Stock Sentiment Analyzer

Sanidhya Raghuvanshi, B Sai Gunavanth
Bennett University

ABSTRACT

The purpose of our project research is to implement machine learning to extract informative results in form of stock price trends using sentiment analysis. Scraping through news headlines from leading financial news websites, Stock Sentiment Analyzer pulls out raw data to its own database, and upon the customers requests it advices them whether it is a good time to invest in the organization, hold stock, or cash out with the money. We use different classification algorithms to train a model with an astronomical stock market data of more than past five years. Through vast training of the model, we have achieved a successful prediction rate of 75%. These results illustrate the need of our project to invest smartly in the stock market.

***Index Terms*—Sentiment Analysis, Natural Language Processing, Stock Trends, Web Scraping, Machine Learning.**

I. INTRODUCTION

With the boom of the stock market, there are many upcoming amateur investors joining in who neither have proper knowledge in the field nor have enough time. With abundance of information, it is hard to keep track of useful info which helps us in choosing the right stock or company depending on our needs and risk appetite.

Upon initial research, one of our team members stumbled upon some stock prediction website that used the stock prices in the past to train the model and hence predict the stock price for the future. This, as based only on past prices, was not as accurate as a model should be, and hence gave us the idea to integrate the prediction with sentiment analysis.

It is evident that the stock market depends on public opinion which is not covered by previous stock prices. Instead of past stock prices, a better way to predict a company's state would be how well it is performing and its internal statistics. Ultimately, this gave us the idea of using the most relevant state and accurate of information present in the business field, i.e., the financial news.

We aim on gathering information from online financial news websites like 'marketwatch.com' and 'economic times' and try on various models for machine learning available and settle on one with best results.

To use such stock predictors, he must pay some base fee, and then when he invests some amount on stocks advised by the stock predictor, he still has a considerable chance of losing money.

This is where our project comes in. Stock sentiment analyzer is the idea that uses natural language processing (NLP), machine learning, and other data analysis techniques to analyze and derive effective and quantitative results from news websites, integrates the data with the stock prices in the past, and conclusively advices the user whether to invest in a stock or not. Our idea is to use abstraction effectively, with which we emphasize that we do not intend to give the probability whether the stock will rise or fall (as of now), but rather a judgment whether the user should invest in the specific stock or not.

- Background Knowledge

The research and creation of a sentiment classification model [Pang, Lee, and Vaithyanathan (2002)] trained a model by feeding movie reviews and ratings by the sentiment of the critic.

Some well-known examples of sentiment analysis

can be seen in the Natural Language Processing (NLP) community, where words that are relevant to sentiment are automatically learned from the data. Moreover, Non-linear algorithms like decision trees and support vector machines are also used for mapping the words in a text to its classification.

We have seen many different use cases for sentiment analysis such as movie and product reviews and tweet emotion analysis. Sentiment analysis can also be used by economists as a way of measuring how optimistic customers feel about their finances and the state of the economy.

Papers and projects on tweet sentiment analysis are built everyday as parts of projects in many institutions. Stock market prediction has been an active area of research and has been somewhat successful yet and could use a breakthrough. In this document we will discuss on how the linking of sentiment analysis to a ML model in the stock market field can prove to be a major convenience and how can we find a correlation between the sentiments of financial news and stock market variation.

II. RELATED WORKS

Kalyani et al. (2016) The project analyses a company's financial news stories and applies news sentiment classification to estimate its stock movement in the future. They noticed that news stories have an impact on the stock market's movement. For this, they used a dictionary-based technique. General and finance-specific sentiment carrying phrases are used to create positive and negative dictionaries. They used this data to develop categorization models. According to the results, Random Forest (RF) and Support Vector Machine (SVM) perform well in all tests. [1]

Khedr, Salama, and Yaseen (2017) want to construct a model with a low error ratio that can forecast future stock market changes and improve prediction accuracy The K-NN, and naïve Bayes algorithms were used to get the results, which are based on sentiment analysis and historical stock market prices. The model can be broken down into two parts. The naive Bayes approach is utilized in the first stage to determine if the news is good or bad, and the K-NN algorithm is used in the second stage to forecast the future stock trend based on the first stage's results and the processed historical numeric data. [2]

