Stock Price Movement Using News Analytics

Wolves of 10^{th} Street
Aditya Aggarwal, Anna M. Riehle, Emily T. Huskins,
Manish Mehta, Ravi P. Singh and Sudhanshu R. Singh

December 06, 2018

1 Introduction

Stock price trend prediction is an active research area, as more accurate predictions are directly related to more returns in stocks. It is a well-established fact that stock prices are driven by the market sentiment [1]. It is also reasonable to hypothesize that news does have a role in shaping this sentiment. Validating this hypothesis and making use of the news for stock price prediction can have tremendous repercussions for the finance world. Apart from the obvious benefits to the investors and the analyst community, this technique has the potential to be of immense utility to firms to protect the interests of their shareholders. While this field has attracted significant interest, a model with reasonable accuracy still evades the seekers. The primary motivation for this project comes from the challenge hosted by the company Two-Sigma on Kaggle [2]. We aim to develop a model which can predict the movement in stock prices using news sentiments and financial market data.

2 Literature Survey

In recent years, significant efforts have been put into developing models to predict the future trend of a particular stock or overall market trend. Research has shown that there is a strong relationship between a news article about a company and that company's stock price fluctuations.

In their research, Nagar and Hahsler[3] presented an automated text mining based approach to aggregate news stories from various sources and create a News Corpus. Yu et al[4] present a text mining based framework to determine the sentiment of news articles and illustrate their impact on energy demand. Recent research[5] has also studied the impact of relevance level to financial forecasting.

Studies have shown that earlier stock prices aren't good predictors for future stock prices[6] and that data obtained can be used for predicting the slope of short-term stock price movements [7], [8]. Further researches and advancements suggested that in addition to building Concept Maps and checking for correlation between stock prices and sentiments[4], one can come up with finer predictive models[9], such as time-series based regression[10][11], Artificial Neural Networks[12], and Dual Sentiment Classifiers and Predictors[13].

Traditional algorithms such as decision tree (DT) [14], logistic regression(LR) [15], Naive Bayes (NB) [16][17] , artificial neural network(ANNs) [18][19], and support vector machines(SVMs) [19] have been shown to be effective for financial forecasting.Technical indicators were used to predict daily maximum and minimum of the stock prices using multilayer perceptron(MLP) classifier [18] and SVM [20]. Weng et al. [20] combined stock market data with crowdsourced data obtained from Wikipedia and Google News to predict the daily stock movements.

For more stabilized and detailed data, researchers have reported stock price movements with 81.27% accuracy using Artificial Neural Networks (ANN)[21] and 78.81% accuracy using Random Forests[22]. However, use of such machine learning algorithms on news sentiment data to predict stock price movement remains largely unexplored. The focus of this project is to implement various machine learning techniques to identify the impact of news sentiments on stock price movements.

3 Motivation

The focus of this project is to implement various machine learning techniques to identify the impact of news sentiments and features of news articles on stock price movements. Through our project, we hope to answer the following questions:

- How does news sentiment impact stock price movement?
- How powerful of a predictor are the features of news articles for stock price movement?
- What are the most important news article features for predicting stock price movement?
- For how long do news articles impact stock price movement?

4 Data Collection

Our data is primarily collected from the challenge hosted on Kaggle by an investment management company, Two Sigma. The data provided contains daily information on 3,510 stocks and is divided into market data provided by Intrinio, and daily news data provided by Thomson Reuters. The market data originally consisted of 4,072,955 rows and 16 columns, including such features as opening and closing stock prices and stock volume. The news data consisted of 9,328,749 rows and 35 columns and contained the results of sentiment analysis done on news items as well as other features of the news items such as word and sentence counts. The news dataset was so much larger than the market dataset because there were typically multiple news items per stock on any given day.

5 Data Cleaning

The data originally spanned from 2007-2016. We limited it to 2010-2016 to avoid impact from the Financial Crisis of 2008. We used the data from 2010-2015 for training/validation and the data from 2016 for testing.

The majority of our data cleaning and feature engineering was done primarily on the news data. We calculated the delay between when a news item was created and when it became available and removed rows where the delay was one day or more. We also removed rows with blank headlines, as we believed these would provide little value. Finally, we removed rows with an urgency of 2, as these comprised less than 0.001% of the data and could be easily ignored.

