



Predicting Stock Market Movements with Social Media And Machine Learning

Paraskevas Koukaras¹^a, Vasiliki Tsihli¹ and Christos Tjortjis¹^b

¹ School of Science and Technology, International Hellenic University, 14th km Thessaloniki–N. Moudania, Thessaloniki, Greece
{p.koukaras, v.tsihli, c.tjortjis}@ihu.edu.gr

Keywords: Social Media, Prediction, Machine Learning, Data Science, Stocks

Abstract: Microblogging data analysis and sentiment extraction has become a popular approach for market prediction. However, this kind of data contain noise and it is difficult to distinguish truly valid information. In this work we collected 782.459 tweets starting from 2018/11/01 until 2019/31/07. For each user day, we create a graph (271 graphs in total) of user that tweets and their followers and we utilize this graph to obtain a PageRank score for each user. This score is then multiplied with the sentiment data. Findings indicate that using an importance-based measure, such as PageRank, can improve the scoring ability of the models. On average the PageRank dataset achieved a lower mean squared error than the economic dataset and the sentiment dataset. Finally, we tested multiple machine learning models, showing that XGBoost is the best model, with the random forest being the second best and LSTM being the worst.

1 INTRODUCTION


Stock market forecasting is an important academic topic, which has attracted academic interest since the early 1960's (Fama, 1965). Although a lot of effort and time has been spent on predicting financial time series, the results of the research are not robust. In recent years a lot of researchers have shifted their focus from classical econometric approaches to machine learning approaches. With the rise of microblogging platforms, such as Twitter, StockTwits and others, information is more available than ever before and given that emotions can have a significant effect on economic decisions (Bollen et al., 2011), alongside with herding phenomena (Devenow and Welch, 1996), one can assume that mining information through such microblogging platforms might be the key to achieve better results in predicting stock market movements.


Stock market forecasting has drawn a lot of academic attention since the 1960's. The first model that revolutionized how the stock was evaluated is the Capital Asset Pricing Model (or CAPM for short). CAPM was developed¹ by William Sharpe (Sharpe, 1964) who built on top of Markowitz's diversification

theory. The model is fairly simple and is based on the sensitivity that a stock's returns exhibit over the systemic risk (or market risk), which is expressed quantitatively through the use of a factor, called beta and which is symbolized by β .

CAPM measures the return of a stock in accordance with the market risk. Every other risk that stems from the stock itself can be diversified as Markowitz proved in the portfolio theory and thus, there is no point in measuring it. Although CAPM has been a fundamental tool with which asset managers make decisions, it has been criticized a lot by academics because by its nature, it has a lot of problems. It has been proven that the model is not robust, and that it fails to give accurate results consistently. Fama and French (Fama and French, 1993) stated that the model is not robust and that a model that takes into account the size and the ratio of accounting over stock market value is more accurate.

Fama's and French's research gave the incentive to start looking for other factors that may be affecting the returns of a stock and this gave birth to a whole new way of evaluating a stock, which is called technical analysis. Technical analysis is not an academic principle, rather it is based on ratios and indicators that capture the momentum of the stock market. Although technical analysis is not based purely on academic research, it is used extensively and it is a com-

^a  <https://orcid.org/0000-0002-1183-9878>

^b  <https://orcid.org/0000-0001-8263-9024>

¹There is a dispute on who deserves credit about CAPM, for more information check (Treynor, 1962)

mon practice.

In recent years there has been a lot of effort to construct indicators or ratios based on the information of the microblogging community. Essentially, those indicators provide an overall sentiment over the market or a particular stock, and thus, the trader can have a more objective metric about the "feelings". Moreover, this data might contain useful information that would be unavailable otherwise. On the other hand, this approach contradicts one of the most fundamental economic theories, which is the Efficient Market Hypothesis. As Fama (Fama, 1965) suggested in his seminal paper, the price of a given stock embodies all the prior information available and thus it is impossible to forecast future values since the current ones reflect everything. Moreover, in efficient-market hypothesis (EMH), it is believed that the market adjusts the prices instantly as the news spread, and as Fama (Fama, 1965) noted, the most probable future price is the current price.

Nevertheless, recent empirical research provided evidence that sentiment plays an important role and can act as a determining factor of the stock market returns.

2 BACKGROUND

This section provides a review of the relevant literature of sentiment analysis on microblogging platforms and machine learning techniques to predict stock market returns. Sentiment analysis on financial news or forum posts is not a relatively new idea (Tetlock et al., 2008), but it has recently gained a lot of attention since more data are available and the tools for processing the data are becoming more and more trivial. There, exist numerous papers that examine this subject, with a plethora of methodologies. Thus, we opted to break the literature in two main parts. In the first part, we provide the reader with an overview of the studies which use statistical approaches, such as correlations and OLS Regression. The other part examines the literature in which some machine learning approaches are used, such as decision trees, neural networks, etc.

To have a far-reaching variety of papers to examine, we decided to use the ACM Digital library, the IEEE Xplore Digital Library, the Science Direct, and the Springer Link (Figure 1). The keywords we used were "Stock Market Sentiment Analysis" and "Stock Prediction Sentiment".

To present a substantial, but manageable number of papers, we decided to pose a restriction on the year

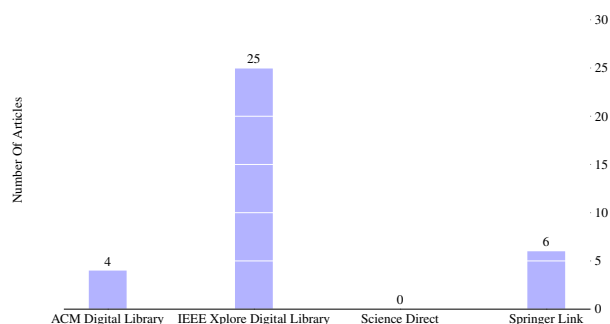


Figure 1: Final Results Per Digital Library.

of publication; we chose the years 2016–2019² (Figure 2), but we did not limit our research to any type of publication, with the exceptions that it had to be written in English and be accessible to us.

