stock market [15, 17]. In particular, since the link between social media and the stock market has been widely recognized, the derived theory is defined as behavioral finance, which is a theoretical rationale behind investment decision making [18]. Behavioral finance claims that, besides the well-known model of efficient markets traditionally applied to study financial decisions, some less rational factors such as investors’ sentiment and public opinions are influential and worth more consideration in the era of social media and big data technologies.

Following this line of research, social media becomes a valuable source of information for researchers and practitioners to further explore the relationship between social media sentiment and the stock market. As pointed out in past research, the relationship between social media information transmission and stock performance has not been sufficiently defined and researched [18].

Big data processing, or sentiment analysis, is one of the most important and challenging research topics in the era of Web 2.0. Sentiment analysis and opinion mining require various computational and statistical techniques, attempting to discover, extract, and classify human emotions, public opinions, and collective behaviors from textual information in social media environment [19, 20].

There are different sentiment analysis approaches. Among them, one of the most commonly used techniques is to label emotional words, by using an existing lexicon, and then quantify and classify sentiment embedded in online content [21, 22, 23]. With the sentiment extracted from social media data, such as a large sample of tweets, sentiment variables can be further utilized to predict various behaviors of the society.

1. Date collection

Tweets that contain a keyword corresponding to each ticker of AAPL, GOOG, MSFT, and NFLX have been filtered and collected from Twitter by using a text mining tool developed in R, named “rtweet”, and the official Twitter streaming API. The raw data has been downloaded during the time frame of 15 days (from 03/26/2021 to 04/09/2021), then cleaned and transformed into four datasets, one at each stock ticker. The number of APPL, GOOG, MSFT, and NFLX is 35,202, 13,463, 35,637, and 11,337, respectively. For the four stocks, datasets are processed separately. Besides Twitter data, we have collected stock performance data from Bloomberg database, containing hourly and daily price information within the same time frame as Twitter data.

1. Sentiment analysis

Sentiment analysis is an NLP technique for extracting and mining public opinions embedded in textual information gathered from a social media platform or a website. In this study, we adopt a learning approach for each stock based on existing lexicons, including 1) AFINN, 2) bing, and 3) NRC. These lexicons are all based on unigrams, or single words. For many of the English words contained in these lexicons, a score will be assigned to each for positive or negative sentiment.

In particular, the bing lexicon categorizes words into two types, positive and negative [24]. The AFINN lexicon assigns words with a score ranging from -5 to 5, with negative scores indicating negative sentiment and positive scores indicating positive sentiment [25]. Last, the NRC lexicon not only identifies words as positive or negative, but also categorizes them into several types of sentiment, such as anger, anticipation, disgust, fear, joy, sadness, surprise, and trust [26].

The R package “tidytext” allows us to use specific lexicons as discussed above. For the sentiment analysis of Twitter data, we first want to use these different lexicons and compare their results, in order to select the one that can best serve our research purpose. It is noteworthy that, the lexicons mentioned above have been constructed by human subjects through crowdsourcing platforms (e.g., Amazon Mechanical Turk) or by experts with domain specific knowledge, and further validated using various analytical tools. Thus, in general these methods for text analysis and opinion mining are valid. But still, we need to further validate how these lexicons can be efficiently and effectively applied, in order to help us better understand social media sentiment regarding financial information. For this reason, we think our work is significant and meaningful. Figure 1 explains how lexicons are used for sentiment analysis based on stock-related tweets.

|  |  |  |
| --- | --- | --- |
| Tweets  Tweets  Stock-related Tweets | Pre-processing | Transformation |
| Stock prediction | Sentiment scores | Word matching and labeling |
|  |  |  |

Figure 1. A framework of lexicon-based sentiment analysis

As shown below, we use AAPL as an example to demonstrate the results of sentiment analysis by using these lexicons. First, while using bing to analyze the sentiment of stock-related tweets of APPL, we aggregate tweets on an hourly basis and score the sentiment of each day’s aggregated tweets, according to the following equation.

*Sentiment of an hourly tweet = (Total score of positive – Total score of negative)*

Figure 2 shows the normalized sentiment scores, obtained by the bing lexicon, of AAPL’s daily tweets over the 15 days of time frame. Second, in a similar manner the sentiment scores of AAPL have been created by using the AFINN lexicon, as shown in Figure 3.

Using the KS-test to compare the score distributions based on the two lexicons has shown significant difference (*p* < .05), indicating that the bing and AFINN can yield different results for sentiment analysis of our sample tweets.