6 Feature Engineering

The news data contained an Urgency column, indicating whether news items were alerts (1) or articles (3) and a Sentiment Class column indicating overall sentiment toward a stock (1 for positive, 0 for neutral, or -1 for negative). Since the feature set contains categorical and numerical variables and need to be aggregated to merge with stock price data, we classed certain features of the news data on urgency and sentiment class to create features which can be aggregated and hence minimizes the information loss. the format is as as follows: numericalVar1_categoricalVar1_categoricalVar2.
For Example: wordCount_3_1 represents the word count of news item with urgency being article(3) and sentiment being positive(1). We performed this transformation on the original features as well as on the derived feature set:

- wordCount: The number of words in the news item
- sentenceCount: The number of sentences in the news item
- sentimentWordCount: The number of words in a news item that are relevant to an asset
- firstMentionSentence: The first sentence in which an asset is mentioned
- relevance: A number between 0 and 1 indicating the news item's relevance to the asset
- rel_FirstMention: Where in a news item a stock, or asset, was first mentioned. (For example, a rel_FirstMention value of 0.5 indicates that the asset was first mentioned halfway through the article.)
- rel_SentCount: What proportion of the words in a news item had sentiments affecting a given stock

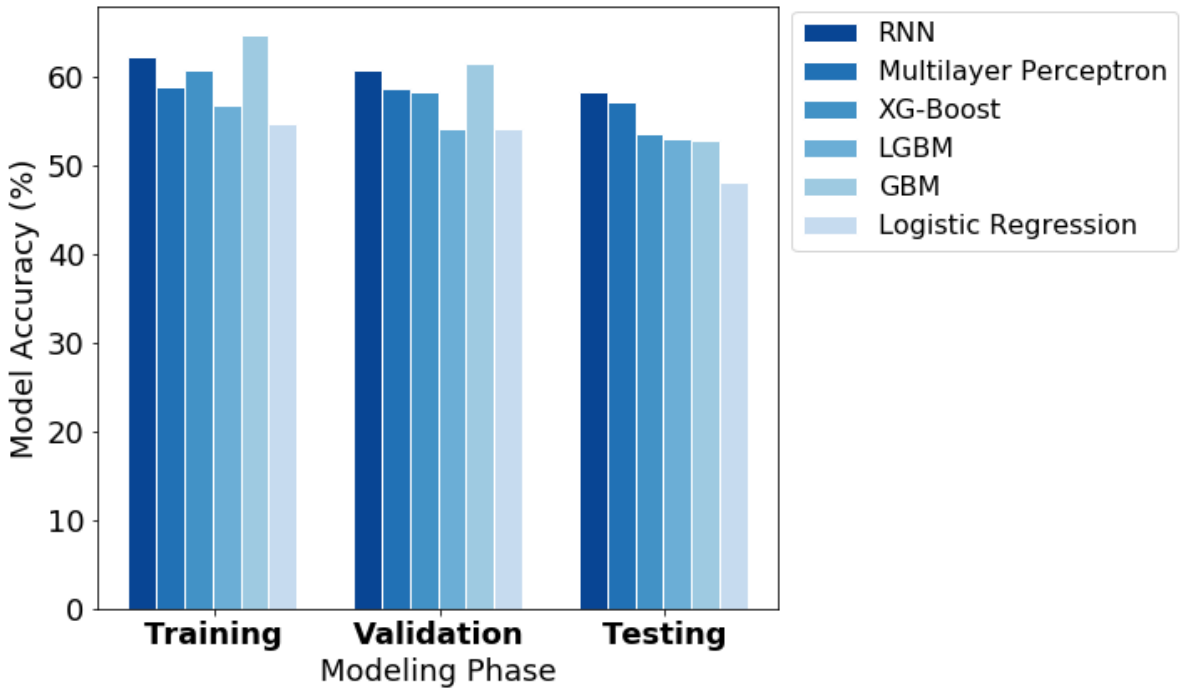
To keep the effect of news data to minimal, we only used the following two feature information from the market data to train the model:

- returnsOpenPrevMktres1_dir: Direction of change of the market adjusted stock price during last day
- returnsOpenPrevMktres10_dir: Direction of change of the market adjusted stock price during previous 10^{th} day

We grouped the news data by asset code and date to get mean features of all news items at the daily level, then merged the transformed news data with our daily market data to create a merged dataset with 2,606,934 rows and 73 columns.

