An Entropy-based Evaluation for Sentiment Analysis of Stock Market Prices using Twitter Data

Andreas Kanavos*, Gerasimos Vonitsanos*, Alaa Mohasseb†, Phivos Mylonas‡
*Computer Engineering and Informatics Department
University of Patras, Patras, Greece
{kanavos, mvonitsanos}@ceid.upatras.gr
†School of Computing
University of Portsmouth, Portsmouth, UK
alaa.mohasseb@port.ac.uk
‡Department of Informatics
Ionian University, Corfu, Greece
fmylonas@ionio.gr

***Abstract*—Stock markets prediction is considered a considerably demanding task due to its notable returns as well as due to the high randomness within the stock market. Moreover, stock price alternations are primarily related to the capital circumstances and hot occasions/events. Nowadays, researchers have sufficiently improved prediction accuracy by taking into consideration news and social media. However, the existing strategies do not employ the different impacts that events may pose. Streaming data proves to be a perpetual real-time source of data analysis as information from different web sources can be carried. In this paper, we explore whether estimations, in terms of sentiment analysis derived from Twitter posts, can be correlated to the stock market prices. Initially, the daily Twitter posts are analyzed and different n-grams along with two strategies that are utilized to increase the accuracy of the classification, are applied. Spark streaming has been employed for the processing of Twitter data, while Apache Flume has been utilized for the analysis.**

Keywords*—Stock Market, Social Networks, Sentiment Analysis, Spark Streaming, Apache Flume, Apache Cassandra

I. Introduction

Twitter has experienced enormous development over the past few years. Specifically, more than 200 million users are registered, while 50 million daily users and 400 million monthly visitors, are considered. Approximately 1 billion tweets are generated by Twitter users every five days. With so many people exchanging opinions about their different conclusions regarding an abundance of different subjects, Twitter is considered a rich source of real-time data with respect to current societal trends and opinions.

Behavioral economics gives us an insight regarding individuals that are not rational consumers and individual behaviors and decisions are greatly influenced by polarities, and indeed by the opinions of others. This ought to remain true for societies at large; that is, society can involve mood states influence affect their collective decision-making. So, if each tweet is considered a condensed outline of a person's mood or opinion about a certain subject, in following, the total number of tweets about the subject should express the collective mood. Public mood should be associated with or even predict of economic indicators [7].

The prediction of stock market prices is considered a classic yet challenging problem, drawing the attention of computer scientists as well as economists and financial analysts [1], [5], [8]. More to the point, linear and machine learning tools have been investigated for the past decades aiming at developing an efficient prediction model. Concretely, in 2019, the esteem of global equities surpassed $\$85$ trillion[1]. One critical feature regarding markets existence is the fact that investors are constantly searching for ways to procure companies information within the market for improving their investment returns. In the past, investors depended upon their personal encounter to distinguish market patterns, but in our days, this is not doable due to the markets estimate and the trades execution speed. A simple statistical analysis of financial data can provide various insights, however, in recent years, investment companies have progressively utilized various AI systems to seek patterns in massive amounts of real-time financial data.

The theoretical and empirical review on the efficient markets model is demonstrated in [11]. The empirical work considers the alteration of security prices to three important data sets; *weak form* tests, where the dataset consists of historical prices only, *semi-strong form* tests, where it is reviewed if prices effectively adjust to other data that is freely accessible, e.g., announcements of yearly profit, etc., and *strong form* tests that review if certain investors have strong access to any data pertinent for price configuration.

This study presents a methodology that forecasts the movement of the initial price (opening price) of the shares for a specific company. A distributed framework for efficiently collecting, aggregating and moving massive quantities of streaming data, i.e., Apache Flume, in collaboration with a NoSQL database, i.e., Apache Cassandra, are employed to take into consideration sentiment analysis in the Twitter posts for predicting stock market prices. Also, the appliance of different n-grams along with two strategies that are utilized to increase the accuracy of the classification, is proposed.