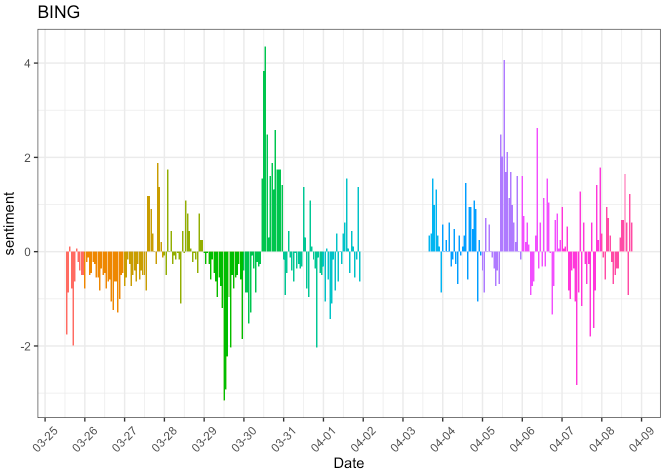


Figure 2. The bing-based normalized sentiment score of APPL tweets

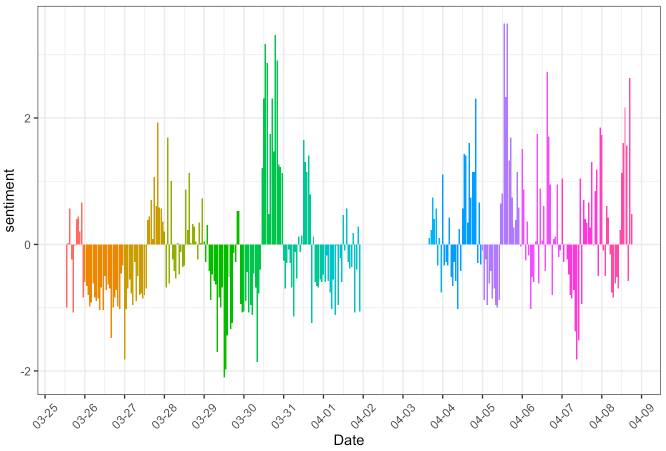


Figure 3. The AFINN-based normalized sentiment score of APPL tweets

Next, because the NRC lexicon allows us to diversify the sentiment extracted from tweets and quantify various types of sentiment as the predictive variables to be further used in future modeling process, we are more optimistic about the sentiment scores created by this lexicon. In the following section, we discuss in detail how sentiment scores are used in a predictive model of stock performance, by comparing these lexicons. Figure 4 demonstrates the frequency distribution of sentiment scores on hourly basis, based on APPL.

Table 1 shows two sample tweets of AAPL, one of which is overall more positive with significant feelings of Anticipation, Trust, and Joy and the other one more negative with significant feelings of Sadness, Anger, and Fear. In addition, in Figure 5 and Figure 6 we present the top 10 most frequent words that contribute to each type of sentiment based on APPL tweets.

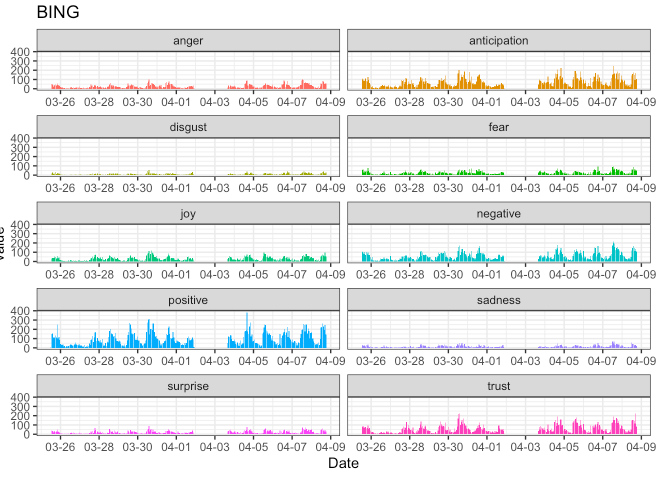


Figure 4. The NRC-based sentiment score of APPL tweets

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sample tweet | Sentiment score | | | | |
| *WIMI patience will pay off big dividends. Tons of huge news on major technological fronts are coming. Remember hologram , 5G mobile, gaming, any big partnership license or customer win gap this up min to double digits if we get a TSLA or AAPL partnership.*  *(Positive example)* | Pos | Neg | Ang | Ant | Dis |
| 12 | 2 | 0 | 8 | 0 |
| Fea | Joy | Sad | Sur | Tru |
| 0 | 2 | 0 | 0 | 8 |
| *Apple Inc. (AAPL) surprised the market with Q3 result. Merrill Lynch changed the rating to Hold, as Apple Inc. (AAPL) has negative EPS revisions (down 3%) and not enough upside to warrant a Buy rating.*  *(Negative example)* | Pos | Neg | Ang | Ant | Dis |
| 0 | 4 | 3 | 1 | 1 |
| Fea | Joy | Sad | Sur | Tru |
| 3 | 0 | 4 | 1 | 0 |

Table 1. Sample tweets and NRC-based sentiment scores of AAPL

*Note: Pos = positive, Neg = negative, Ang = anger, Ant = anticipation, Dis = disgust, Fea = fear, Joy = joy, Sad = sadness, Sur = surprise, and Tru = trust*