Because the initial weight of the random selection issue is easily prone to erroneous predictions, traditional neural network algorithms may incorrectly predict the stock market when researching the influence of market factors on stock prices. Pang et al. (2020) use deep learning to create word vectors to exemplify the concept of a "stock vector." Rather than a single index or single stock index, the input is multi-stock high-dimensional historical data. To predict the stock market, they advocate employing a deep long short-term memory neural network (LSTM) with embedded layer and a long short-term memory neural network with automatic encoder. [3]

Krishnamoorthy (2017). His work is a hierarchical sentiment classification model based on association rule mining that predicts the polarity of financial writings as positive, neutral, or negative. He compared the model to other current dictionary and machine learning-based techniques. [4]

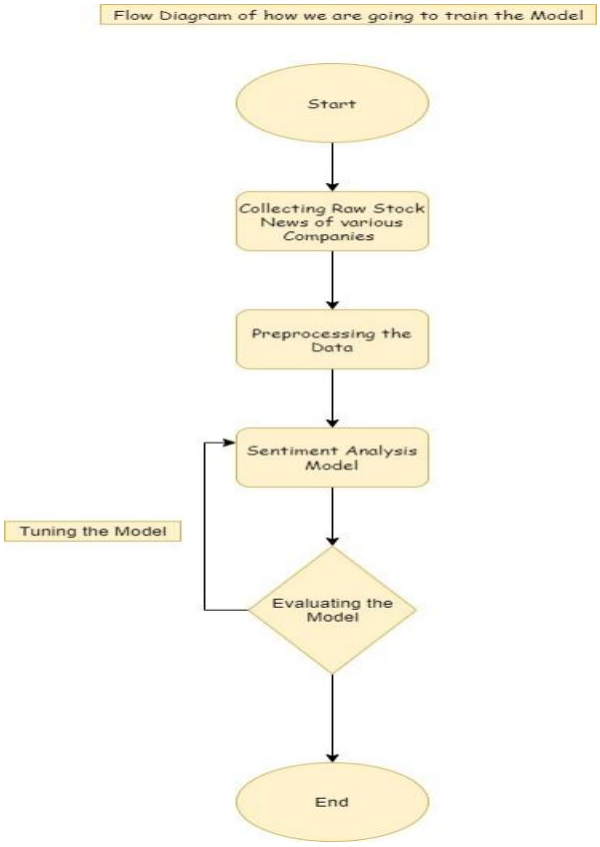
Anita et al (2020). Their research focuses on using sentiment analysis to analyze financial news. Initial estimates of document semantic orientation are made by fine-tuning an existing financial domain approach. In terms of selecting representative phrases that effectively communicate the text's sentiment, the existing method has limitations. The researchers evaluated two alternative strategies: one that uses noun-verb pairings and the other that is a hybrid. [5]

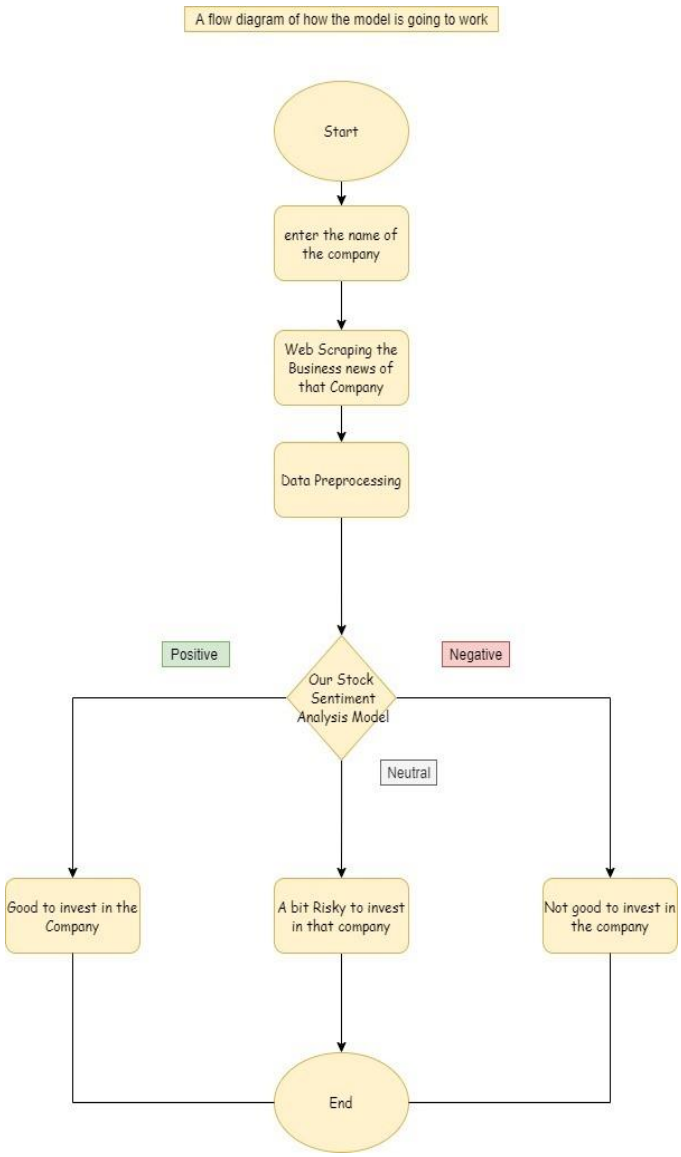
Kirange and Ratandeep's research focuses on the impact of emotional classification of financial news on stock market price prediction. They compared the feelings of two businesses (Infosys and Wipro) over a ten-year period to see if there was a link between sentiment projected by news and original stock ice. Various classifiers are tested, including Naive Bayes, K-NN, and SVM. [6]