The proposed research is organized in the following mode. Section II describes the related works concerning stock mar-

[1] https://www.cnbc.com/2019/12/24/global-stock-markets-gained-17-trillion-in-value-in-2019.html

ket prediction and sentiment analysis. Moreover, Section III introduces the process of data representation from the Twitter platform as well as the overall architecture of the system and the features utilized, while in Section IV, the implementation details are presented. Furthermore, Section V demonstrates the experimental results and conclusions retrieved from the current study. Ultimately, in Section VI, the synopsis of the proposed paper is concluded and directions for future work are drawn.

II. Related Work

Sentiment analysis of stock market prices using machine learning is considered a popular field of data mining with many references. Initially, authors in [3] investigated whether the estimations of mood states retrieved from large-scale Twitter posts are related to the esteem of the Dow Jones Industrial Average (DJIA) over time. Twitter posts texts were analyzed by two mood tracking tools and divided in 6 dimensions (Alert, Calm, Happy, Kind, Sure and Vital). Authors in [17], authors applied machine learning algorithms for sentiment analysis to discover the relationship among "public sentiment" and "market sentiment". Twitter information were used to predict the public mood and this predicted mood along with past days' DJIA values try to forecast stock market alternations.

Furthermore, supervised machine learning strategies for predicting tweets sentiment analysis are applied as the relationship among a company's stock market movements and tweets sentiment is effectively analyzed in [19]. The experiments introduced a solid association between stock prices rises and falls with the tweets sentiments. In [16], authors built a system for stock prices prediction and proposed an approach that represents numerical price information by technical indicators by means of technical analysis. A deep learning model is developed to learn the sequential information within the market snapshot series.

The Granger causality test is used in [21] where the relations between financial markets and Twitter for a 15 months period are investigated; concretely, the Twitter volume along with sentiments of the 30 large stock companies that form the Dow Jones Industrial Average (DJIA) index. Results depicted a relatively low Pearson correlation and Granger causality between the corresponding time series over the entire time period. However, the experiments showed that sentiment polarity of Twitter peaks suggests the direction of cumulative abnormal returns. Similarly, an investor sentiment proxy extracted from Twitter to investigate whether investor sentiment, as expressed in daily happiness, has predictive power for stock returns in 10 international stock markets, is utilized in [27]. To account for complex correlations between stock returns and sentiment, a Granger non-causality test in quantiles is employed.

Twitter posts for a 6 months period were collected, and only a randomized sub-sample of the total number of tweets has been used in [28]. The emotions of hope and fear were measured on a daily basis with the aim of analyzing the relationship between these indices and stock market indicators. Authors showed that sentiment tweet percentage was negatively correlated with stock markets, but shown a noteworthy positive association to VIX. In addition, in [25], approximately $250,000$ tweets related to stock prices were analysed and the results introduced a relationship among stock returns and tweets sentiment, trading and message volume, and finally volatility and disagreement.

Also, authors in [18] proposed a model for forecasting the movement of stock prices by incorporating the sentiments of the company's specific topics, derived from social media, into the stock prediction model. Comparing the average accuracy of 18 stock companies in transactions of a one year period, when only comparing the strategies for the stocks that are complicated to predict, the proposed method accomplished 9.83% better accuracy in contrast with the historical price method and 3.03% better accuracy than the human sentiment method. Twitter's capacity of predicting consumer purchases, by noticing the association among societal Twitter trends and hourly stock prices of the top gainers and top losers of 10 companies concerning the technology sector, is examined in [7]. Experimental results depicted that the movements of stock prices are more rapidly predictive for Twitter sentiment movements. In addition, there is no significant prescient control of trending negative sentiment scores on stock markets concerning a particular subject.

Moreover, authors in [22] introduced a novel methodology to determine investor sentiment retrieved from social media messages. More to the point, the connection between real-time investor sentiment and intra-day stock returns was investigated. A words lexicon was initialized by employing a dataset with messages posted on the microblogging platform, namely StockTwits; there, the terms are utilized by investors when they share opinions about the bearishness or the bullishness of the stock market. Authors in [26] identify directions for future ML stock market predictions based upon a review of current literature. A similar to the current work is the one explored in [10], where authors explore the effectiveness of social network analysis as well as sentiment analysis in predicting trends by mining publicly available online data sources. In our previous works, we have utilized cloud-based architectures aiming at creating sentiment analysis tools for Twitter data, based on Apache Spark framework [2], [4], [14].