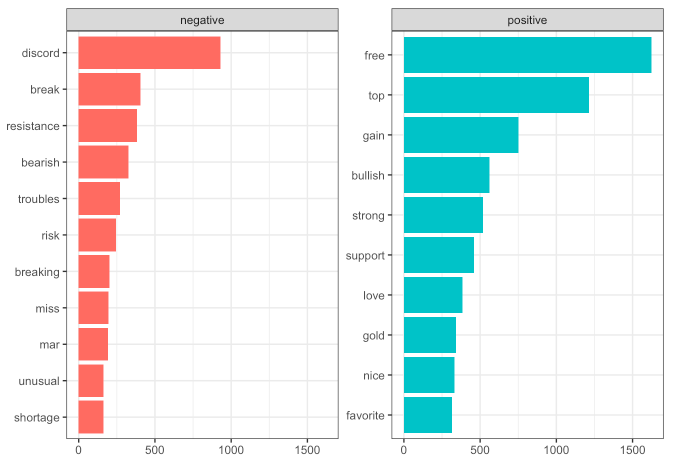


Figure 5. The top 10 frequent words contributing to positive and negative sentiment

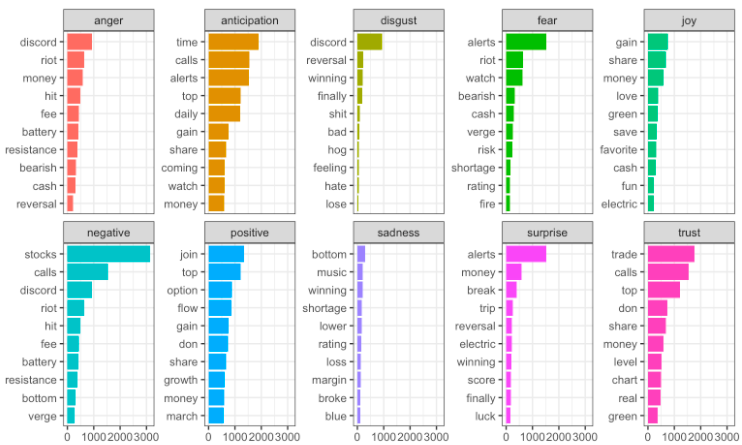


Figure 6. The top 10 frequent words contributing to each specific emotional category

Once we have obtained the sentiment scores of hourly stock-related tweets by using the three lexicons, we need to further decide which one is appropriate our research purposes. Later, when we are preparing predictive models we rely on sentiment scores as predictive variables in three separate models for comparison.

1. Predicting stock performance based on sentiment

A common sense of predicting stock performance is to develop and test a multiple linear regression as it is simple and the results are easy to interpret. So, we first use a regression model including NRC-based sentiment scores as independent variables and stock price as dependent variable. Actual stock price of AAPL, GOOG, MSFT, and NLFX has been collected from Bloomberg database, which is considered as ground truth for stock performance measure.

Moreover, we need to clarify that the relationship between sentiment and stock performance can vary from opening hours to closing hours of stock market. Thus, besides an overall regression model, containing hourly-based stock price and sentiment, we also develop two separate regression models, in the same fashion, and then compare their results.

The results indicate that, in general a multiple linear regression model is acceptable for predicting stock price (adjusted *R2* = .14, *p* < .001). Interestingly, the model for opening hours (adjusted *R2* = .16, *p* < .05) seems to be slightly more robust than the one for closing hours (adjusted *R2* = .11, *p* < .05). This is reasonable, to our understanding, because when the stock market is open people are more actively engaged in social media information and become more likely to be influenced by the content-based sentiment. Their judgment about a stock to rise or fall can determine the movement of stock price afterwards. On a side note, the performance of the regression models based on AFINN- and bing-based sentiment scores to predict stock price is poor (*p’s* > .05), and thus not worth any further discussion in this paper.

Then, taking APPL as an example, we use the data collected from 03-25 to 04-09 as training set and a new dataset starting from 04-10 till 04-15 for testing. For the testing data, the predicted stock price is compared with the actual price. The correlation graphs of the four stocks are shown in Figure 7. The predicted stock price of GOOG is less fitted than some other stocks, such as APPL and MSFT, which is a finding similar to that of a past study [27].