Goel et al used sentiment analysis and machine learning concepts to discover the relationship between "public sentiment" and "market sentiment" in their article. They used Twitter data to forecast public sentiment, then combined that forecast with the previous day's readings to forecast stock market moves. To validate their findings, they used Self-Organizing Fuzzy Neural Networks (SOFNN) to achieve 75.56% accuracy. [7]

III. PROPOSED METHODOLOGY

- Flow Diagram





- Description

There are two flow diagrams for this project. The first flow diagram represents of how we are going to train our natural language processing model. The first step of training the model is to collect the data from various sources in the internet like Kaggle and news headlines from news websites. So, after collecting the information we are going to pre-process the data. Pre-processing is a very important step because it filters out the information that is unnecessary and the information that is not in correct format. To fit the data into a machine learning model, pre-processing is a mandatory step so that our model wouldn't give any wrong outputs/results. Thereafter we would try out different machine learning algorithms like Naïve Bayes, Support Vector Machines and Decision Trees. We will try out Algorithms with

different parameters and evaluate which algorithm Performs the best. Our model is going to analyze or predict whether the news is positive or a negative one. This is how we would train our model.

The second flow diagram represents how the final product is going to work, the first step is performed by the user in which he/she provides the name of the company they want to know about. So, after the user provides the name, we are going to scrap the recent news of that company from various websites on the internet. Web Scraping is a very powerful technique which helps us to gather information from the news websites or any website it could be. So, once we gather all the information, we'll start by pre-processing the data.

Pre-processing is a critical step since it removes superfluous information and information that is not in the proper format. Pre-processing is required to fit the data into a machine learning model so that our model does not provide incorrect outputs/results. So once the pre-processing is done the information will be sent to the model which we have trained before, so this model is going to analyze the data very carefully and tells the information is a positive one or a negative one. If it analyses and says that a business's news is positive, it suggests that it is a good time to invest in that company; if it analyses and says that the company's news is bad, it shows that it is not a good time to invest in that company. When the percentages of negative and positive news are identical, the models conclude that investing in that company is neither risky nor profitable. This is how the model's is going to work.

- News collection

For fetching the live data, we have used an API called news API which fetches the headlines of a company that we have mentioned for a period of last two months. At first, we have logged in to the news API website and they have provided us an authentication token to fetch the data through the URL they provided. The data is of json type. So, we had to parse the data by using some programming libraries like json decoder. Now using the decoder, we have decoded the data and we have put that into a text file. This text file will be used to determine the fate of the company.

- Preprocessing

We stored the data collected from web-scraping to a csv file on the system. And then, we start with preparing our data for preprocessing.

Using Pandas we imported our dataset in a pandas data-frame. The features were dates, headlines, and our target variable label. After this we start with preprocessing our data.

Firstly, we mapped our target variable from string object to int by converting positive to 1, neutral to 0 and negative to -1.

Then we cleaned our headlines using Pythons Regular Expression library. We removed special characters, any unwanted spaces, non-English words, and lowered all words to remove redundancy for our model. We used TF-IDF vectorizer because it gives the importance and high value to the most frequent words. After preprocessing we did some exploratory analysis on our data.

Using python's matplotlib.pyplot, we plotted a pie chart based on our data we also created a bar plot for better visualization. Apart from that we explored the most frequents words in our dataset

After all this work, our dataset was ready to move to next step which is Model creation.