Several works have investigated the role of deep learning models in stock prices prediction. Authors in [13] give a review of recent progress by surveying more than 100 related published articles in the past three years in order to show the rapid utilization of deep learning models. A similar study was carried out in [9], where the public mood derived from Twitter posts can be used to forecast the movement of particular stock prices through batch processing by employing the long short-term memory (LSTM) recurrent neural networks. Authors in [6] deployed a method to study and analyze communication dynamics in the blogosphere to decide the relationships with stock market movement. Information roles as well as contextual attributes for 4 technology companies were identified and modelled as a regression problem in a Support Vector Machine framework. Authors in [20] utilized a deep LSTM Neural Network (NN) with an embedded layer and a long short-term memory NN with an automatic encoder with the aim of performing more efficient stock market predictions. The experimental results, like visualizations and analytics, aim at demonstrating the Internet of Multimedia of Things for stock analysis.

Finally, the trust between users, which can be considered as a filtering and amplifying mechanism for social media to increase its relationship with financial information in the stock market, is proposed in [23]. Real stock market information was used as ground truth for the proposed trust management system. To justify the trust management system, the Pearson correlation test was introduced for an 8 months period. A deep convolutional long short-term memory (DeepConvLSTM) model is trained by using the proposed Rider-based Monarch Butterfly Optimization (Rider-MBO) algorithm in [15]. Authors examined a stock market prediction system that can predict the state of the stock market. This corresponding Rider-MBO algorithm is the integration of the Rider Optimization Algorithm (ROA) and MBO.

III. Proposed Architecture

In this section, we focus on stock market forecasting as a particular data processing architecture. This concrete architecture is designed with the aim to handle massive quantities of information by incorporating two different strategies, namely batch and stream processing. This design endeavors to balance fault-tolerance, latency as well as throughput by utilizing batch handling for supplying accurate and extensive views of batch information, whereas at the same time utilizing real-time stream processing to provide online data views.

The Lambda architecture constitutes of 3 layers: initially, the **Batch Layer** continuously receives new data as a feed to the proposed system. This layer is considered as input to the next layer in the form of batch views and it checks as well as corrects, if needed, the entirety of data at a single step. In following, the **Serving Layer** records the batch views in a way that they can be queried in low-latency on an ad-hoc basis. Finally, **Speed Layer (Stream Layer)** handles the data that are not yet delivered in the batch view because of the latency of the first layer. This layer employs the latest data so as user can be provided with complete information real-time views. The Lambda architecture is illustrated in Figure 1.

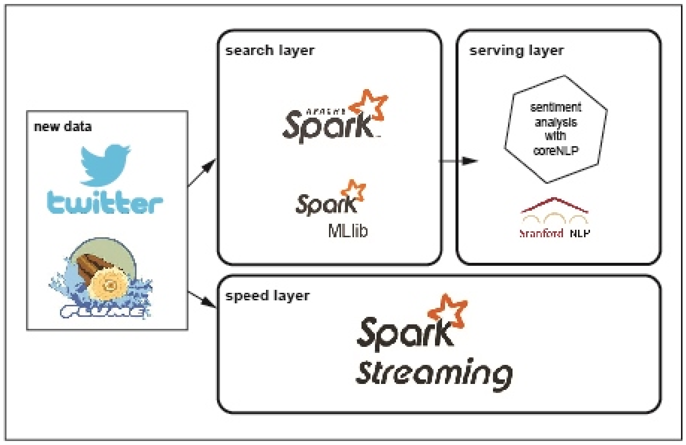


Fig. 1. Information Flow Analysis Architecture

A. Apache Flume

Apache Flume[2] is considered a distributed service for initially collecting, in following aggregating, and finally transfering massive quantities of streaming data into HDFS in a rather efficient way. One common example use case of Flume is the collection of log data present in log files from web servers and in following its aggregation in HDFS for analysis. Flume's plain and flexible architecture, when dealing with streaming data flows, acquires reliability mechanisms along with many recovery mechanisms.

