Time Series for Forecasting Stock Market Prices Based on Sentiment Analysis of Social Media

Babu Aravind Sivamani, SSN College of Engineering, Chennai, India
Dakshinamoorthy Karthikeyan, SSN College of Engineering, Chennai, India
Chamundeswari Arumugam, SSN College of Engineering, Chennai, India
Pavan Kalyan, SSN College of Engineering, Chennai, India

ABSTRACT

This paper attempts to find a relation between the public perception of a company and its stock value price. Since social media is a very powerful tool used by a lot of people to voice their opinions on the performance of a company, it is a good source of information about the public sentiment. Previous studies have shown that the overall public sentiment collected from sites like Twitter do have a relation to the market price of a company over a period of time. The goal is to build on their research to improve the accuracy of predictions and determine if the public perception surrounding a company is a driving factor of its stock growth.

KEYWORDS

Apple, Correlation, LSTM, Natural Language Processing, Random Forest, Sentiment Analysis, Social Media, Stock Growth Prediction, Stock Market, Time Series Analysis, Twitter, Word2Vec

1. INTRODUCTION

Previous studies (Mankar et al., 2018) on the effect of social media on the stock market have shown that the aggregate public mood towards a company over a short time span has a relation to the closing price of that company at the end of the time span. Studies have been able to utilize data collected from any one reputed social media site (Ex: Twitter, Stocktwits, Weibo, etc.) to produce a model that predicts stock market prices with ~70% accuracy (Acosta et al., 2017). This paper attempts to more accurately gauge the public sentiment of a company from social media websites such as Twitter by implementing time series analysis at minute intervals to find correlations that will likely produce a better stock estimate.

Stock price of a company is determined by a large number of independent traders all over the world. Previous studies have not taken into account the reasons why an individual trader makes the decision to buy or sell. As social media has been shown to offer an insight into the mindset of people, it was realized that the posts online may be an indication of how the market at large is inclined towards
a company. The main objective of this paper is to find whether the public sentiment surrounding a company is able to determine the growth of its stock price. Here in this paper, the company Apple (NASDAQ: AAPL) was selected because it is prominent in the public spotlight and hence ideally suited for an analysis of this kind.

First the selected social media platform is queried for posts in the time period containing any of the keywords in our search term. The search term must be carefully selected to ensure that the number of off-topic posts is limited, while not missing out on any messages with important content. Then any irrelevant posts which passed through the search query are found and filtered out. Data pre-processing procedures such as the removal of non-English characters, stop words, hashtags and user mentions is carried out. Sentiment analysis is performed on the pre-processed text data and each post is classified as positive, negative or neutral corresponding to whether the market for Apple is bullish, bearish or not having any effect. Finally, the aggregate sentiment values from all collected websites will be fed into the model which would use a machine learning algorithm to produce a correlation between the media posts and the stock market price which can then be used to predict the closing market value, given the opening price and overall public sentiment.

The organization of this paper proceeds as follows. Section 2 discusses the literature survey, while Section 3 elaborates on the proposed methodology. Section 4 details with the result and discussion, and Section 5 details the conclusion and future work.

2. LITERATURE SURVEY

Venkata et al.(2016) used Word2vec and N-gram representation of text to train a classifier model to predict the stock market movements and picked Word2vec representation due to its high accuracy in large datasets. Rakhi et al.(2018) collected the sentiment data, and the stock price data to predict stock market price using a support-vector machine (SVM) classifier and observed that if the data size increases the accuracy obtained will also increase. Scott et al.(2017) used smart user classification to filter the tweets by computing scoring weights based on number of likes, number of followers count and how often the user is correct. Further, they used Tf-Idf vectorizer for textual representation and linear regression classifier for the sentiment prediction. Zhaoxia et al.(n.d.) used the sentiments of the news data to predict the stock market price using neural networks.