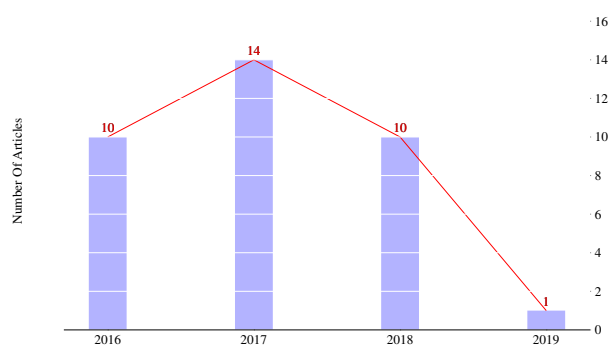


Figure 2: Results Per Year.

2.1 Statistical Approaches

The sentiment is an opinion of a view on a subject that is carried by a person. In recent years, and because of the more available data from social media, blogs, and forums, sentiment analysis has attracted a lot of academic research. In recent years, there has been a plethora of studies examining the prospect of sentiment analysis as a predictive factor of the stock market. It started from the seminal papers of Bollen et. al (Bollen et al., 2011) and Tetlock et. al (Tetlock et al., 2008). In (Bollen et al., 2011), the authors used the price of the Dow Jones Industrial Average and obtained the sentiment by OpinionFinder and GPOMS. GPOMS variables described the moods of the public, which allowed the authors to have a more accurate result. They chose a period in which both elections and Thanksgiving were included. Afterwards, all the variables were normalized.

The Granger causality (Granger, 1969) showed

²Cited papers from years before those, do not originate from our search, but from the papers we collected.

that the time lag that has the biggest predictive power is 3 days. Moreover, when the authors dropped the OpinionFinder variable and examined only the calm variable from GPOMS, the score was improved. Also, they showed that using the happy variable does improve the MAPE of the model but it drops the direction accuracy, which, according to the authors, indicates that there is a non-linear relationship between calm and happy variables. Moreover, authors in (Tetlock et al., 2008) analyzed financial news about specific firms and calculated a ratio of the negative words contained in the news articles. The stock prices seem to underreact to the information contained in the negative words. Lastly and more importantly, the best predictor for stock market prices is the ratio of the negative words that came only from the news that focus on the fundamentals.

An effect that has been extensively examined is if there is causation between sentiment and stock market movements. Authors in (Wong and Ko, 2016) used the Granger Causality test to test this hypothesis. They employed a similar to (Park et al., 2017) procedure in preprocessing to extract the sentiment. Afterwards, they classified the emotions into eight categories, namely "afraid", "amused", "angry", "annoyed", "don't care", "happy", "inspired", and "sad". After constructing the eight variables, they tested for Granger causality between those variables and the Korean Stock Price Index. The results indicated that different emotions affect stock market behavior with different time lags and different metrics. For example, the emotions classified as "amused" and "happy" affect the next day's stock price, whilst other emotions affect the next day's trading volume and the day's after.

Granger causality tests have been utilized by many authors to test if there is a relationship between microblogging activity and stock market movements. For example, authors in (Zhang et al., 2012) use this test. The authors collected data for the period between November 15, 2010 and April 20, 2011. The restrictions that were placed were that all tweets had to be a retweet only, that they had to contain the words "Hope" and "Fear" or "Worry", and that the location of the users had to be in the United States. Lastly, they limited their research only for tweets of economic sentiment by using keywords such as "dollar", "\$", "gold", "oil", "job", and "economy". Afterwards, they proceeded to the statistical analysis. The most interesting observation was that the gold keyword does not have a causation relationship with the gold but that it does have one with the exchange rate.

Many authors do not examine the relationship between the returns of stocks' but the volatility. Volatili-

ty is a term used in finance to indicate the fluctuations of stock returns – it is measured with the variance. Because time series data exhibit a phenomenon called volatility clusters, simple regression techniques have been proved to provide spurious results. That is because different fluctuations in different periods result in residuals with non-constant variance, which is a necessary and sufficient condition to have robust results.

Examining the relationship between volatility and microblogging data, like authors in (Li et al., 2017) did, solves this problem. The authors constructed an emotional index based on microblogging activity in Chinese forums and examined the relationship with the volatility of the SSE Composite Index for the second semester of 2016. Their results indicated that there is a strong relationship between those two indices.

2.2 Machine Learning Approaches

Statistical approaches focus on determining if there is a causal relationship between microblogging data and stock market movements and do not put a lot of effort into predicting. That is because of the volatility and the problems that it creates. So in recent years a lot of researchers use machine learning approaches, such decision trees, XGBoost and neural networks. In this section of the literature, we are going to review papers, which use those or similar techniques.

Decision trees have been proven useful in predicting the direction of stock market movements. In (Mahajan et al., 2008), the authors used a stacked classifier with Decision Tree and SVM to forecast the Bombay Stock Market. At first, they used the latent Dirichlet allocation algorithm to extract relevant topics from a financial news repository. They extracted 25 such topics and disregarded texts that did not correspond to those topics. From those 25 topics only 8 exhibited high correlations with the BSE Sensex index or with its volatility. Lastly, they used the stacked classifier to forecast the stock market. The best accuracy of their system was 62.02% with a training sample that ranged from August 2005 to December 2007 and a testing sample that ranged from January to April of 2008.

Another model that is extensively used to forecast the stock market is Naive Bayes (Schumaker and Chen, 2009). Many authors have used Naive Bayes either to classify the sentiment of the text or to study the movement of the stock market. For example, in (Suman et al., 2017), authors used Naive Bayes as a classifier to study the relationship between sentiment and Apple's stock. The authors mined data from a

specific site for financial news called Stocktwits and classified them using Naive Bayes. Their results indicated that there is a relationship between the sentiment and the stock market movement, which becomes strong when investors exhibit a bearish³ behavior.