7 Initial Models

We used a binary response to indicate whether or not stock prices would go up 10 days after the day of observation and explored a variety of classification models for predicting this response.
Initially, we used all the features from both the market and news data to build our model. However, we quickly realized that all the top features in these models were market features. While this is intuitive, it defeats the purpose of our project, which is to predict stock price movement using news data. Hence, we removed most of the features from the market data, keeping only our two engineered features. We used these two features along with all the news features to develop our model.

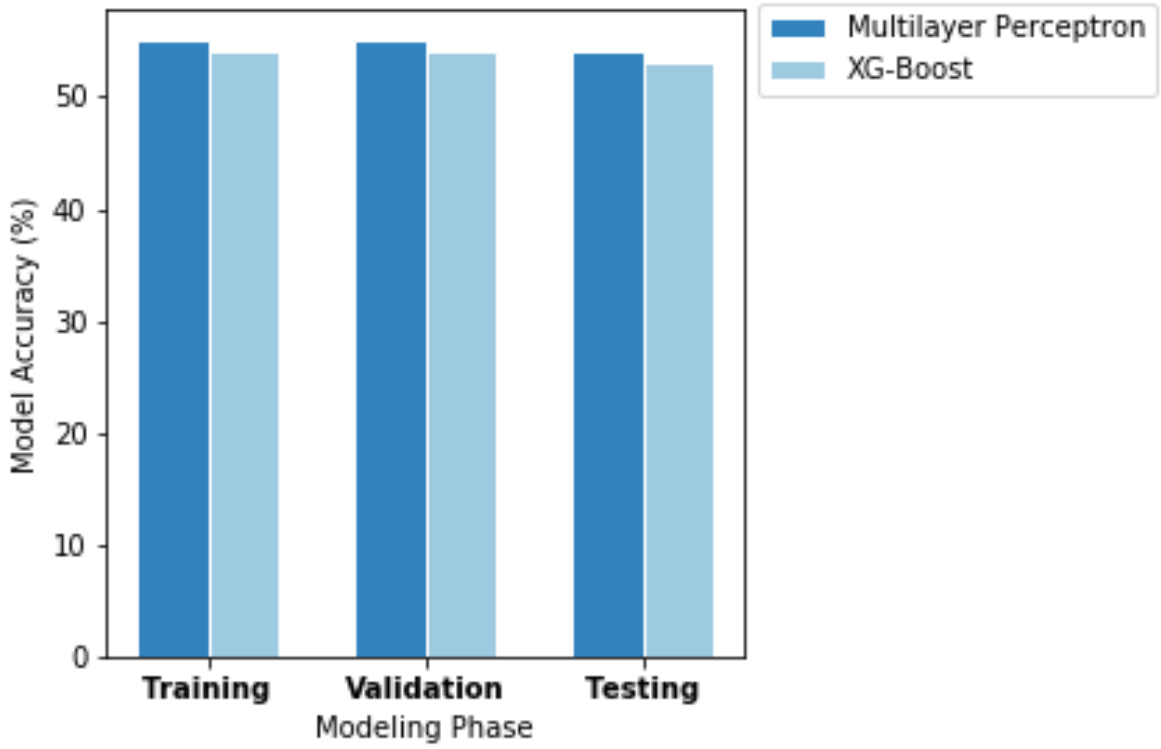


While we briefly explored logistic regression, we found that neural networks and gradient boosting machines were much more powerful for predicting stock price movement. We therefore focused primarily on neural networks and gradient boosting machines for our project.
Within neural networks, we looked primarily at recurrent neural networks (RNNs) and multilayer perceptrons (MLPs). We got similar accuracies of around 54% for both these models. Based on research, we know that MLPs are better for working with tabular data and predicting time series and classification. We therefore chose MLPs as our primary neural network.
Within Gradient Boosting Machines, we looked at Light Gradient Boosting Machines (LGBM) and Extreme Gradient Boosting Machines (XG-Boost). LGBM has a very fast training speed and is therefore particularly well-suited for working with large datasets. It uses leaf-wise splitting, rather than the level-wise splitting done by other gradient boosting machines, and can therefore be prone to overfitting. [25] In our project, we found that LGBM predicted positive stock price movement for around 94% of stocks. However, prediction for negative stock price movement was true only for 2% of the stocks.
While not as fast as LGBM, XG-Boost is still relatively fast when compared to other

Gradient Boosting Machines. It is particularly useful for regularized and predictive classification. [26] On our data, the XGBoost model gave stable accuracies of approximately 54% across train and test data. In addition, we had good prediction for both negative (TNR) and positive (TPR) stock price movement for around 56% of stocks.