Sreelekshmy et. al.(2017) applied Recurrent Neural Networks(RNN), Long short-term memory(LSTM) and Convolutional Neural Networks(CNN) - sliding window architecture for stock price prediction of Infosys, TCS and Cipla and concluded that CNN outperforms the other two models in the stock market analysis due to the irregular changes that happen in the stock market. Few works have used the previous stock market data to predict the movements of the stock market while another few used the sentiments from social media to predict the same using SVM, random forest and other machine learning algorithms. Also it is clear that Word2Vec representation of text will be ideal for data that is fed into the neural network layers for building the classifier that predicts the trends of the stock market.

Stock market predictions have become an interesting research area, Correlation of social sentiment data about a company and its stock values there exists research papers that provide solid efficacy to perform a time series analysis on prediction of stock prices and ensemble models that increases the accuracy of the prediction by performing a sentiment analysis on the co-related socio-economic data of that particular company, though its limitation was performed on a 24-hour interval; This research paper extends this notion by performing minute-wise stock price sentiment analysis that gives you a more through window for predicting stock rise and stock fall.
3. PROPOSED METHODOLOGY

Twitter was considered to be the source for the dataset because many companies practice public relations via tweets and also it provided a concrete API with filtering that would prove imperative to the selection criteria for a specified category of text data. The collected twitter dataset will be pre-processed for any missing inconsistencies, and cleaned using our custom data-cleaning libraries. After preprocessing, a subset of that dataset is manually labeled with a sentiment-value. A Random Forest Classifier is used to classify the rest of the sentiment based on the labeled dataset. For stock price prediction, the stock market data was downloaded from Finam and after it undergoes pre-processing, the processed prices dataset along with the labeled sentiment dataset is run through a LSTM model. A graphical overview of the system structure is shown in Figure 1.

3.1. Data Collection

Data collection is defined as “the process of acquiring raw, unprocessed data and storing in a mutable format”. The data collection period was a little over three months, and approximately two million tweets were scraped for the last quarter of 2018. For collection of tweets from Twitter, a python module: ‘TwitterScraper’ was used, it supports querying of the Twitter database with advanced search parameters and operators (Available parameters include followers_count, friends_count, and also the logical operators AND, OR and NOT)(Standard search operators available in the Twitter search query field, n.d.) that limits results to tweets that match our query, as well as additional metadata such as ensuring that the time of the tweet fits inside our selected time period. The exact search query given to the module is ‘apple OR ((bullish OR bearish) AND (AAPL OR apple))’. This query has proven effective in filtering out the majority of completely unrelated tweets from the result set.

The result object returned by this module is a JSON array of tweet objects, where each tweet is a JSON Object with the following fields: username, user id, html, text, likes, retweets, comments, timestamp, profile-picture, profile display-name, etc. An Example of the raw Tweet data is represented in Figure 2.

The fields user_id, text, and timestamp are extracted from the tweets and other unwanted fields are deleted.

There are several services that provide access to historical intraday stock prices for NASDAQ listed companies like Apple (Publicly Available sources of Intra-day stock market data for listed

![Figure 1. Proposed system structure](image-url)
companies, n.d.). Finam (Finam.ru, n.d.) is a Russian website that provides data for the stock, futures, ETF and Forex markets for research and analysis purposes. The data is available for only very highlight capitalized securities, however for these one can avail several months worth of tick data. A representation of the finam stock dataset is shown in Figure 3.

3.2. Data Pre-Processing

Data preprocessing is a technique which is used to transform the raw data in a useful and efficient format. In this section unnecessary data or noise is removed from the raw text twitter data. Firstly, the raw text data is converted to lower-case. Secondly, text data which contain words that begin with #(hashtags), @ (user mentions) are simply replaced with the actual word content of the hashtag and username. Thirdly, long URLs are replaced with just the domain name of the URL. For example, https://techcrunch.com/2019/10/19/the-new-iphone-is-ugly/ is replaced with techcrunch. The identification of these words is implemented through regex matching. Then the special symbols like non-english characters are removed.