One of the most interesting machine learning models used is Support Vector Machines (SVMs). SVMs have been used considerably (Schumaker and Chen, 2009; Chiong et al., 2018). In (Schumaker and Chen, 2009) the authors considered the same models but also introduced a sequential minimal optimization. The data used covered a period of one month, and they were mined from credible news sources, such as The Wire, Bloomberg, CNN, etc. They trained 3 models, one with only the extracted news, one with extracted news and the baseline of stocks' prices and the regressed estimate for stocks' prices. The authors evaluated the result based on the mean squared error and directional accuracy. The model with the best score was the one that utilized the strength of both the public sentiment data and the economic variables, with an MSE of 0.04621 and a directional accuracy of 57.1%.

Although SVM can be a very powerful tool, more advanced techniques may yield better results. For example, authors in (Sun et al., 2017) compared multiple models as well. Their selection included Logistic Regression, LSTM (Long short-term memory neural networks), SVM, Naive Bayes SVM, and an ensemble of methods. The most accurate method in sentiment classification was the ensemble of methods, which consisted of a weighted model of the Naive Bayes SVM and the LSTM models. The model achieved an accuracy of 71.3%. Lastly, the authors used classified information on stocks to optimize their portfolio strategy. The average total return for seven months was 19.54%, and the authors noted that their strategy was more stable than the market itself, which also indicated less risk.

There is a vast and growing literature that uses artificial neural networks comparing them to other machine learning techniques. In most of this literature, the best results are achieved by ANNs. For example in (Plakandaras et al., 2015), the authors compare multiple techniques to test if the efficient market hypothesis (EMH) holds in exchange rates. More specifically, their data consisted of four exchange rates, the USD/EUR, the USD/JPY, the USD/GBP, and the USD/AUD for the period between January 2, 2013 and December 26, 2013. The models used were Logistic Regression, SVM, Naive Bayes, KNN, Decision Trees with boosting (AdaBoost and LogitBoost),

³Bull/Bear: Financial terminology to indicate when the investors feel positive/negative respectively towards a stock.

and ANN. Moreover, to test the efficiency of using sentiment data, they constructed 5 different input sets. Their results indicated that there is not a methodology that consistently outperforms all of the others, rather than the outcome is susceptible to the exchange rate. On the other hand, KNN, SVM, and ANN exhibit higher forecasting accuracy than the other methodologies.

Lastly, many major financial companies, such as Bloomberg and Thomson Reuters, are now providing sentiment metrics. The authors in (Vanstone et al., 2019) used 6 different metrics from the Bloomberg database to test the importance of sentiment in 18 stocks. The variables included several news articles for each company, the number of positive articles and the number of negative articles. They tested two artificial neural network models, one that included the sentiment variables and one without them. Regarding root mean square error when forecasting the price of the stock, the model with the sentiment variables outperformed the basic model.

2.3 Graph Theory

The noisy nature of Twitter data has been noted by a lot of authors (See-To and Yang, 2017; Alshahrani Hasan and Fong, 2018).

A more efficient way to handle these data is to identify the importance of its author using graphs. Graphs have been utilized to map importance in very efficient ways. The most famous of all is PageRank (Page et al., 1999). PageRank was created by Sergey Brin and Lawrence Page (Brin and Page, 1998) and is the basis that Google was built on. What PageRank does is map every web page in the world wide web as a node. Each node gets a ranking according to the edges that lead up to it. What this algorithm achieves is that when a web page is referenced by other non-important web pages, its score is lower and thus, it will not be in the top suggested pages. A similar mapping can be used to model Twitter users.

Since the microblogging activity has been increased enormously, many new algorithms have been implemented. For example, in (Zhang et al., 2017), the authors provided an improved LeaderRank algorithm. More specifically, the authors used both Wikivote and Twitter networks to identify influential users. To test the results of their algorithm, they also extracted the LeaderRank and the PageRank from each graph. Their algorithm identified users who affect more nodes than the other two algorithms. Another known algorithm is the HITS (Hypertext-Induced Topic Search). The authors designed an HITS algorithm that is based on the topic-decision method and

then employed an LDA model that identified the critical events and the influential spreaders. As the authors noted, their approach largely reduced the impact of unrelated posts, which in turn increased the efficiency and accuracy of identifying critical events.

In (Bae and Lee, 2011), authors utilized graphs to test the sentiment significance with times series. The authors modeled only 13 very influential users, such as Barack Obama. Afterwards, the model took into account all of the users that replied, retweeted or mentioned any of these 13 users. The final graph consisted of 499,756 nodes. Lastly, using the users who interacted with Barack Obama and the sentiment from Barack Obama's tweets, they tested if those correlate with the Job Approval rating. They found that the sentiment of those tweets can be used to landscape the offline phenomena.

Although graphs have not been used extensively to model stock market prediction, the literature suggests that modeling the Twitter opinion space as a graph and extracting features, such as PageRank, can provide a solution to noisy data and also, act as estimators.

3 PROBLEM STATEMENT

One of the biggest problems encountered by the researchers that used data from Twitter of other relevant sources is that they are noisy (See-To and Yang, 2017; Alshahrani Hasan and Fong, 2018), thus yielding spurious results. To deal with that problem, the authors either choose a specific news source, such as MarketWatch (Hájek, 2018) or Thomson Reuters (Mittermayer and Knolmayer, 2006), but this approach might lead to overlooking important information. Another issue is that they use a lot of data which might hinder their research in terms of efficiency and statistical robustness (Antweiler and Frank, 2004).