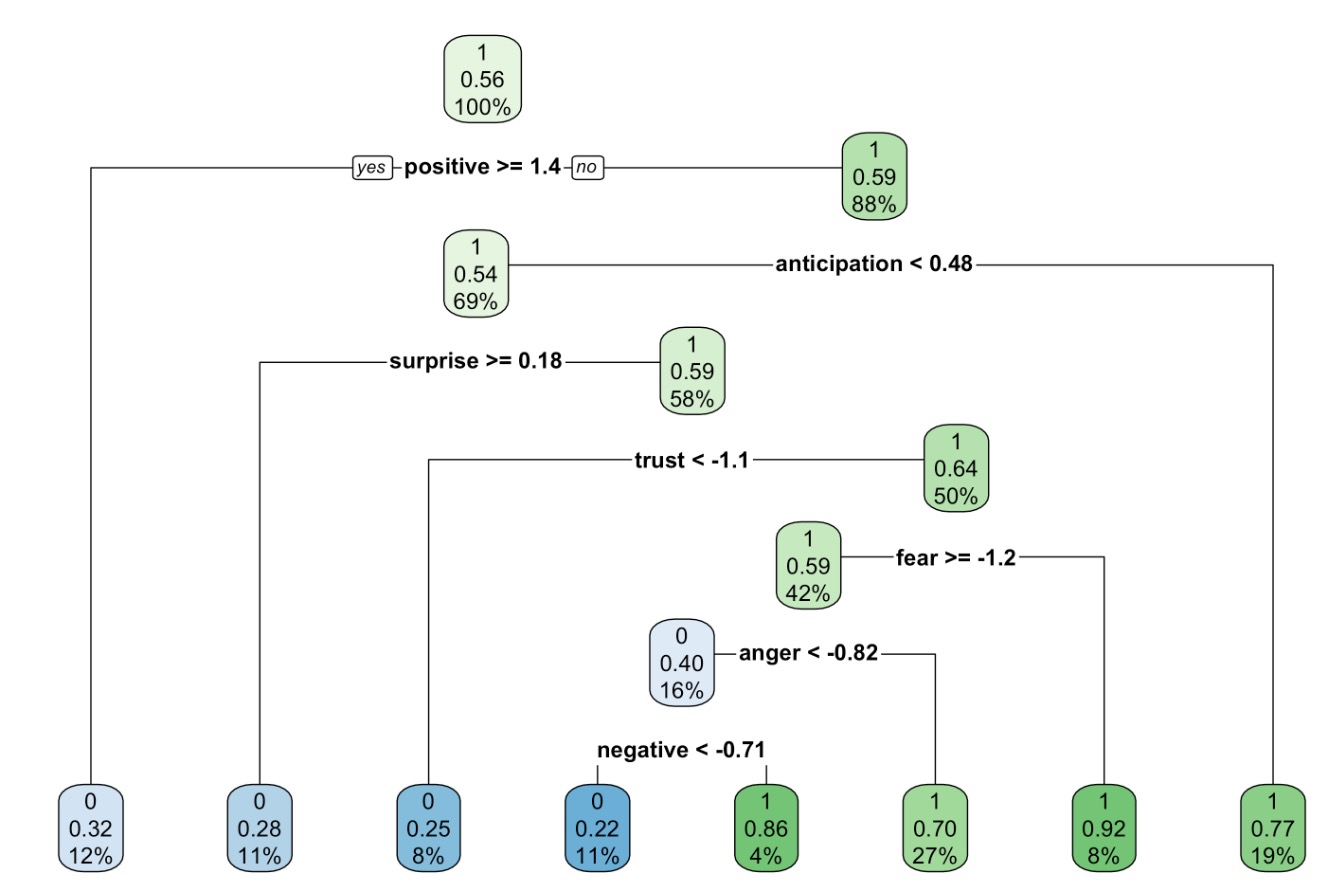
|  |  |
| --- | --- |
|  |  |
|  |  |

Figure 7. The correlation between actual and predicted stock price of a random sub-sample

For other companies, GOOG, MSFT, and NFLX, the results of multiple linear regression for prediction of stock price are mostly acceptable but not quite consistent across companies. To explain, for GOOG the regression model performs best for the stock price in opening hours (adjusted *R2* = .21, *p* < .05) but very poor in closing hours, not even acceptable. For MSFT, the overall regression model for opening and closing hours (adjusted *R2* = .12, *p* < .001) outperforms the two separate predictions for opening and closing. For NFLX, the regression model for closing hours performs best (adjusted *R2* = .17, *p* < .01) but the one for opening hours is not significant.