B. Pre-processing Modules

Several pre-processing steps should be employed in order to accelerate the mining process of the retrieved data [12]. Naive Bayes, i.e. one popular machine learning algorithm with high accuracy, was selected for classification along with the Stanford Core NLP[3] for sentiment analysis. The modules are the following:

Regular expressions: These expressions were utilized in order to remove urls, references and other characters that match an already defined pattern as well as pointless spaces. Also, the redundancy of characters harnessed for emphasizing the meaning of a sentence, which do not attach sentiment significance, were subtracted.

Punctuation marks: These marks do not facilitate any sentiment magnitude to the text and thus, should be subtracted.

Part-of-Speech (POS) Tagging: It is the process of classifying words into their parts of speech and labeling them accordingly, e.g. noun, verb, adverb. This process uses the text's context to be analyzed as well as a number of aggregated components (corpus).

Lemmatization: In this module, terms' lexical and morphological examinations are considered with the aim of removing composite suffixes and in following recovering their lexical form. It is utilized after the aforementioned POS tagging for sentiment analysis using a plethora of ML algorithms.

Tokenization: In this module, the sentences are separated into a set of symbol terms that can be employed for alternating the initial text. All the terms of the text are stored within the same token list and the text's features appear in a list with their appearing order.

Stop words: The whole attempt of pre-processing is to examine meaningful words that define the general sentiment manifested in a text. These terms frequently appear without expressing a sentiment polarity, and as a result, should be subtracted.

C. Features

One of the most traditional and popular structures used in NLP as well as text mining area are the n-grams. They are considered a corpus of co-occurring words in terms of a particular window. Three data representations, utilized by particular techniques in terms of the proposed research, are unigrams, bigrams, and trigrams. The primary used are the bigrams, which look for one specific term, the trigrams, which look for two specific terms, and the general n-grams, which look for $n-1$ terms.

To increase classification's accuracy, the n-grams that do not indicate any sentiment must be discarded. As these n-grams can be uniformly identified across all datasets, two

[2] https://flume.apache.org/

[3] https://stanfordnlp.github.io/CoreNLP/

strategies are presented. The first strategy is, given different datasets, to calculate the entropy H of the probability distribution of an n-gram [24]:

$$H(g) = -\sum_{i=1}^{n} P(x_i) \log P(x_i) \qquad (1)$$

where g is a particular n-gram, $x_1, \ldots, x_n$ are the possible outcomes which occur with probability $P(x_1), \ldots, P(x_n)$, $\sum$ denotes the sum over the variable's possible values and n is the number of sentiments (in our paper, $n = 3$ for positive, negative, neutral).

High value of entropy indicates that the appearance distribution of an n-gram in different datasets when considering sentiment analysis tends to become uniform. Therefore, such an n-gram does not contribute to the classification analysis. On the other hand, low value of entropy indicates that an n-gram appears in some datasets more often than in others and therefore, it can emphasize more a specific sentiment. Thus, to increase the accuracy of the classification, we prefer for consideration n-grams with low entropy values.

Regarding the second strategy, we introduce a semantic term, e.g. salience, calculated for each n-gram, where it takes a value between 0 and 1. The low value indicates a low majority of n-grams and such an n-gram should be distinguished:

$$S(g) = \frac{1}{n} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} 1 - \frac{min(P(x_i), P(x_j))}{max(P(x_i), P(x_j))} \qquad (2)$$

IV. Implementation

A. Model Overview

Current paper proposes a model that consists of two main modules, namely data collection and data analysis. The first module, i.e. data collection, is utilized in order to crawl a number of posts from Twitter with use of Apache Flume streaming engine. In following, these posts/tweets are stored in Cassandra, which is a NoSQL database, efficient in terms of scalability. The second module considers the storing procedure, where the system performs a sentiment analysis procedure to forecast the correlation between stock market prices and tweets sentiment analysis.