Our objective is to provide a more efficient way of handling those massive data, by looking for and distinguishing those data that matter the most. To achieve that, we use graphs that are constructed based on users and their data accordingly. We believe that our approach solves the problem of noisy microblogging data, without disregarding any useful information that might exist. Given our hypothesis we expect that the dataset which accounts for the noise in the data have a better score than the simple sentiment dataset.

4 RESEARCH DESIGN

4.1 Data

This section presents the data. First, we provide an overview of the economic variables we chose and the reasons behind these choices. Afterwards, we present how we gathered Twitter data.

4.1.1 Economic Variables

Economic variables can act as predictors. There is a humongous number of such variables ranging from fundamental analysis of a company's balance sheet to technical indicators specially designed to capture specific events. In this work, we chose to use technical indicators for multiple reasons. Firstly, technical analysis is based on examining a stock's trend, thus it constitutes a more robust tool for prediction. Moreover, one of the core principles of technical analysis is that a stock's price reflects all the available information, thus it is focused more on past behavior of the market. Although technical analysis has been dismissed by academics (Malkiel and Fama, 1970), many of the leading trading companies use technical indicators to identify signals and trends on time.

For these reasons, we concluded that technical indicators are more suited for our research. Since technical indicators do not focus on news events, our final dataset will be more balanced and have features that try to capture different aspects of trading. From all the available technical indicators, we opted for 5 of the most common ones. Those are:

1. The Aroon Oscillator is a trend indicator that measures the power of an ongoing trend and the probability to proceed by using elements of the Aroon Indicator (Aroon Up and Aroon Down). Readings above zero show an upward trend, while readings below zero show a downward trend. To signal prospective trend changes, traders watch for zero line crossovers (Mitchell, 2019).
2. The CCI was created to determine the rates of over-bought and over-sold stocks. This is done by evaluating the price-to-moving average (MA) relationship or, more specifically, by evaluating ordinary deviations from that median (Kuepper, 2019).
3. On-balance volume (OBV), is a momentum indicator that measures positive and negative volume flows (Staff, 2019).
4. The RSI is a momentum index, measuring the magnitude of the latest price modifications, that is used to assess which stocks are over-bought or

over-sold. The RSI is an oscillator. Traditionally, traders interpret a score of 70 or higher as a sign that a stock is overbought or overestimated, which might lead to a trend reversal. An RSI of 30 or lower signals that a stock is undervalued (Blystone, 2019).

5. The Stochastic Oscillator attempts to predict price turning points by comparing the last closing price of a security to its price range. It takes values from 0 up to 100. A value of 70 or higher signals an overbought security.

These indicators were chosen for two main reasons. Firstly, they are very robust and are used extensively in the industry. Secondly, these indicators belong to a special category which is called "Oscillators". Oscillators are indicators that fluctuate within a range and are used to capture short term trends. Our sample period ranges from December, 1st of 2018 to July, 31st of 2019. This is a period that is characterized by high fluctuations and small but powerful shocks (Trade War, No Deal Brexit, etc.), thus we believe that using such variables will provide more accurate results than using fundamental analysis. Finally, to collect the economic variables, we used the API of Alpha Vantage.

4.1.2 Twitter Data

We are interested in two categories of data, the tweets and the users that wrote those tweets. The main problem reported in the literature is the noisy nature of Twitter data (Rousidis et al., 2019; Koukaras et al., 2019; Koukaras and Tjortjis, 2019; Belevelis et al., 2019; Oikonomou and Tjortjis, 2018). To overcome this problem, we used the "cashtag" or "\$" in the tweets, which as (Chakraborty et al., 2017) notes, is more suited for gathering stock related data.

The module for gathering Twitter data is built upon a library called Twint. This library can provide tweets, users' statistics (followers, following, likes, etc.) and also, it can gather users' followers. Moreover, it also has a built-in function for storing those data directly to a database. From the downloaded tweets, we take all the tweets authors' usernames and gather metrics for them. These metrics are used when we are checking the validity of our data. We also gather all users' followers, a metric that is going to be used in the graph module.

4.2 Methodology

This section includes a summary of the main processes, such as how we identified the most influential users, describing the utilized algorithms. The

methodology splits into three main parts. At first, we designed a Graph for the users to obtain their importance using the PageRank algorithm (Page et al., 1999). Afterwards, we analyzed the tweets that were obtained using two different lexicons and lastly, we estimated five different machine learning models.

4.2.1 Identifying Influential Users

Graphs have not been used in the literature extensively, although most of the literature recognizes the problem with the noisy Twitter data. Because of the complexity of the project, we generated a graph, we computed the PageRank score for each edge, and the hub and authority scores. The Graph class is fairly simple and is based on the NetworkX library (Hagberg et al., 2008). Moreover, the PageRank and HITS algorithms are also implemented in the NetworkX library (Hagberg et al., 2008).

PageRank and HITS are two algorithms that are often used to measure the importance of nodes on directed graphs. Both of the algorithms were designed to rank websites. The PageRank algorithm is a recursive algorithm, where an internet page is important if and only if important pages are linked with it. As it is usually described, a website's score is the probability that any random person who is browsing on the web will end up on this website. This is by definition a Markov Process. Markov Processes have been used extensively to model recursive phenomena, such as the weather. The PageRank algorithm starts with a set of websites (denoting the number of those websites with N). On each website, we assign a score of $1/N$. Afterwards, we sequentially update the score of each website by adding up the weight of every other website that links to it divided by the number of links emanating from the referring website. But if the website does not reference any other website, we distribute its score to the remaining websites. This process is executed until the scores are stable.

The HITS algorithm was developed around the same time with PageRank (Kleinberg, 1999). HITS stands for Hypertext-Induced Topic Search and provides two scores, the "Authority" and the "Hub". We tried to compute the HITS algorithm, but the algorithm never achieved convergence. Since we wanted to compute the hubs and the authorities for each day in our sample, the recursiveness of the algorithm poses a significant barrier. On the computing part, for each date, we needed to create a graph that references the follower relationships of the users that have tweeted on that specific date, which is from 2018 - 12 - 01 to 2019 - 07 - 31, creating 242 graphs.