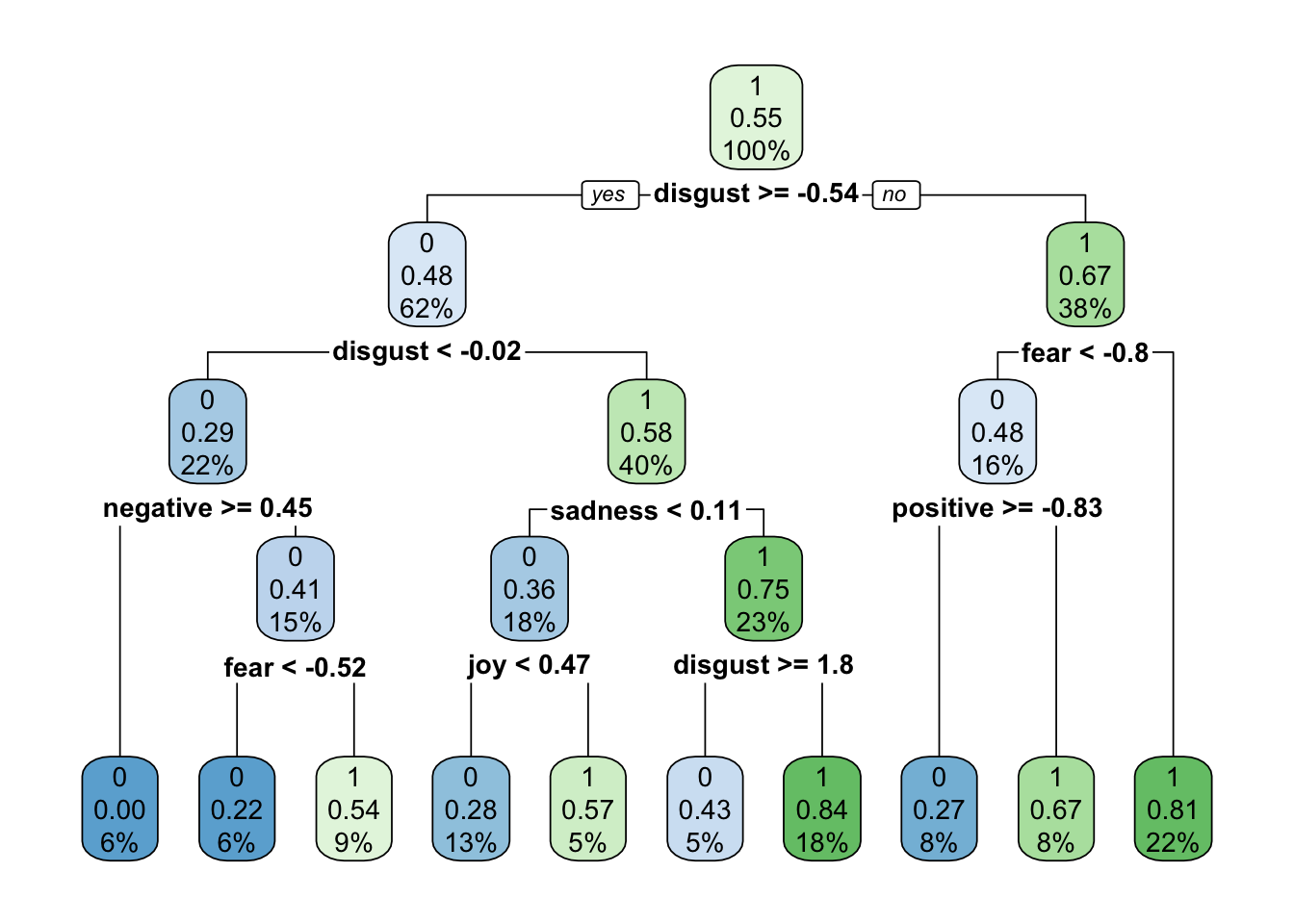
Considering that NRC-based sentiment can provide multiple independent predictors and stock prices of different companies can vary dramatically, we further develop a decision tree approach to capturing the likelihood for a stock to rise (*y* = 1) or fall (*y* = 0). On an hourly basis, a stock is increasing if the price at a time of *T (e.g., 11 a.m.)* is higher than the price at a time of *(T - 1)* (e.g., 10 *a.m.*) while it is decreasing if the price at *T* is lower than an hour ago. This variable indicates stock price movement, which can reflect investors’ earlier bidding behavior and influence their later buying or selling decision.

Using APPL as an example, the decision tree shown in Figure 8 indicates that tweets with positive sentiment greater than or equal to 1.4, which is a normalized sentiment score, is likely to result in an decrease of stock price (*probability* = 0.32). A score of 1.4 is equivalent to a numeric sentiment score of 198.5 (Mean = 99.7, Standard Deviation = 70.6). The remaining tweets, 88% among all, are more likely to relate to an increase of stock price (probability = 0.59). Furthermore, for AAPL those tweets with relatively high scores of anticipation, surprise, and trust as well as low scores of fear, anger, and negative sentiment are more likely to come up with an increase of stock price. The rationale based on AAPL’s decision is easy to interpret and has met our expectation that positive sentiment might increase stock price while negative sentiment might result in a decrease. The decision tree models of other stocks are shown in Figures 9, 10, and 11 respectively.