8 Final Model: XG-BOOST

Because they are based on decision trees rather than hidden layers, Gradient Boosting Machines are much more explainable and interpretable than Neural Networks. They are also faster to fit than Neural Networks. For these reasons, and because it provided testing accuracy roughly equivalent to that of MLP, we chose to move forward with XG-Boost as our primary model.



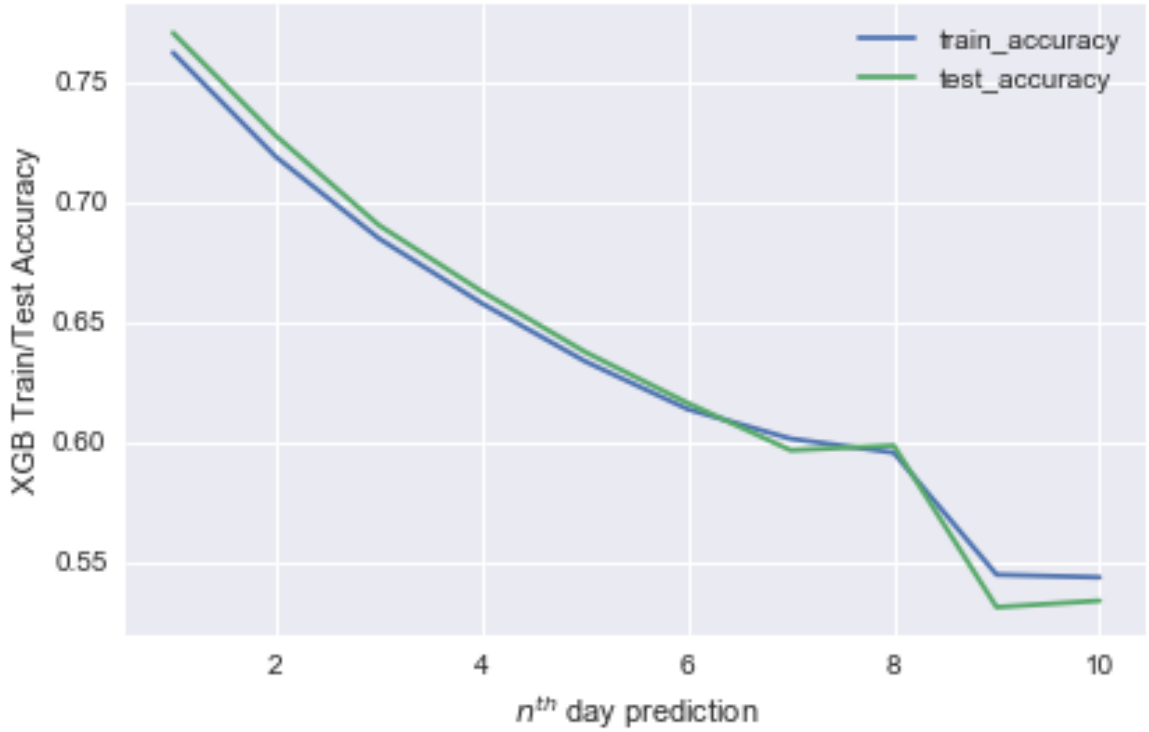
Since we had time series data, we used TimeSeriesSplit from scikit learn to do cross-validation with 5 folds inside RandomizedSearchCV to find the best set of hyperparameters. TimeSeriesSplit helps to do cross validation with time series data by maintaining temporal restrictions while creating folds. We found that the optimal hyperparameters were:

- **n_estimators**(Number of trees to build): 5000
- **max_depth**(Maximum depth of a tree): 8
- **subsample**(Percent of data to sample before building a tree): 0.7
- **colsample_bytree**(Percent of columns to sample before building a tree): 0.8
- **min_child_weight**(Minimum sum of instance weight needed in a child): 10

- **gamma**(Min loss reduction required to make a further partition on a leaf node of a tree): 2
- **reg_lambda**(Regularization term on weights determining how conservative model is): 1
- **reg_alpha**(Regularization term on weights determining how conservative model is): 2
- **learning_rate**(Value used to prevent overfitting in model): 0.01