4.2.2 Sentiment Analysis

As noted by (Sohangir et al., 2018), lexicon analysis outperforms other methodologies. In our approach, we used VADER (Valence Aware Dictionary and sEntiment Reasoner) (Hutto and Gilbert, 2014) and TextBlob. Both of these tools are part of the nltk library and are pretty easy to use. VADER analyzer returns four scores, the negative, the positive, the neutral, and the compound score, whilst TextBlob returns two scores, the polarity (which should be very close to the compound score) and the subjectivity. We decided to use all of these variables as features in our models, which allows us to compare those two analyzers. Furthermore, to achieve better accuracy on the scores, the tweets must be stripped from any special characters, etc. More specifically, tweets often contain Unicode characters such as the non-breaking space. These characters should be normalized so as not to negatively affect the scoring of the analyzers.

4.2.3 Machine Learning Models

Decision Tree. The Decision Tree (DT) builds regression or classification models in the form of a tree structure. This means that the model breaks the dataset into smaller subsets by asking different questions each time. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches (Decisions), each one representing values for the attribute that was tested. Leaf nodes (Terminal Nodes) represent decisions on the numerical targets. The questions and their order is determined by the model itself using Information Gain (for classification) or ID3 (for regression) (Tzirakis and Tjortjjs, 2017; Tjortjjs and Keane, 2002).

For each question, the model must make a strategic split using a criterion. Decision trees have a lot of advantages for example they are not affected by missing values or outliers. They can handle both numerical and categorical values, and they are very easy to understand. Also, trees can capture non-linear relationships. On the other hand they also have some disadvantages. The most important one is that they tend to overfit to the training sample. A small difference in data might produce a completely different tree. Lastly, there is no guarantee that the tree will be the global optimal.

Random Forest. Random Forest (RF) is another method that uses a tree structure to solve a regression or a classification problem. A random forest is a collection of decision trees, with each tree voting on the final decision. In the training phase, each tree on the

forest considers only a random sample of the data. In the predicting phase, each tree will make a prediction and the average of all of the trees will be considered as the final value.

XGBoost. Boosting and bagging are two methods commonly used in weak prediction trees, such as decision trees, to improve their performance. Those two methods work sequentially, meaning that a new model is added to correct the error of the existing models until no further improvements can be made. XGBoost stands for eXtreme Gradient Boosting, which is a method where new models are created that predict the residuals or errors of existing models and then, added together to make the final prediction. Its name comes from the algorithm used to minimize the loss function, which is called gradient descent.

k-Nearest Neighbors. k-Nearest Neighbors (k-NN) is one of the most basic and essential machine learning algorithms. Like the trees, it belongs to the supervised machine learning algorithms. k-NN is a non-parametric method, meaning that it does not make any assumptions about the distribution of the data. k-NN is a fairly simple model that calculates similarities based on the distances between the data points. When a new entry needs to be classified, the algorithm measures the distances between the new data and all the already classified data. The new entry is then assigned to the class that has the minimum distance to the new data point. There are multiple methods to measure the distance, such as the Euclidean or the Manhattan distance.

LSTM. Simple neural networks cannot understand the context and the order of the data. For that, we need some sort of memory. Recurrent neural networks are a special form of neural networks where their units are connected between each other so the values depend not only on all the units (Hochreiter, 1991). RNNs are extremely important and have been successfully used in a lot of applications, such as speech recognition. But, RNNs suffer from the vanishing gradients problem. This problem refers to the hidden neuron activation functions that are used. If those functions are saturating non-linearities, like the tanh function, then the derivatives can be very small, even close to zero. Multiplying many such derivatives leads to zero, which means that the neural network cannot propagate back for too many instances.

Hochreiter & Schmidhuber (Hochreiter and Schmidhuber, 1997) introduced another kind of recurrent neural networks, the long short term memory (LSTM). Those models have the same "chain" like

structure, but the module responsible for the "repetition" part has a different structure. In a classic RNN, the repetition module is a neural network with a hidden layer, usually with tanh as the activation function. On an LSTM, instead of having a single hidden layer, there are four. On the first stage or gate, as it is called, the neural network decides which information to throw away from the cell state. Continuing to the second stage, the model incorporates the new information and decides what to keep and what to throw away. The model updates the old cell state into the new cell state. In the third stage, the model throws away the old information and adds new information. In this stage, the candidate values are estimated. Lastly, the output values depend on the state of the first and the third layer.

5 RESULTS & EVALUATION

This section presents the results of this research. We present the feature selection and the summary of the results per dataset and per model.

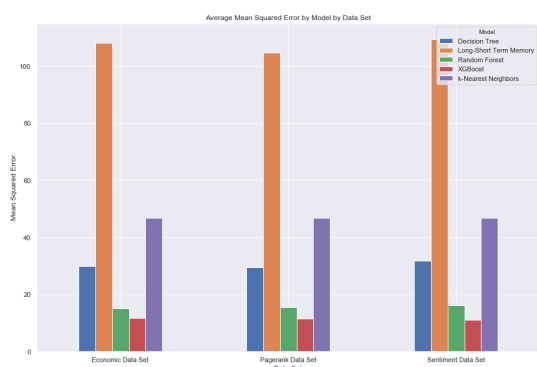


Figure 3: Average Mean Squared Error per Model per Dataset.

All of the scores refer to the mean squared error, thus the best score is the lowest (Figure 3). Lastly, we evaluate our results using a naive trading strategy and comparing it across all datasets regarding our stocks portfolio (Table 1).

5.1 The Trading Strategy

For evaluating results we utilize a naive trading strategy comprising the following points:

1. At the end of each day of our testing sample we sell the stocks that are predicted to have a loss in the next day.
2. We buy the stocks that are predicted to have a positive return.