To show the relation between stock trend and news, we have created different algorithms to predict the sentiment of news as positive, negative or neutral.

In most of the cases naive bayes, random forest, svm classifications performs well in text classification. we executed all three algorithms and checked accuracy of each algorithm, and after tuning model and tweaking parameters svm and random forest gave better accuracy compared to naive bayes.

IV. RESULT AND ANALYSIS

Starting with our dataset, initially we faced a major issue where around 60% of our data was neutral. After research, we used under sampling to improve

the class distribution.
But because of this we faced a problem, first the data size was reduced and thus our results were not prominent.

By using naïve bayes initially our accuracy came to be around 60%. Then, random forest the accuracy increased slightly to 65%. Finally we tried SVM, which gave the best result across all algorithms at 73%.

We used Stratified K-fold instead of K-fold because the former distributes the label equally to the splits.

In Random Forest we used GridsearchCV to tweak and find the best parameters.

#Accuracy	70% Data Split	80% Data Split
Random Forest	71.82 %	72.45 %
SVM	76.85 %	77.17 %

#Precision	70% Data Split	80% Data Split
Random Forest	0.71	0.71
SVM	0.75	0.75

#Senstivity	70% Data Split	80% Data Split
Random Forest	0.56	0.60
SVM	0.69	0.70

#Specificity	70% Data Split	80% Data Split
Random Forest	0.83	0.81
SVM	0.83	0.82

5-Cross Validation	Maximum	Minimum	Overall
Random Forest	75.6 %	71.65 %	73.23%
SVM	78.8 %	75.22 %	77.4 %

10-Cross Validation	Maximum	Minimum	Overall
Random Forest	77.1 %	73.4 %	74.7 %
SVM	82.14 %	75 %	78.11 %

15-Cross Validation	Maximum	Minimum	Overall
Random Forest	78.2 %	70.1 %	73.4%
SVM	83.1 %	73.9 %	78.1 %

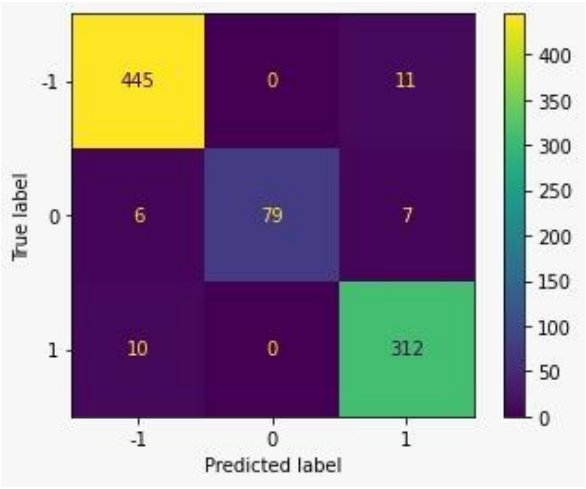


Fig: Random Forest Confusion Matrix

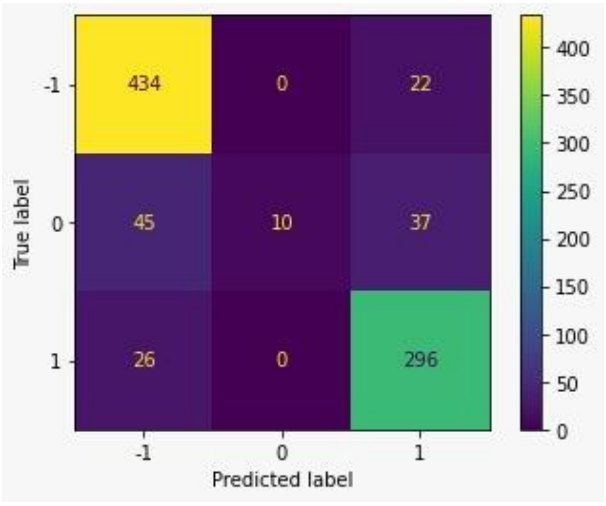


Fig: SVM Confusion Matrix

The exploratory analysis of our dataset includes pie chart, bar plot.