B. Dataset

The data were stored in Cassandra database sorted out by each day in different collections in order to simplify the standard procedures of fetching, management and analysis. The experiments were utilized with use of a cluster consisting of 10 nodes, where each node is equipped with quad-core Intel(R) Core(TM) i5-2400 CPU@3.10 GHz and 4GB RAM.

The stock prices of both Apple and Microsoft shares have been retrieved using Quandl's platform[4] that is a powerful source for financial and economic, serving investment professionals. The estimated stock prices for the test data is compared to the actual prices derived from the dataset.

Furthermore, to test our proposed method, several Twitter posts related to these two huge companies were downloaded. To guarantee that only relevant tweets are taking into consideration, all downloaded posts with appropriate keywords and hashtags were filtered; the keywords are Microsoft, Apple and the hashtags are #MSFT, #AAPL. The filtered dataset resulted in $156,000$ tweets from $01/05/2016$ to $30/06/2017$.

V. Results

The evaluation of our proposed method has been conducted with a number of experiments using classification methods and analysis on prediction prices. We evaluated our method by measuring the performance of our methodology in terms of accuracy and decision, where accuracy is calculated as the number of corrected classifications divided by the total number of classifications; and decision is defined as the number of retrieved documents divided by the total number of documents.

A. Accuracy

Initially, we have measured the effect of different n-grams on classifier performance. The results are shown in Figure 2. As we can observe, bigrams achieve the best performance. In other words, bigrams are considered the best despite the coverage of the unigrams and the ability to capture sentiment patterns of trigrams.

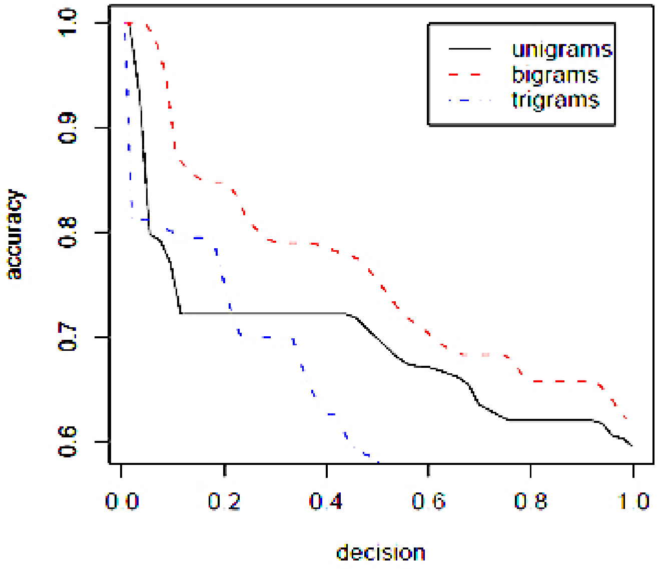


Fig. 2. Classification accuracy when using unigrams, bigrams and trigrams

In following, we measure the effect of adding negative words when n-grams are formed as depicted in Figure 3. Concretely, we can achieve high accuracy even with a low decision value. So, if the classifier regarding the sentiment analysis will be used, then the results will be precisely accurate.

Except the examination of the effect of dataset on the system performance, we have also measured the performance by using the F_β metric:

$$F_\beta = (1 + \beta^2) \times \frac{precision \times recall}{\beta^2 \times precision + recall} \qquad (3)$$