Table 1: Initial Portfolio.

Ticker	Quantity	Price	Amount
AAPL	1	204,5	204,5
CAT	1	139,09	139,09
HD	1	217,26	217,26
UNH	1	264,66	264,66
XOM	1	75,93	75,93
IBM	1	143,53	143,53
TRV	1	154,59	154,59
V	1	179,31	179,31
BA	1	362,75	362,75
INTC	1	49,17	49,17
GS	1	215,52	215,52
JNJ	1	132,5	132,5
WBA	1	55,81	55,81
DOW	1	52,32	52,32
VZ	1	57,41	57,41
JPM	1	115,12	115,12
PG	1	115,89	115,89
KO	1	52,14	52,14
MSFT	1	137,08	137,08
CVX	1	124,76	124,76
MRK	1	81,59	81,59
CSCO	1	57,62	57,62
UTX	1	133,19	133,19
MMM	1	176,49	176,49
WMT	1	114,76	114,76
MCD	1	213,72	213,72
PFE	1	42,85	42,85
AXP	1	128,06	128,06
DIS	1	144,3	144,3

3. We choose to buy the one that maximizes the return and we do not take into account variance, estimated error or diversification effects.
4. In the next day we firstly update the prices and then we calculate our gains or losses.

5.2 Feature Selection

This section describes the features we created, as well as the descriptive statistics of those features per ticker. We have to note that all of the variables are not available for the day we want to predict, thus all the features created are values of previous days. Since there is no consensus on the literature on which time lag is the most important, for every variable we created the lags from 1 to 3 days prior (Bollen et al., 2011).

One major aspect of this work is to determine if the sentiment data are noisy, and how this can be redeemed, we decided to create three different datasets. The first dataset contains only the lagged economic variables and the lags of previous days' closing prices. The second dataset contains all the features of the economic dataset as well as the sentiment data. Lastly, the PageRank dataset contains all of the features from the sentiment dataset, but the sentiment variables are all multiplied by the PageRank value of the respective user.

One major drawback of calculating daily PageRank values for each user is that the algorithm does not always estimate the importance for all of the users. Thus, we decided to fill all those dates with the mean of each user. After that process, we fill all the residual non-estimated PageRank values with 0. We do that since we wanted to have a timely measure of the importance of the user, and in the cases where this was not feasible, we theorized that the number of the followers of a user does not alter significantly from day to day, so it was a logical assumption to fill any missing values with their respective mean. Lastly, if there was no mean, it means that the PageRank algorithm did not find any importance in the user for any day, thus we filled the residual empty values with 0, marking them as noisy and not important.

5.3 Economic Dataset Evaluation

We began with the evaluation of the economic dataset results, on the XGBoost model. Our predictions suggested that we should sell MRK, MCD, MSFT, V, PFE, DOW, JNJ, WMT, DIS, BA, HD, AXP, CAT, IBM, TRV, MMM, JPM, AAPL, NKE, KO, CSCO, GS, and PG and buy 4 shares of Intel's stock. Our predictions proved correct and Intel's stock recorded a gain, so our portfolio now had a total evaluation of 4.041,61\$. Our decisions for 2019/7/18 also proved correct and, again, we recorded a gain of 0,30%. On the contrary, for 2019/7/19 our decisions lead to a negative return of -0,47%. The biggest gain was observed on 2019/7/30 with a daily return of 1,87%, whilst our worst day was the next day, where we lost most of our gains (-75,99\$). Finally, our cumulative return for the whole period was positive, 0,75%.

5.4 Sentiment Dataset Evaluation

In the sentiment's dataset we began by selling most of our portfolios' stocks and buying only one. More specifically, we sold 23 stocks and bought WBA's stock. This decision was wrong, as we sold Intel's stock, which as we have seen in the previous dataset leads to a significant gain. These decisions naturally lead to a significant loss of -1,91%. Although the next day (2019/7/18) our predictions resulted in a daily positive return of 0,26%, that was not enough to overturn the cumulative negative return. Our best return was 2019/7/29 of 1,39%. Even that return could not reverse our losses, thus the final result of this dataset was a cumulative loss of -3,05%.

5.5 PageRank Dataset Evaluation

For the PageRank dataset in the first day, we sold the following stocks, V, MRK, PFE, JNJ, HD, AXP, WMT, MCD, NKE, CAT, TRV, CVX, JPM, MMM, CSCO, INTC, IBM, KO, PG, DIS, and GS. This decreased the value of bought stocks to 1.343,65\$ and increased the available funds to 2.686,87\$. At this point, 10 units of ticker UNH were bought at 264,66 per unit. This updated the value of bought stocks to 3.990,25\$ and the available funds to 40,27\$. Since we were still on the same day, the evaluation of the portfolio had not changed, because we had not updated the prices yet. On the next day, after updating the prices, we saw that our portfolio had a value of 4.051,48\$, which meant that our strategy and predictions resulted in a positive return of 1,5%.

On the second day, we decided to sell the stocks of VZ, AAPL, and UTX and buy 3 units of Nike's stock. This decision resulted in a loss of 75,88\$ and a total return of -1,3%. The decision was based on the prediction that Nike's stock would have a positive return. On the contrary, the actual result was a loss of -1,07%. We followed the same strategy for every day. We ended up having two stocks, that of XOM's and Intel's on 25/7/19. From then and onwards, the predictions showed that Intel's stock would have a positive return, so according to our strategy we held on to our stocks. This never happened, and our overall return was negative, resulting in a loss of -122\$ or -3,03%.

Table 2 aggregates daily transactions to top daily losses and gains for the investigated period (2018/11/01 until 2019/31/07) as well as the cumulative returns per dataset. Positive values stand for gains and negative values for losses.

Table 2: Top daily Losses & Gains and Cumulative Returns per dataset.