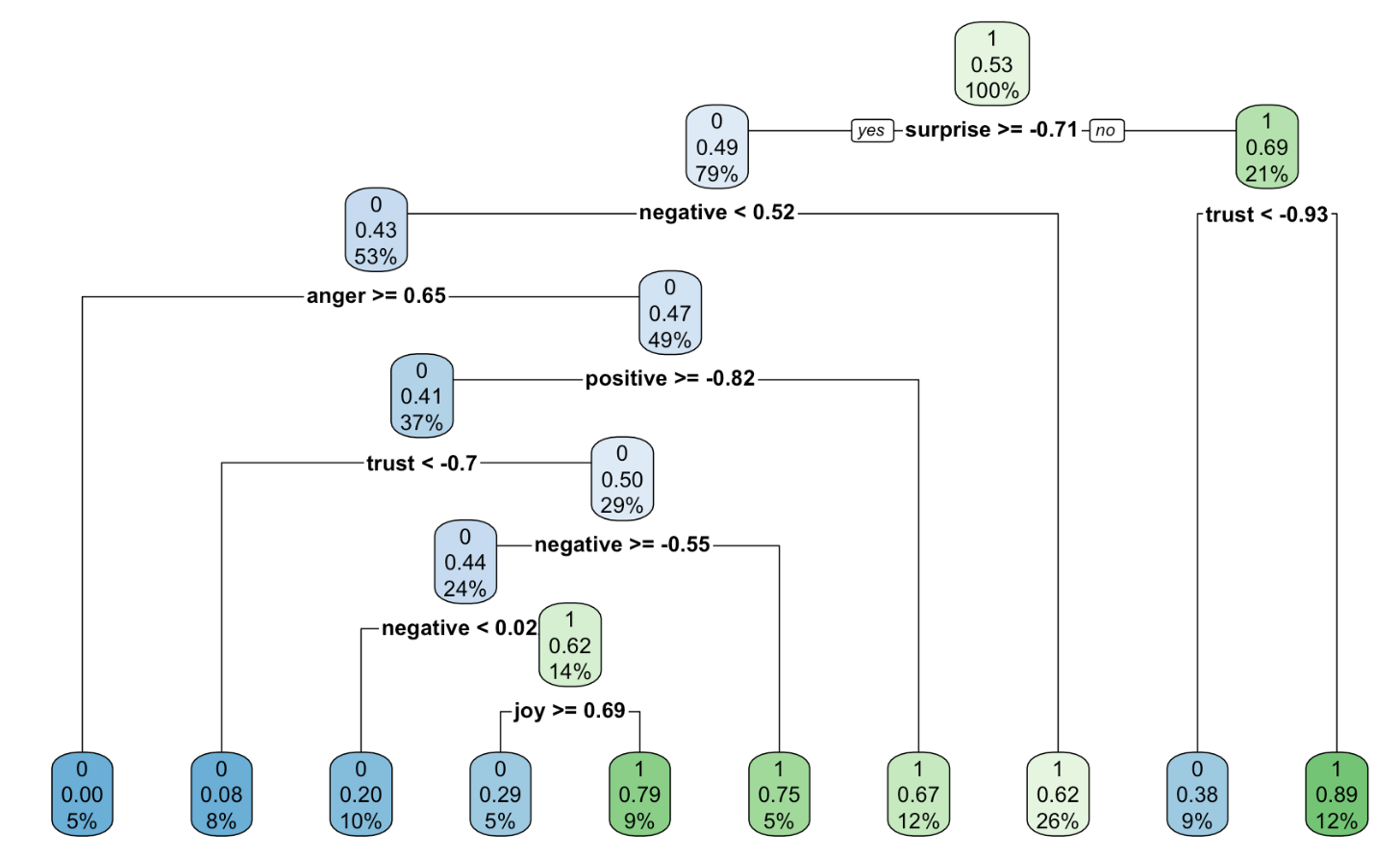
**Apple (AAPL)**

Figure 8. A decision tree model of APPL



**Google (GOOG)**

Figure 9. A decision tree model of GOOG



**Microsoft (MSFT)**

Figure 10. A decision tree model of MSFT

**Netflix (NFLX)**

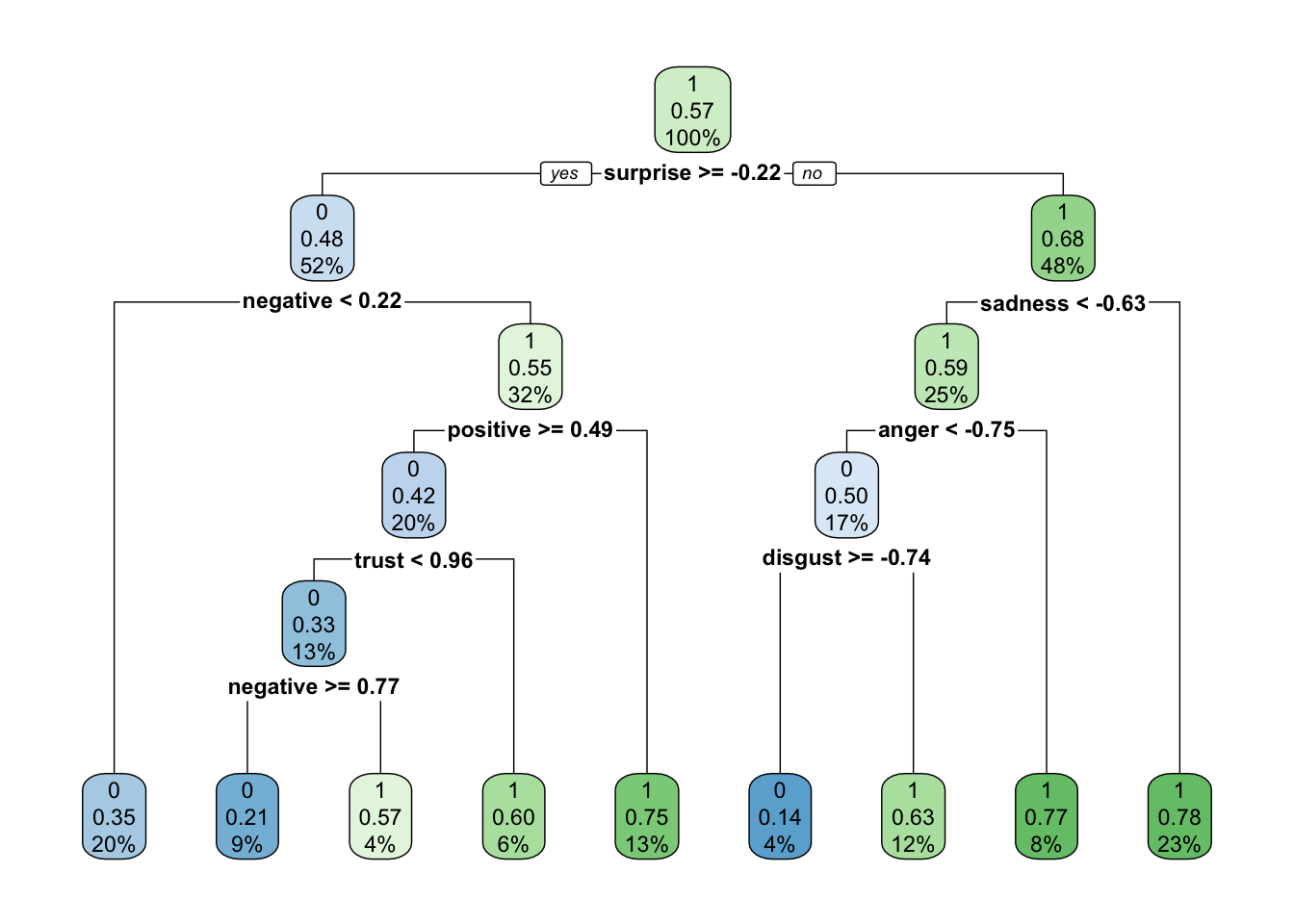


Figure 11. A decision tree model of NFLX

To summarize, the decision tree model performs better than multiple regression model in this work. Based on the data collected from Twitter, we have found that a decision tree model can present acceptable accuracy of prediction of the direction of stock price movement. Unfortunately, data is not available at certain timelines due to the failure of data download using Twitter API, thus, the performance of the proposed model may not always be satisfactory.

1. Conclusion and future work

Predicting future stock prices is a complex but interesting topic for both researchers in Finance and Economics and practitioners in financial industry and related areas. Especially, the connection between public mood, expressed and diffused in social media, and stock performance has received significant attention in recent years. Following Behavioral Finance theory, we have collected empirical evidence to show that public opinions matter to stock price movement. Our results indicate that sentiment analysis based on social media content can be further automated and can produce meaningful results to help investors make better decisions.

In particular, the results of this work suggest that public opinions extracted from stock-related tweets can be used to forecast movements of stock price. In this study, we have used the stocks of Apple (APPL), Google (GOOG), Microsoft (MSFT), and Netflix (NFLX) as examples. The approaches that we have taken, such as using Twitter streaming API for data collection and using existing lexicons for sentiment analysis of social media content, can be easily replicated and smoothly implemented for future research in the related areas.

We propose that a decision tree approach is novel and feasible, as well as easy for investors to interpret which type of sentiment should be addressed and further considered to make a decision of buying or selling stocks. Although it seems to be a common sense that sentiment can influence investors’ opinions and decisions, in this study we demonstrate how sentiment is related to stock performance. Particularly, our results demonstrate that not all the kinds of sentiment are equal. For example, we have found that trust and anticipation are more important than other positive sentiment, in relation to stock price movement. Also, fear and anger are more significantly associated with stock performance.

We can anticipate that some applications can be developed to automatically capture social media sentiment regarding stock price forecasting, so that investors can use these tools to manage the portfolio more effectively and conveniently. In that sense, social media can be not only a communication tool for stock-related information exchange, but also a decision support system that enable investors gather and process information for better forecasting and decision making.