[4] https://www.quandl.com/

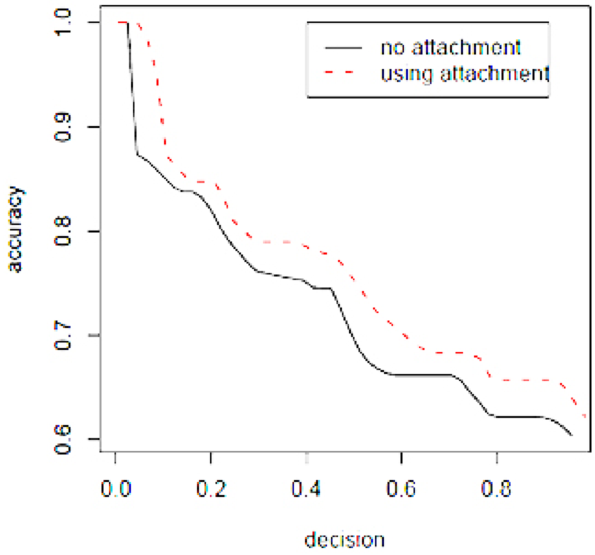


Fig. 3. The impact of attaching words that express negative sentiment

This metric uses a positive real factor β; this factor is chosen in a way that recall is considered β times as important as precision. However, in our experiments, we have replaced the precision with the accuracy as well as the recall with the decision, because we are dealing with multiple categories rather than binary classification:

$$F_\beta = (1+\beta^2) \times \frac{accuracy \times decision}{\beta^2 \times accuracy + decision} \qquad (4)$$

where $\beta = 0.5$. In this experiment, no n-grams filtration is utilized as presented in Figure 4. As the sample size is increasing, the performance of our system is improved. However, at the certain point when the dataset is large enough, this improvement can not be achieved just by increasing the size of the training data.

Furthermore, we have considered two filtering strategies for common n-grams, namely, salience (semantics) and entropy. The following Figure 5 shows that the use of semantics provides better accuracy and therefore, semantics distinguishes common n-grams in a better way than the entropy.

B. Prediction Price Analysis

As above mentioned, the stock prices of Apple and Microsoft shares have been retrieved with use of Quandl's platform and the estimated stock prices for the test data are compared to the actual prices derived from the dataset. More to the point, the training set comprises of an equal number of correctly classified as positive as well as negative reviews about the corresponding company.

Furthermore, some additional Twitter data were collected to be considered as test datasets. The forecast graph for each dataset in terms of its history was designed, showing this way the daily polarity of the tweets for the company. Figure 6 illustrates the association among the actual and the forecast price of Apple and Microsoft shares, respectively.