The final step in pre-processing of text is stop-word removal, which is the removal of words in the text that do not contribute to the overall meaning of the post. Examples of such words include

|   | A  | B     | C       | D     | E    | F     | G    |
|---|----|-------|---------|-------|------|-------|------|
| 1 | <DATE> | 20180926 | <TIME>     | 100100 | <OPEN>   | 161.26 | <HIGH>  | 161.3 | <LOW>   | 160.74 | <CLOSE> | 160.88 | <VOL> | 267350 |
| 2 |     | 20180926 |          | 100200 | 160.91 | 160.98 | 160.52 | 160.76 | 163520 |
| 3 |     | 20180926 |          | 100300 | 160.67 | 160.86 | 160.65 | 160.84 | 40670  |
| 4 |     | 20180926 |          | 100400 | 160.75 | 160.88 | 160.72 | 160.87 | 34280  |
| 5 |     | 20180926 |          | 100500 | 160.84 | 160.86 | 160.64 | 160.7 | 61070  |
| 6 |     | 20180926 |          | 100600 | 160.65 | 160.78 | 160.64 | 160.67 | 51490  |
| 7 |     | 20180926 |          | 100700 | 160.65 | 160.68 | 160.49 | 160.49 | 69370  |
| 8 |     | 20180926 |          | 100800 | 160.53 | 160.54 | 160.37 | 160.37 | 48950  |
| 9 |     | 20180926 |          | 100900 | 160.48 | 160.48 | 160.15 | 160.15 | 113360 |
| 10|     | 20180926|          | 101000 | 160.15 | 160.45 | 160.15 | 160.45 | 103510 |
a, an, the, I, for, etc. The text of each post is tokenized and compared with any publicly available curated list of stop words (Curated list of English stop-words extracted from Python’s NLTK library, n.d.). The above preprocessing steps were repeated for the remaining two million raw text data. The data preprocessing outcome of an instance is displayed in Table 1.

### 3.3. Sentiment Analysis Module

After collecting a large twitter dataset, sentiment analysis is performed on the text. For this purpose, the library Word2vec (Mikolov, 2013) is used, which is an advanced Natural Language Processing (NLP) technique for mapping words to a vector representation of any dimension. A 200-dimension vector is used for generation in this case. When run on the dataset of text, Word2vec will generate a unique vector for every word in the dataset which will exactly preserve the context of the words and the relation between similar meaning words in vector space. Then the word2vec representations along with around 15000 messages manually labelled as positive(1), neutral(0) or negative(2). An android app was developed with google’s firebase backend systems. The app was distributed to a group of trained people for labeling the tweets. The app contains three buttons for inputting the sentiments below the text data. An instance of the tweet in the app is shown in Figure 4.

| Before data pre-processing | #Apple could have made all their products’ current design years ago, now their products are now left behind by @Samsung and others http://gizmo.do/Twg8I8i |
|----------------------------|-----------------------------------------------------------------------------------------------------------------------------------------|
| After data pre-processing  | apple could made products current design years ago products left behind samsung others gizmodo                                          |

**Figure 4. An instance of app tweet data with sentiment labels**

![Tweet labeler](thumbnail.png)
The output of the app tweet data is then stored within a firebase datastore along with the labeled sentiment. This is highlighted in Figure 5.

The manually labeled 15,000 tweets are first split into a training and validation set to train the random forest classifier. On the validation set, the random forest reached a precision score of 90%, Recall of 88% and F1 score 90%. The trained model is used to predict the sentiments for all the two million tweets in the datastore. To determine the stock price at a future point of time, the random forest classifier predicts social sentiments, and calculates the total sentiment for each one-minute interval as the number of positive minus negative sentiments. A sample output for the random forest classifier is shown in Figure 6.