9 Model Evaluation

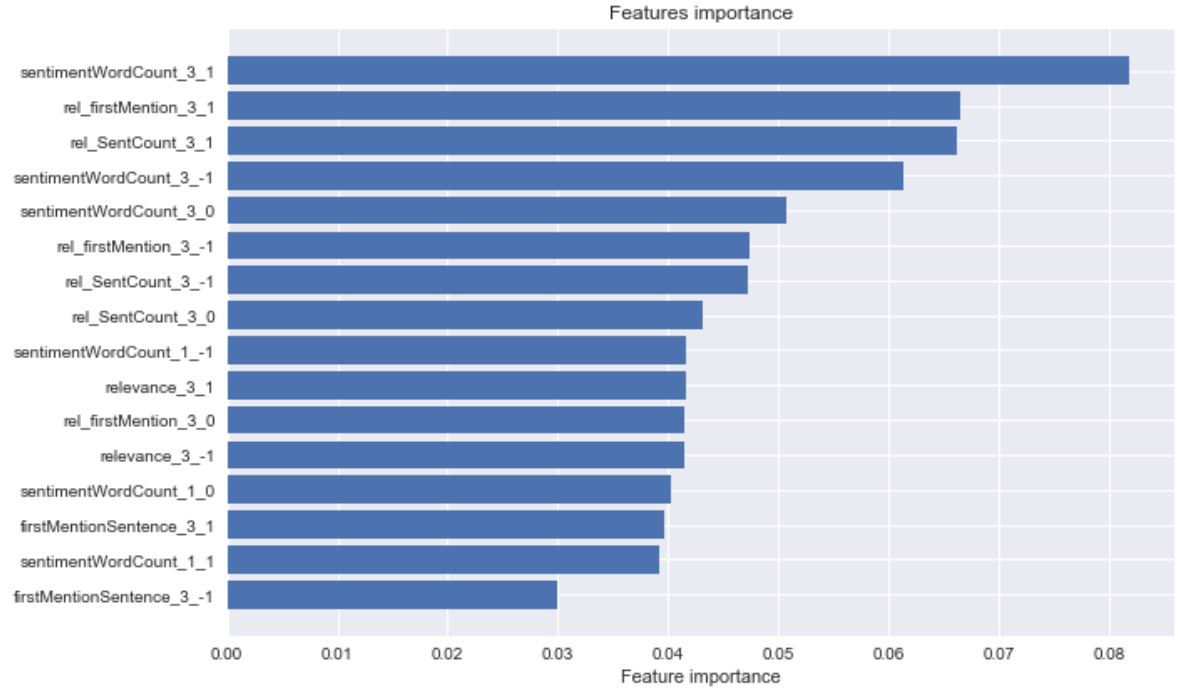
In keeping within the scope of the Kaggle competition, we began by attempting to predict stock price movement 10 days in advance. For 10-day prediction using XG-Boost, we got roughly equal proportions of True Positive, True Negative, False Positive, and False Negative values. We also got a True Positive Rate of 55%, a True Negative Rate of 53%, and an overall Accuracy of 54%.
We wanted to see whether shortening the window of prediction would improve our model's predictive power. We therefore built a series of models to predict 1-10 days in advance. We found that our model's next-day predictive accuracy was approximately 75%, and that the predictive accuracy decreased from there.



10 Feature Importance

Like Random Forests, Gradient Boosting Machines allow us to easily plot feature importance. We chose to look exclusively at how news features impacted stock price

movement because we found that when we included historical stock market data, the market features overwhelmed the news features. Using solely the news data, we found the following feature importances:

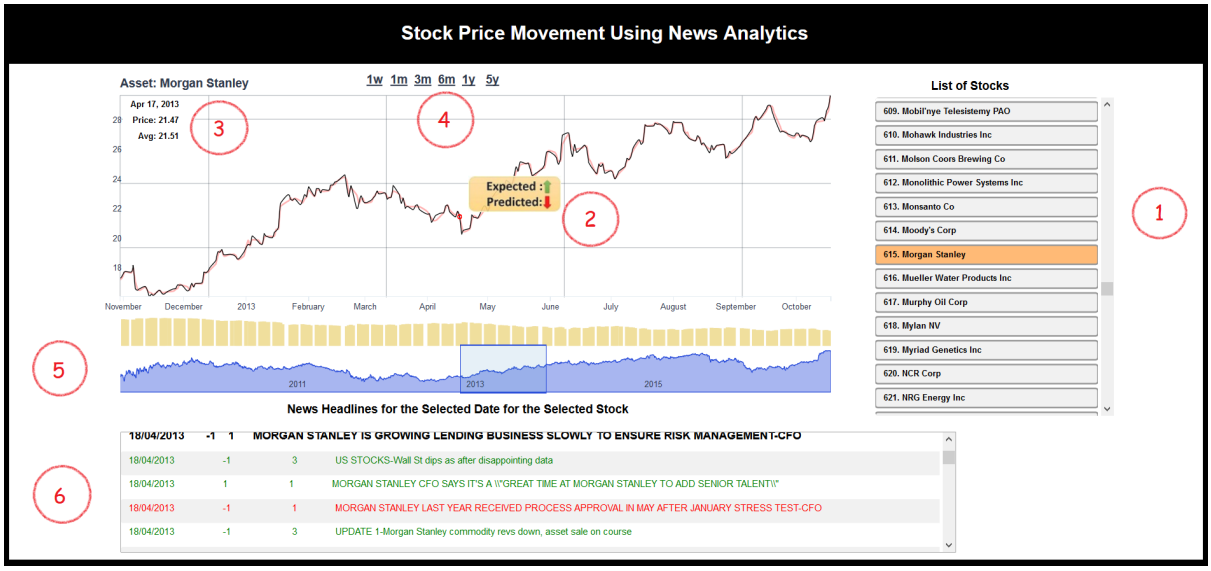


From the above two charts, we could conclude the following:

- People respond to articles more than to alerts.
- Negative sentiment has a stronger impact on the stock market than positive sentiment.
- Individual news items have an impact on stock price movement for short periods, but that impact fades with time.
- The proportion of words or sentences in a news item that relate to an asset are more important than the actual word or sentence counts.
- The relative position of where an asset first appears in a news item is more important than the actual sentence number of that first appearance.

11 Innovative Visualization

An innovative part of our project is the construction of an interactive, Flask-based web application in D3. Most research into stock price movement done to-date has included little if any visualization. We built an interactive web application to show stock price and volume over time. Our app also shows which news items impact stock price movement on a particular date.
Our application consists of two pages: a front page that introduces our topic, explains how to use the app, and provides information on news features deemed important in modeling; and a separate page for our app itself.



The app contains six sections:

1. List from which the user can select a stock of interest
2. Our app lets you easily see stock price movement trend and the tooltip tells the Expected and Predicted stock price movement. The Predicted arrow shows the direction our model expected the stock to move. The Expected arrow shows what actually occurred.
3. The ticker on the left lets the user to track the actual price and the volume of the selected stock on a particular date. The date changes as the user moves the cursor over the line plot.
4. The legend at the top provides the user to choose the time interval for which he is interested to see the stock price pattern.
5. The blue box allows the user to select the custom time window and also to scroll over it.
6. Click on a date of interest in the top graph to see a list of news items from that day related to your chosen stock. The app shows the sentiment value with 1 being positive, -1 being negative, and 0 being neutral, urgency of the news item whether it is an article (3) or an alert (1) and the color of the news shows the positive (Green) and negative (Red) impact on stock price movement.

12 Project Evaluation

Given time restrictions, it was impossible for us to evaluate how well our project would do in the real world. However, we plan to roll this out to investors. Our model provides information on whether the price of a particular stock will increase or decrease in the next 10 days. Based on this information, investors can make their decision of buying or holding the stock. How our model's predictions compare to actual stock price movements over 10 days and how much money an investor is able to make by using our model will provide real-world evaluation for our project.

13 Conclusions and Future Research

Currently, most research focuses on predicting stock price movement based solely on historical stock market data. Earlier research using this method has produced accuracies of around 78% for next-day prediction. Our model, using only news data, gives an accuracy of 75% for next-day prediction, verifying that news plays an important role in stock price movement. Additionally, since news data has an effect for longer durations than random fluctuations in market stock price data, this model could be useful to investors to make relatively better long-term investment decisions.
Areas for future research could include seeing whether keywords within articles impact stock price movement or whether certain news sources have greater influence than others. For this, we could use the News API. Additionally, our data currently stops at the end of 2016. It could be interesting to see how our model fares with more recent articles and stock market data.