In our future work, we will need to test our model by using a large sample of data that comes from various social media platforms. Also, we are interested to use other lexicons or even for different languages to see how effectively these analytical tools can be used. Last, although using an existing lexicon to extract public opinions from tweets is much convenient, in the future we are hoping to conduct experimental studies, in which human subjects are recruited to read tweets and score sentiment before they estimate the likelihood for a stock to be bullish or bearish as well as their investment decision based on the given information. By doing so, we will be able to compare the automated and human-generated results, in terms of reliability and prediction accuracy, to better understand stock trend and investment behavior.

References:

[1] Malagrino, L. S., Roman, N. T., & Monteiro, A. M. (2018). Forecasting stock market index daily direction: A Bayesian Network approach. *Expert Systems with Applications*, *105*, 11-22.

[2] Zhang, J., Cui, S., Xu, Y., Li, Q., & Li, T. (2018). A novel data-driven stock price trend prediction system. *Expert Systems with Applications*, *97*, 60-69.

[3] Shah, D., Isah, H., & Zulkernine, F. (2018, December). Predicting the effects of news sentiments on the stock market. In *2018 IEEE International Conference on Big Data (Big Data)* (pp. 4705-4708). IEEE.

[4] Wu, D. D., Zheng, L., & Olson, D. L. (2014). A decision support approach for online stock forum sentiment analysis. *IEEE transactions on systems, man, and cybernetics: systems*, *44*(8), 1077-1087.

[5] Das, S., Behera, R. K., & Rath, S. K. (2018). Real-time sentiment analysis of twitter streaming data for stock prediction. *Procedia computer science*, *132*, 956-964.

[6] Oliveira, N., Cortez, P., & Areal, N. (2017). The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices. *Expert Systems with Applications*, *73*, 125-144.

[7] Yang, S. Y., Mo, S. Y. K., & Liu, A. (2015). Twitter financial community sentiment and its predictive relationship to stock market movement. *Quantitative Finance*, *15*(10), 1637-1656.

[8] Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business horizons*, *53*(1), 59-68.

[9] Java, A., Song, X., Finin, T., & Tseng, B. (2007, August). Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis* (pp. 56-65).

[10] Thomson Reuters. (2017). More financial executives forgo business decisions as personal liability seems more likely, Thomson Reuters Culture and Conduct Risk 2017 Survey finds. https://www.thomsonreuters.com/en/press-releases/2017/may/more-financial-executives-forego-business-decisions-as-personal-liability-seems-more-likely.html

[11] Ruiz, E. J., Hristidis, V., Castillo, C., Gionis, A., & Jaimes, A. (2012, February). Correlating financial time series with micro-blogging activity. In *Proceedings of the fifth ACM international conference on Web search and data mining* (pp. 513-522).

[12] Bollen, J., Mao, H., & Pepe, A. (2011, July). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 5, No. 1).

[13] Chen, M. P., Chen, P. F., & Lee, C. C. (2013). Asymmetric effects of investor sentiment on industry stock returns: Panel data evidence. *Emerging Markets Review*, *14*, 35-54.

[14] Chen, K., Li, X., Luo, P., & Zhao, J. L. (2021). News-Induced Dynamic Networks for Market Signaling: Understanding the Impact of News on Firm Equity Value. *Information Systems Research*.

[15] He, W., Guo, L., Shen, J., & Akula, V. (2016). Social media-based forecasting: A case study of tweets and stock prices in the financial services industry. *Journal of Organizational and End User Computing (JOEUC)*, *28*(2), 74-91.

[16] Tirunillai, S., & Tellis, G. J. (2012). Does chatter really matter? Dynamics of user-generated content and stock performance. *Marketing Science*, *31*(2), 198-215.

[17] Das, S. R., & Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management science*, *53*(9), 1375-1388.

[18] Bukovina, J. (2016). Social media big data and capital markets—An overview. *Journal of Behavioral and Experimental Finance*, *11*, 18-26.

[19] Fang, X., & Zhan, J. (2015). Sentiment analysis using product review data. *Journal of Big Data*, *2*(1), 1-14.

[20] Godsay, M. (2015). The process of sentiment analysis: a study. *International Journal of Computer Applications*, *126*(7).

[21] Annett, M., & Kondrak, G. (2008, May). A comparison of sentiment analysis techniques: Polarizing movie blogs. In *Conference of the Canadian Society for Computational Studies of Intelligence* (pp. 25-35). Springer, Berlin, Heidelberg.

[22] Pang, B., & Lee, L. (2008, August). Using very simple statistics for review search: An exploration. In *Coling 2008: Companion volume: Posters* (pp. 75-78).

[23] Russell, M. A. (2011). *Mining the social web: Analyzing data from Facebook, Twitter, LinkedIn, and other social media sites*. " O'Reilly Media, Inc.".

[24] Hu, M., & Liu, B. (2004, August). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 168-177).

[25] Nielsen, F. Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*.

[26] Mohammad, S., & Turney, P. (2010, June). Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text* (pp. 26-34).

[27] Das, S., Behera, R. K., & Rath, S. K. (2018). Real-time sentiment analysis of twitter streaming data for stock prediction. *Procedia computer science*, *132*, 956-964.