**Figure 5. An instance of labeled tweet data**

```markdown
```

**Figure 6. Five samples of random forest classifier output**

| text                           | timestamp   | sentiment |
|-------------------------------|-------------|-----------|
| liveblog news apple making brooklyn event bit ... | 2018-10-27 14:55:00 | 1         |
| spotify apple music           | 2018-11-16 19:56:08 | 0         |
| ready upgrade tv room ultimate home theater en... | 2018-11-08 21:22:05 | 0         |
| technews trending apple launches portal users ... | 2018-10-17 18:14:20 | 1         |
| depends type apple watch want come different s... | 2018-12-25 20:40:32 | 1         |
4. RESULTS AND DISCUSSION

A time series forecasting method using LSTM is used here as both the social media posts as well as the financial stock price dataset has a time component. LSTM was trained using Google’s open source TensorFlow libraries which comes with an implementation of LSTM. As the dataset contains two distinct features namely the sentiment and the stock price, a multivariate version of the dataset was created for training. In this process, overlapping sliding windows of length 720 minutes (12 hours) are applied on the dataset. To get a single step dataset, the starting point of the window is set to the point immediately after the start of the previous window, i.e. a new window begins every minute. Finally the dataset contains 35634 minutes of data, which is divided into a testing and validation set (80-20% ratio).

The first prediction of the trained LSTM model at a future time from the validation set is shown in Figure 7. The blue line shows a subset of the stock price history passed as input to the LSTM. The green circle represents the price predicted by the model (151.79), whereas the red cross shows the actual value that came to pass at that point of time (151.93). From Figure 7, it is observed that the model is accurately predicting the sentiment of the tweet.

The second prediction of the trained LSTM model at a future time from the validation set is shown in Figure 8. The predicted price was 152.81 whereas the observed actual price was 152.15. From Figure 8, it is observed that the model is accurately predicting the sentiment of the tweet.

Figure 7. First sample prediction of LSTM

![Figure 7. First sample prediction of LSTM](image)

Figure 8. Second sample prediction of LSTM

![Figure 8. Second sample prediction of LSTM](image)
The third prediction of the trained LSTM model is shown in Figure 9. The predicted price was 151.75 whereas the observed actual price was 151.25. From Figure 9, it is observed that there is a small deviation between the model prediction and the actual value. Here we see a rare instance of inconsistencies with the prediction.

Though there are minor deviations in some single step predictions, this process usually consists of a large sample of 60 single step predictions which allows the algorithm to make an accurate prediction overall, marginally on the error of 0.5.

It is observed from Figure 10 that as the step size and history size decreases, the Mean Squared Error (MSE) also reduces, indicating an improvement in the accuracy rate (Table 2).

The above data is taken from a random portion of the validation set. It is observed that the LSTM is able to learn and accurately predict with errors less than 1 in the vast majority of cases at each minute interval. The system is well trained over various curves, rises, falls over the last quarter of 2018.

Though there are a multitude of factors which could determine the exact values of stock prices, based on the outcome of the system. The public image of the company seems to be one of the driving forces. The LSTM Neural Network was accurate enough to forecast the stock values, which indicates the existence of a correlation between the sentiments and stock prices. This suggests that the public image of a company has a bearing on the market performance of a company. The visualisation found between the stock prices and sentiments are shown in the Figure 11. The y-axis for the blue line shows the market price and the total sentiment is shown in the orange graph. The x-axis represents the time component, i.e. the number of minutes since the start of Q4 2018.
5. Conclusion and Future Work

In this paper, the rise and fall of stock prices were predicted at every minute interval. Intraday (1 minute intervals) stock market data on Apple Inc. (NASDAQ: AAPL) was collected for Q4, 2018. Twitter was scraped to find all tweets related to Apple over the same time period. Total sentiment in each one-minute interval was calculated and combined with market price histories to forecast the future prices.

The accuracy of the sentiment analysis model could be improved by performing more manual labelling of tweets to improve the quality of training dataset for sentiment analysis and by exploring other ML models for classification of tweets. The usage of streaming APIs could result in a narrower time frame for the time series analysis, possibly for each continuous second.