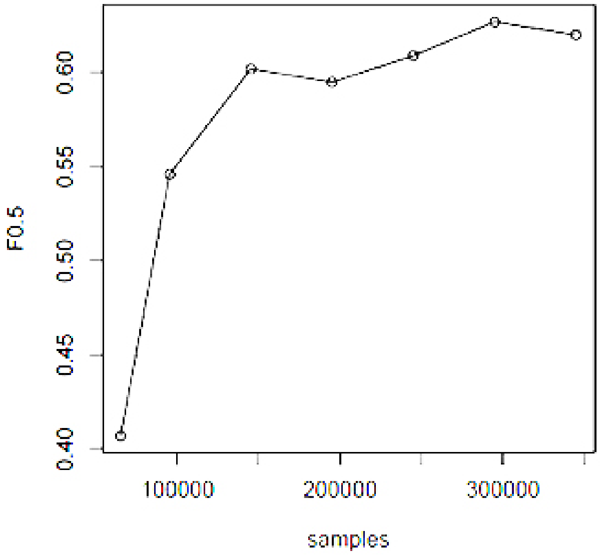


Fig. 4. The effect of increasing the dataset size on the F_β metric

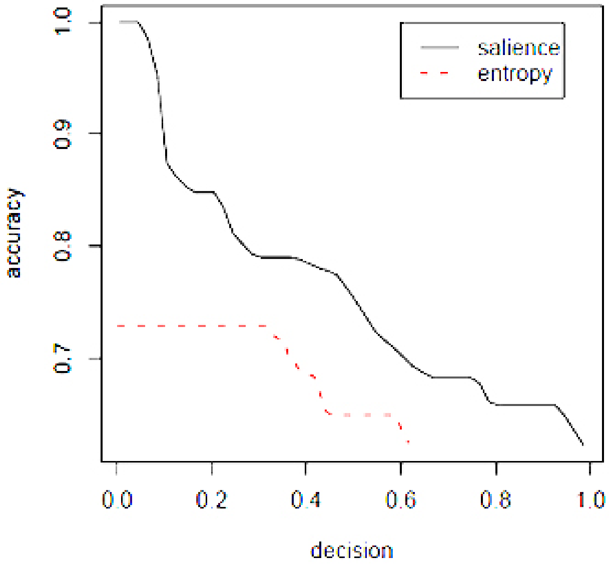
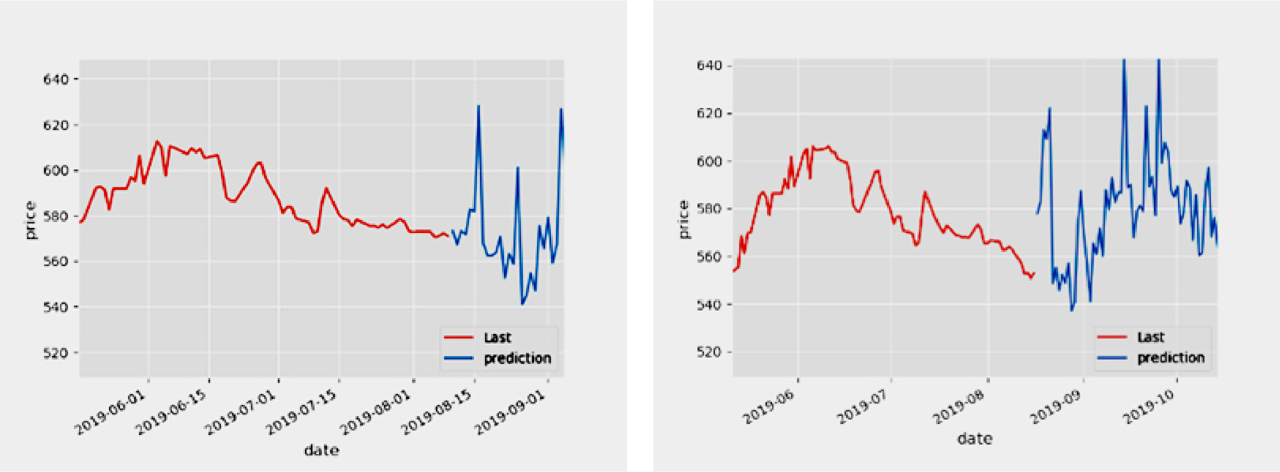


Fig. 5. Semantics (salience) vs entropy for distinguishing common n-grams

VI. Conclusions and Future Work

This work focuses on developing a methodology that forecasts the movement of the initial price of the shares and can be applied to different companies. A distributed framework for efficiently collecting massive quantities of streaming data, Apache Flume in combination with a NoSQL database for managing huge amount of data, i.e. Apache Cassandra, are utilized to efficiently implement sentiment analysis in the Twitter posts for predicting stock market prices. The accuracy of the classification is increased with the application of different n-grams along with two utilized strategies.

Regarding future work, the proposed methodology can be augmented with a module that can overlook the fact that stock price variations are not available when the stock markets are closed while on the other hand, Twitter posts can be



(a) Apple shares

(b) Microsoft shares

Fig. 6. Correlation between forecast and actual price of shares

generated during any time. In addition, Twitter posts can be comprised of images or even hyperlinks to particular websites, which is not taken into account in this paper and this can be considered an open problem that may affect our outcome. The introduction of deep learning algorithms classifiers can also be utilized to predict stock information. Finally, a limitation that illustrates the restrictions of our proposed methodology and suggests room for improvement in future studies is the sentiment analysis module, where non-literal phrases such as sarcasm, cannot be identified by the current tool.

References

[1] W. Antweiler and M. Z. Frank. Is all that talk just noise? the information content of internet stock message boards. *The Journal of Finance*, 59(3):1259–1294, 2004.

[2] A. Baltas, A. Kanavos, and A. Tsakalidis. An apache spark implementation for sentiment analysis on twitter data. In *1st International Workshop on Algorithmic Aspects of Cloud Computing (ALGOCLOUD)*, pages 15–25, 2016.

[3] J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[4] A. Bompotas, A. Ilias, A. Kanavos, C. Makris, G. Rompolas, and A. Savvopoulos. A sentiment-based hotel review summarization using machine learning techniques. In *16th International Conference on Artificial Intelligence Applications and Innovations (AIAI)*, volume 585, pages 155–164, 2020.