Dataset	Loss (%)	Gain (%)	Return (%)
Economic	-1,83	1,87	0,75
Sentiment	-1,91	1,39	-3,05
PageRank	-1,87	0,86	-3,01

6 CONCLUSIONS

6.1 Summary

This work addresses the problem of predicting stock market data using Twitter data. Sentiment data can have a significant positive impact on the forecasting ability of the models. However, many authors noted

the noisy nature of these data. To redeem that, we proposed a new methodology. By using graphs, we obtained a daily importance measure for all of the users and we weighted their tweets.

Table 3 summarizes the results for the computed errors of all of the stocks. The PageRank dataset performed better than both the economic and the simple sentiment dataset. Moreover, we were able to confirm that the most important feature, on the sentiment data, is the negative score of the tweet. However, we were not able to confirm which time lag is the most important, since results are highly dependant on the feature.

Table 3: Best Dataset Per Ticker.

Ticker	PageRank	Sentiment	Economic
AAPL			✓
AXP		✓	
BA		✓	
CAT	✓		
CSCO		✓	
CVX	✓		
DIS	✓		
DOW	✓		
GS			✓
HD	✓		
IBM		✓	
INTC			✓
JNJ			✓
JPM		✓	
KO	✓		
MCD			✓
MMM	✓		
MRK	✓		
MSFT			✓
NKE	✓		
PFE		✓	
PG	✓		
TRV	✓		
UNH			✓
UTX			✓
V	✓		
VZ			✓
WBA	✓		
WMT	✓		
XOM	✓		

Five different models were tested. For each stock and for each dataset, we estimated a Decision Tree, a Random Forest, an XGBoost, an LSTM, and a k-Nearest Neighbors.

For 15 out of 30 stocks the PageRank dataset performed better than the other datasets. The most important feature of the sentiment data was the negative score. For 13 out of 30 stocks the XGBoost performed better than the other models. We could not confirm which time lag is the most important, as this feature was highly depend and on the stock.

Table 4 presents a summarized version of the results in the PageRank dataset. The best model was XGBoost because it achieved the lowest scores at 13 stocks. Furthermore, it was the most robust model, having the lowest average error and the lowest standard deviation.

Table 4: Best Dataset Per Model on PageRank Dataset.

Ticker	DT	k-NN	LSTM	RF	XGBoost
AAPL	✓				
AXP				✓	
BA					✓
CAT	✓	✓			
CSCO				✓	
CVX					✓
DIS		✓			
DOW	✓				
GS				✓	
HD			✓		
IBM			✓		
INTC		✓			
JNJ					✓
JPM				✓	
KO	✓				
MCD					✓
MMM					✓
MRK					✓
MSFT					✓
NKE			✓		
PFE					✓
PG			✓		
TRV					✓
UNH					✓
UTX					✓
V					✓
VZ				✓	
WBA				✓	
WMT					✓
XOM			✓		

Although PageRank’s dataset provided the best scores for most of the stocks, on the evaluation the economic dataset proved the only profitable (0,75%). The other two datasets recorded of loss of -3,05% and -3,01% for Sentiment and PageRank.

6.2 Limitations

This study acts like a proof of concept that microblogging data can be a powerful feature in predicting stock market data, if we can determine and distinguish the important ones. This is feasible but the required data pose a tremendous obstacle.

Since all of our data come from the Twint library, and not from the official Twitter API, we could collect a specific amount of tweets. Moreover, this library is significantly slower than the official, thus it was very difficult to obtain data for a longer period. We believe

that if we had two years' worth of data and all the tweets per day, then our results would be significantly better.

Lastly, on the evaluation part, we choose a greedy strategy and not an optimal one. The optimal solution would require an extra module that would implement diversification according to Markowitz's Portfolio Theorem (Markowitz, 1991) and the extraction of optimal weights per stock. Moreover, every transaction should move us alongside the efficient frontier.

6.3 Further Research

There are a lot of aspects of our research we want to explore in the future. Firstly, we could estimate more models, such as SVM, which in the literature was used a lot. Another dimension we would like to explore is the economic variables we can choose. There are other useful economic variables that we should embed in our research. Moreover, we could expand our methodology to other financial instruments to explore the possibility that sentiment data can act as features on government and corporate bonds, or even on derivatives. Lastly, as we observed in some models, there were cases where the mean squared error was low, but the fit between the actual and the predicted price was not good. Thus, it would be very helpful if we could define a new measure that can capture the fit better.

REFERENCES

- Alshahrani Hasan, A. and Fong, A. C. (2018). Sentiment Analysis Based Fuzzy Decision Platform for the Saudi Stock Market. In *2018 IEEE International Conference on Electro/Information Technology (EIT)*, pages 0023–0029, Rochester, MI. IEEE.
- Antweiler, W. and Frank, M. Z. (2004). Is all that talk just noise? the information content of internet stock message boards. *The Journal of finance*, 59(3):1259–1294.
- Bae, Y. and Lee, H. (2011). A Sentiment Analysis of Audiences on Twitter: Who Is the Positive or Negative Audience of Popular Twitterers? In Lee, G., Howard, D., and Ślezak, D., editors, *Convergence and Hybrid Information Technology*, volume 6935, pages 732–739. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Belevesslis, D., Tjortjis, C., Psaradelis, D., and Nikoglou, D. (2019). A hybrid method for sentiment analysis of election related tweets. In *2019 4th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNM)*, pages 1–6. IEEE.
- Blystone, D. (2019). Overbought or oversold? use the relative strength index to find out.
- Bollen, J., Mao, H., and Zeng, X.-J. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8. arXiv: 1010.3003.
- Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1-7):107–117.
- Chakraborty, P., Pria, U. S., Rony, M. R. A. H., and Majumdar, M. A. (2017). Predicting stock movement using sentiment analysis of twitter feed. In *2017 6th International Conference on Informatics, Electronics and Vision & 2017 7th International Symposium in Computational Medical and Health Technology (ICIEV-ISCMT)*, pages 1–6. IEEE.
- Chiong, R., Fan, Z., Hu, Z., Adam, M. T. P., Lutz, B., and Neumann, D. (2018). A sentiment analysis-based machine learning approach for financial market prediction via news disclosures. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion on - GECCO '18*, pages 278–279, Kyoto, Japan. ACM Press.
- Devenow, A. and Welch, I. (1996). Rational herding in financial economics. *European Economic Review*, 40(3-5):603–615.
- Fama, E. F. (1965). The behavior of stock-market prices. *The Journal of Business*, 38(1):34–105.
- Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of financial economics*, 33(1):3–56.
- Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438.
- Hagberg, A., Swart, P., and S Chult, D. (2008). Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States).
- Hochreiter, S. (1991). Investigations on dynamic neural networks. *Diploma, Technical University*, 91(1).
- Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Hutto, C. J. and Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*.
- Hájek, P. (2018). Combining bag-of-words and sentiment features of annual reports to predict abnormal stock returns. *Neural Computing and Applications*, 29(7):343–358.
- Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632.
- Koukaras, P. and Tjortjis, C. (2019). Social media analytics, types and methodology. In *Machine Learning Paradigms*, pages 401–427. Springer.
- Koukaras, P., Tjortjis, C., and Rousidis, D. (2019). Social media types: introducing a data driven taxonomy. *Computing*, pages 1–46.
- Kuepper, J. (2019). Timing trades with the commodity channel index.