Table 2. Stock prediction and error

| Date   | Time | Actual Price | Predicted  | Error   |
|--------|------|--------------|------------|---------|
| 20181218 | 155200 | 151.42       | 151.68074  | -0.26074 |
| 20181218 | 155300 | 151.27       | 151.69843  | -0.42843 |
| 20181218 | 155400 | 151.44       | 151.9058   | -0.4658  |
| 20181218 | 155500 | 151.49       | 151.82811  | -0.33811 |
| 20181218 | 155600 | 151.27       | 151.72792  | -0.45792 |
| 20181218 | 155700 | 151.33       | 151.73984  | -0.40894 |
| 20181218 | 155800 | 151.35       | 151.96754  | -0.61754 |
| 20181218 | 155900 | 151.34       | 151.75021  | -0.41021 |
| 20181218 | 160000 | 151.36       | 151.98468  | -0.62468 |
| 20181218 | 160100 | 151.35       | 151.85138  | -0.50138 |
| 20181218 | 160200 | 151.35       | 151.78558  | -0.33558 |
| 20181218 | 160300 | 151.46       | 151.78085  | -0.32085 |
| 20181218 | 160400 | 151.55       | 151.88606  | -0.33606 |
| 20181218 | 160500 | 151.61       | 151.85497  | -0.24497 |
REFERENCES

Acosta, Lamaute, Luo, Finkelstein, & Cotoranu. (2017). Sentiment Analysis of Twitter Messages using Word2Vec. In Proceedings of Student-Faculty Research Day, CSIS. Pace University.

Curated list of English stop-words extracted from Python’s NLTK library. (n.d.). https://gist.github.com/sebleier/554280/raw/7e0e4a1ce04c2bb7bd41089c9821d6cfd60c786c/NLTK’s%2520list%2520of%2520english%2520stopwords

Finam.ru - A website that provides several months of tick data for highly capitalized securities. (n.d.). https://www.finam.ru/profile/stock-market/stocks

Mankar, T., Hotchandani, T., Madhwani, M., Chidrawar, A., & Lifna, C. S. (2018). Stock Market Prediction based on Social Sentiments using Machine Learning. 1–3. Advance online publication. doi:10.1109/ICSCET.2018.8537242

Mikolov, T. (2013). Efficient Estimation of Word Representations in Vector Space. Academic Press.

Publicly Available sources of Intra-day stock market data for listed companies. (n.d.). https://www.quantshare.com/sa-636-6-new-ways-to-download-free-intraday-data-for-the-us-stock-market

Rakhi, B., & Sher, M. D. (2018). Integrating StockTwits with Sentiment Analysis for better Prediction of Stock Price Movement. IEEE International Conference on Computing, Mathematics and Engineering Technologies – iCoMET.

Scott, C., & Preveen, M. (2017). Forecasting Stock Prices using Social Media Analysis. IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress, 1031-1038.

Selvin, Vinayakumar, Gopalakrishnan, Menon, & Kp. (2017). Stock price prediction using LSTM, RNN and CNN-sliding window model. 10.1109/ICACCI.2017.8126078

Standard search operators available in the Twitter search query field. (n.d.). https://developer.twitter.com/en/docs/tweets/rules-and-filtering/overview/standard-operators

Venkata, S. P., Kamal, N. C., Ganapati, P., & Babita, M. (2016). Sentiment Analysis of Twitter Data for Predicting Stock Market Movements. International conference on Signal Processing, Communication, Power and Embedded System, 1345 – 1350.

Wang, Z., Seng-Beng, Ho., & Lin, Z. (n.d.). Stock market prediction by incorporating social media news as sentiment. https://ieeexplore.ieee.org/document/8637365

Pavan Kalyan completed their bachelors at SSN College of Engineering. They are excited to make computers more and more intelligent. This made them develop a huge interest on deep learning. They feel there is always something or another left to explore in this field, and they are always ready to learn and adapt to new technologies.