[5] W. F. M. D. Bondt and R. Thaler. Does the stock market overreact? *The Journal of Finance*, 40(3):793–805, 1985.

[6] M. D. Choudhury, H. Sundaram, A. John, and D. D. Seligmann. Can blog communication dynamics be correlated with stock market activity? In *19th ACM Conference on Hypertext and Hypermedia*, pages 55–60, 2008.

[7] S. Chung and S. Liu. Predicting stock market fluctuations from twitter. *Berkeley, California*, 2011.

[8] P. H. Cootner. *The Random Character of Stock Market Prices*. MIT press, 1967.

[9] S. Das, R. K. Behera, M. Kumar, and S. K. Rath. Real-time sentiment analysis of twitter streaming data for stock prediction. *Procedia Computer Science*, 132:956–964, 2018.

[10] L. Doshi, J. Krauss, S. Nann, and P. Gloor. Predicting movie prices through dynamic social network analysis. *Procedia-Social and Behavioral Sciences*, 2(4):6423–6433, 2010.

[11] E. F. Fama. Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2):383–417, 1970.

[12] S. García, J. Luengo, and F. Herrera. *Data Preprocessing in Data Mining*, volume 72 of *Intelligent Systems Reference Library*. Springer, 2015.

[13] W. Jiang. Applications of deep learning in stock market prediction: Recent progress. *CoRR*, abs/2003.01859, 2020.

[14] A. Kanavos, N. Nodarakis, S. Sioutas, A. Tsakalidis, D. Tsolis, and G. Tzimas. Large scale implementations for twitter sentiment classification. *Algorithms*, 10(1):33, 2017.

[15] A. Kelotra and P. Pandey. Stock market prediction using optimized deep-convlstm model. *Big Data*, 8(1):5–24, 2020.

[16] X. Li, P. Wu, and W. Wang. Incorporating stock prices and news sentiments for stock market prediction: A case of hong kong. *Information Processing & Management*, 57:102212, 2020.

[17] A. Mittal and A. Goel. Stock prediction using twitter sentiment analysis. *Standford University*, 15, 2012.

[18] T. H. Nguyen, K. Shirai, and J. Velcin. Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications*, 42(24):9603–9611, 2015.

[19] V. S. Pagolu, K. N. R. Challa, G. Panda, and B. Majhi. Sentiment analysis of twitter data for predicting stock market movements. In *International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*, 2016.

[20] X. W. Pang, Y. Zhou, P. Wang, W. Lin, and V. Chang. An innovative neural network approach for stock market prediction. *The Journal of Supercomputing*, 76(3):2098–2118, 2020.

[21] G. Ranco, D. Aleksovski, G. Caldarelli, M. Grčar, and I. Mozetič. The effects of twitter sentiment on stock price returns. *PloS One*, 10(9):e0138441, 2015.

[22] T. Renault. Intraday online investor sentiment and return patterns in the us stock market. *Journal of Banking & Finance*, 84:25–40, 2017.

[23] Y. Ruan, A. Durresi, and L. Alfantoukh. Using twitter trust network for stock market analysis. *Knowledge Based Systems*, 145:207–218, 2018.

[24] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

[25] T. O. Sprenger, A. Tumasjan, P. G. Sandner, and I. M. Welpe. Tweets and trades: the information content of stock microblogs. *European Financial Management*, 20(5):926–957, 2014.

[26] T. J. Strader, J. J. Rozycki, T. H. Root, and Y.-H. J. Huang. Machine learning stock market prediction studies: Review and research directions. *Journal of International Technology and Information Management*, 28(4):63–83, 2020.

[27] W. You, Y. Guo, and C. Peng. Twitter's daily happiness sentiment and the predictability of stock returns. *Finance Research Letters*, 23:58–64, 2017.

[28] X. Zhang, H. Fuehres, and P. A. Gloor. Predicting stock market indicators through twitter “i hope it is not as bad as i fear”. *Procedia - Social and Behavioral Sciences*, 26:55–62, 2011.