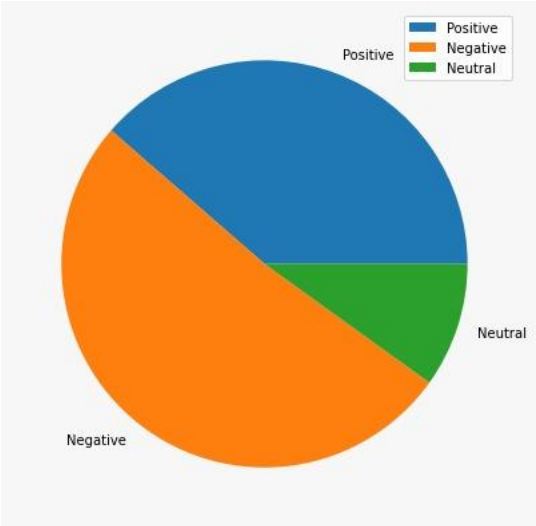


Fig: Pie Chart

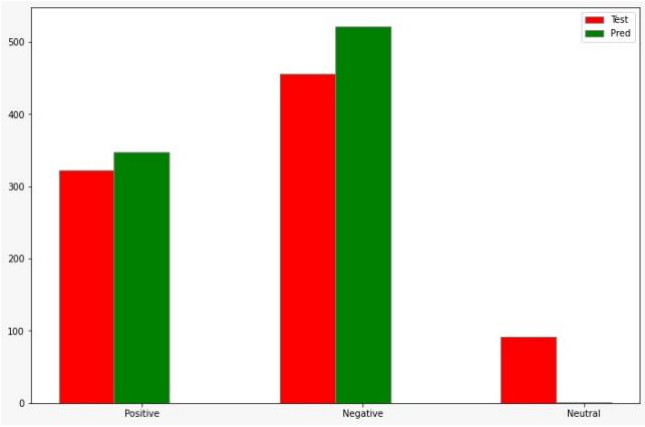


Fig: Bar Plot

V. Conclusion

Prediction of stock prices based on previous trends is a complicated yet demanding task since stock trends do not depend on a single feature. A prerequisite for the research is an assumption that day-to-day ups and downs of the market are correlated with financial news. Overnight changes in management or uncontrollable price variations are reflected in the news prior to having effect on actual prices. Therefore, we have exhaustively studied this relationship and concluded positive about the correlation. We also detected significant differences in the correlation matrices while comparing Random Forest, Naïve Bayes and SVM.

As news articles capture emotions about the current market, the analyzer automates that data and forms a polarity. If this news impact is good in the market, there are a lot of opportunities that the stock price goes up. If there is no concrete evidence or not enough for the model to make a decision, it recommends holding the stock. This algorithm is used to make the training dataset. Based on this data, we created three classification models and tested under different test conditions.

As a result, the Random Forest performed very well in all experimental cases ranging from 65% accuracy. The accuracy of the Naive Bayes is about 60% and SVM 75%. Given any financial news topic, the model is likely to generate a polarity that can continue to predict stock trends.

VI. References

[1] Joshi, K., N, B. H., & Rao, J. (2016). Stock Trend Prediction Using News Sentiment Analysis. International Journal of Computer Science and Information Technology, 8(3), 67-76. doi:10.5121/ijcsit.2016.8306

[2] Khedr, A. E., S.e.salama, & Yaseen, N. (2017). Predicting Stock Market Behavior using Data Mining Technique and News Sentiment Analysis. International Journal of Intelligent Systems and Applications, 9(7), 22-30. doi:10.5815/ijisa.2017.07.03

[3] Pang, X., Zhou, Y., Wang, P., Lin, W., & Chang, V. (2018). An innovative neural network approach for stock market prediction. The Journal of Supercomputing, 76(3), 2098-2118. doi:10.1007/s11227-017-2228-y

[4] Krishnamoorthy, S. (2017). Sentiment analysis of financial news articles using performance indicators. Knowledge and Information Systems, 56(2), 373-394. doi:10.1007/s10115-017-1134-1

[5] Yadav, A., Jha, C. K., Sharan, A., & Vaish, V. (2020). Sentiment analysis of financial news using unsupervised approach h. Procedia Computer Science, 167, 589-598. doi:10.1016/j.procs.2020.03.325

[6] W. P. Risk, G. S. Kino, and H. J. Shaw, "Fiber-optic frequency shifter using a surface acoustic wave incident at an oblique angle," *Opt. Lett.*, vol. 11, no. 2, pp. 115–117, Feb. 1986.

[7] Mittal A, Goel A, "Stock Prediction Using Twitter Sentiment Analysis" *Stanford.edu.*