- Li, R., Fu, D., and Zheng, Z. (2017). An Analysis of the Correlation between Internet Public Opinion and Stock Market. In *2017 4th International Conference on Information Science and Control Engineering (ICISCE)*, pages 150–153, Changsha. IEEE.
- Mahajan, A., Dey, L., and Haque, S. M. (2008). Mining Financial News for Major Events and Their Impacts on the Market. In *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pages 423–426, Sydney, Australia. IEEE.
- Malkiel, B. G. and Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2):383–417.
- Markowitz, H. M. (1991). Foundations of portfolio theory. *The journal of finance*, 46(2):469–477.
- Mitchell, C. (2019). Aroon oscillator definition and tactics.
- Mittermayer, M.-a. and Knolmayer, G. (2006). News-CATS: A News Categorization and Trading System. In *Sixth International Conference on Data Mining (ICDM'06)*, pages 1002–1007, Hong Kong, China. IEEE.
- Oikonomou, L. and Tjortjiss, C. (2018). A method for predicting the winner of the usa presidential elections using data extracted from twitter. In *2018 South-Eastern European Design Automation, Computer Engineering, Computer Networks and Society Media Conference (SEEDA_CECNSM)*, pages 1–8. IEEE.
- Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.
- Park, J., Leung, H., and Ma, K. (2017). Information fusion of stock prices and sentiment in social media using granger causality. In *2017 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, pages 614–619. IEEE.
- Plakandaras, V., Papadimitriou, T., Gogas, P., and Diamantaras, K. (2015). Market sentiment and exchange rate directional forecasting. *Algorithmic Finance*, (1-2):69–79.
- Rousidis, D., Koukaras, P., and Tjortjiss, C. (2019). Social media prediction: A literature review. *Multimedia Tools and Applications*.
- Schumaker, R. P. and Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The AZFin text system. *ACM Transactions on Information Systems*, 27(2):1–19.
- See-To, E. W. K. and Yang, Y. (2017). Market sentiment dispersion and its effects on stock return and volatility. *Electronic Markets*, 27(3):283–296.
- Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance*, 19(3):425–442.
- Sohangir, S., Petty, N., and Wang, D. (2018). Financial Sentiment Lexicon Analysis. In *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*, pages 286–289, Laguna Hills, CA, USA. IEEE.
- Staff, I. (2019). On-balance volume: The way to smart money.
- Suman, N., Gupta, P. K., and Sharma, P. (2017). Analysis of Stock Price Flow Based on Social Media Sentiments. In *2017 International Conference on Next Generation Computing and Information Systems (IC-NGCIS)*, pages 54–57, Jammu. IEEE.
- Sun, T., Wang, J., Zhang, P., Cao, Y., Liu, B., and Wang, D. (2017). Predicting Stock Price Returns Using Microblog Sentiment for Chinese Stock Market. In *2017 3rd International Conference on Big Data Computing and Communications (BIGCOM)*, pages 87–96, Chengdu. IEEE.
- Tetlock, P. C., Saar-Tsechansky, M., and Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance*, 63(3):1437–1467.
- Tjortjiss, C. and Keane, J. (2002). T3: a classification algorithm for data mining. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 50–55. Springer.
- Treynor, J. L. (1962). Jack treynor's' toward a theory of market value of risky assets'. Available at SSRN 628187.
- Tzirakis, P. and Tjortjiss, C. (2017). T3c: improving a decision tree classification algorithm's interval splits on continuous attributes. *Advances in Data Analysis and Classification*, 11(2):353–370.
- Vanstone, B. J., Gepp, A., and Harris, G. (2019). Do news and sentiment play a role in stock price prediction? *Applied Intelligence*.
- Wong, C. and Ko, I.-Y. (2016). Predictive Power of Public Emotions as Extracted from Daily News Articles on the Movements of Stock Market Indices. In *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 705–708, Omaha, NE, USA. IEEE.
- Zhang, X., Fuehres, H., and Gloor, P. A. (2012). Predicting Asset Value through Twitter Buzz. In Altmann, J., Baumöl, U., and Krämer, B. J., editors, *Advances in Collective Intelligence 2011*, volume 113, pages 23–34. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Zhang, Z.-H., Jiang, G.-P., Song, Y.-R., Xia, L.-L., and Chen, Q. (2017). An Improved Weighted Leader-Rank Algorithm for Identifying Influential Spreaders in Complex Networks. In *22017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*, pages 748–751, Guangzhou, China. IEEE.