14 Distribution of Team Member Effort

All team members contributed equally. Specific contributions are detailed below:

Team Member	Contributions and Plan of Activities
Aditya Aggarwal	Data collection, feature engineering , RNN and NN implementation, , model integration and database setup for flask app
Anna M. Riehle	Design and development of poster, html, report, proposal and slide deck; scheduling and room reservations
Emily T. Huskin	Proposal delivery, report drafting
Manish Mehta	Feature Engineering, Neural Network and XGBoost implementation, visualization and poster design ideas
Ravi P. Singh	Data-Preprocessing, Feature-Engineering, GBM and XG-Boost mode implementation, poster design ideas
Sudhanshu R. Singh	Design and development of Flask app, d3 visualisation SQL database integration with the app, report drafting

References

[1] T. J. Dorsey, *Point, Figure & Charting.* Wiley Tradingr, 2007.

[2] T. Sigma, "Two Sigma: Using News to Predict Stock Movements." `https://www.kaggle.com/c/two-sigma-financial-news`, 2008. [Online; accessed 01 November 2018].

[3] M. H. Anurag Nagar, "Using text and data mining techniques to extract stock market sentiment from live news streams," *IPCSIT vol. XX (2012) IACSIT Press, Singapore*, 2012.

[4] B. L. W.B. Yu and B. Guruswamy, "A theoretic framework integrating text mining and energy demand forecasting," *International Journal of Electronic Business Management. 5(3): 211-224*, 2011.

[5] S. C. A. B. Yauheniya Shynkevich, T.M. McGinnity, "Predicting stock price movements based on different categories of news articles," *IEEE Symposium Series on Computational Intelligence*, 2015.

[6] S. V. Bo Pang, Lillian Lee, "Thumbs up? sentiment classification using machine learning techniques," *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2002.

[7] N. R. S. Steven L. Heston, "News vs. sentiment: Predicting stock returns from news stories," *Financial Analysts Journal, Vol. 73, No. 3, 2017*, 2017.

[8] W. W. Kin Yip Ho, "Predicting stock price movements with news sentiment: An artificial neural network approach," *Artificial Neural Network Modeling*, 2016.

[9] L. L. Bo Pang, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval, Vol. 2, No 1-2*, 2008.

[10] K. S. T. H. Nguyen, "Topic modeling based sentiment analysis on social media for stock market prediction," *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th InternationalJoint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, pp. 1354-1364*, 2015.

[11] A. M. . A. Goel, "Stock prediction using twitter sentiment analysis," n.d.

[12] J. X. Y. Gabriel Pui Cheong Fung and W. Lam, "News sensitive stock trend prediction," *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2002.

[13] S. C. Spandan Ghose Chowdhury, Soham Routh, "News analytics and sentiment analysis to predict stock price trends," *International Journal of Computer Science and Information Technologies, Vol. 5*, 2014.

[14] D. D. Basti E, Kuzey C, "Analyzing initial public offerings' short-term performance using decision trees and svms.," *Decis Support Syst; 73: 15-27.*, 2015.

[15] S. S. Dutta A, Bandopadhyay G, "Surveying stock market forecasting techniques–part ii: Soft computing methods," *International Journal of Business and Information; 7: 105-136.*, 2015.

[16] V. K. Atsalakis GS, "Prediction of stock performance in indian stock market using logistic regression.," *Expert Syst Appl; 36: 5932-5941*, 2009.

[17] C. Z. B. Gunduz H, "Istanbul (bist) daily prediction using financial news and balanced feature selection.," *Expert Syst Appl; 42: 9001-9011*, 2015.

[18] P. J. M. W. P. G. Martinez LC, da Hora DN, "From an artificial neural network to a stock market day-trading system: a case study on the bmf bovespa.," *IJCNN; pp. 2006-2013.*, 2009.

[19] K. A, "A neural networks filtering mechanism for foreign exchange trading signals. in: Intelligent computing and intelligent systems," *Intelligent Computing and Intelligent Systems*, 2010.

[20] S. J. B. M. Markovic IP, Stojanovic MB, "Stock market trend prediction using support vector machines.," *Facta Universitatis Series Automatic Control and Robotics 13: 147-158.*, 2014.

[21] S. B. Imandoust and M. Bolandraftar., ". forecasting the direction of stock market index movement using three data mining techniques: the case of tehran stock exchanges," *Int. Journal of Engineering Research and Applications ISSN : 2248-9622, Vol. 4, Issue 6(Version 2), pp.106-117*, 2014.

[22] A. Timmermann and C. W. Granger, "Efficient market hypothesis and forecasting," *International Journal of Forecasting, vol. 20,no., pp. 15- 27*, 2004.