Twitter Sentimental Analysis for Stock Recommendation

TERM PROJECT PROPOSAL (Group 7)

SUBMITTED BY

Rohan Naga Venkata Mayukh Ungarala - 016000127

Harshitha Mohanraj Radhika - 016033849

Janani Ravi Kumar - 016044080

rohannagavenkatamayukh.ungarala@sjsu.edu

harshitha.mohanrajradhika@sjsu.edu

janani.ravikumar@sjsu.edu

Department of Applied Data Science, San Jose State University, San Jose

DATA 228 - Big Data Technologies and Applications

Professor Dr Ming - Hwa Wang

Contents

ACKNOWLEDGMENT

It has been a great honor and privilege to be enrolled in SAN JOSE STATE UNIVERSITY, SAN JOSE for M.S in Data Analytics. We are very grateful to Prof Dr. Ming Hwa Wang for giving his valuable time and constructive guidance in doing this project. We also would like to thank everyone who has helped us, directly and indirectly, to complete this project successfully.

1 Abstract

Stock market prediction is a common and well-known problem of interest. In our day to day life social media is perfectly representing and updating the public sentiments and opinion about current events and incidents. With the real-time information available to us on massive social media platforms like Twitter, we have all the data we could ever need to create these predictions. Twitter in particular has attracted a lot of attention from researchers for understanding the public sentiments. Stock market recommendation based on public sentiments expressed on Twitter has been an intriguing field of research. By using Sentiment analysis, we can take thousands of tweets about a company and judge whether they are generally positive or negative (the sentiment) in real-time data and incidents.

A lot of work is being carried out in the Stock Market domain using Data Analytics. Individual behaviors and decisions are greatly affected by emotions and by the opinions of others. Previous studies have concluded that the aggregate public mood collected from Twitter may well be correlated with commodity prices of individual companies. The thesis of this work is to observe how well the changes in stock prices of a company, the rises and falls, are correlated with the public opinions being expressed in tweets about that company. The paper also suggests the buyer's profit in the stock market by recommending them. The objective of sentiment analysis is understanding the author's opinion from a piece of text. In this paper, we have applied sentiment analysis and supervised machine learning principles to the tweets extracted from Twitter and analyze the correlation between stock market movements of a company and sentiments in tweets. In an elaborate way, positive news and tweets in social media about a company

would definitely encourage people to invest in the stocks of that company and as a result the stock price of that company would increase. At the end of the paper, it is shown that a strong correlation exists between the rise and falls in stock prices with the public sentiments in tweets listed company wise based on the polarity calculated. Stock investors and financial planners use various time series and calculative methods for market prediction techniques and analysis. There is a higher possibility that a stock price will continue to rise and the investor gains profit if the sentiment is positive, and the stock price will fall if it is malicious.

2 Introduction

A **STOCK** is an instrument that gives ownership in a company or corporation and represents a proportionate claim on what the company owns and what the company generates in profits. Stocks may be also known as shares or company's equity.The potential return from deposits,gold, and treasury bonds are less compared to stocks and this is the primary reason why the number of people who invest in stocks have increased over the years. It's easier to buy and sell them. stocks are liquid investments compared to real estate investments.**TWITTER** is a famous American social networking site that is used by people to create posts and interact with messages known as “tweets”.It was launched on March 21st 2006.Twitter’s annual users has increased from 110 million users to 186 million users in the past 5 years.Moreover, social media provides an opportunity for businesses by giving a platform to connect with their customers for advertising. People mostly depend upon user generated content over online to a great extent for decision making.

2.1 Objective

The nature of the stock price movement is non linear.The main motivation of our project is to find a way to generate a consistent profit in the stock market which is a challenging task indeed.People are unsure of what stock they should opt at what time to sell or buy it.The main objective behind this project would be to develop a model that would help users decide on which stocks to opt based on the tweets as there is always a correlation between the moods of public and the way the stock market performs.Since machines can't understand the emotions of the person in this project we use some machine learning algorithms to gauge the mood expressed in tweets.

The end deliverable of this project is a financial model that would be built by analyzing the data collected by using Twitter API.The goal is to provide an understanding to people or stockholders about the important features that determine the stock market.In this process we are going to analyze the retrieved tweets,perform data cleaning,data wrangling,data modeling and present the final study with visualizations.

2.2 Problem Statement

Stock market prediction is one of the most difficult things to do.There are various causes for this, one is market volatility and next one is variety of dependent and independent variables that creates impact on certain stock in the market.These dependent and independent variables make it difficult for any stock market expert to predict the rise and fall of the market with great precision.The area of stock recommendation consists of research which focuses on two areas one is the stock recommendation methods based on comments in social media or public mood and the

second is forecasting the price with comparison to the historical prices.The former method is understandable and is used by the stockholders or investors.In this complicated stock market, the stockholders or investors do not know which one is trustable among the lots of comments and choosing among the comments has a great risk.The next method is difficult for investors to understand,is complex and also needs a lot of mathematical knowledge.A machine learning model that can be used with former method is the solution for this problem.

The emotions and moods of individuals are reflected in their tweets and thus affect the decision making process of a few stockholders or investors, which shows there is a direct correlation between public sentiment and market .The sentiment analysis is performed on Twitter data called tweets to find the public mood.Using the Textblob the global polarity is checked and based on it stock recommendations are made.

2.3 Project Relevance

For this project we will be using two data's, one is a CSV data file that contains the details about the companies and the other data file that contains tweets retrieved from twitter.We would be sourcing these data from twitter for various companies to build the prediction model.The overall size of the dataset is expected to be very large and this makes it a right choice for a Big data analytics project. We would be further performing data pre-processing,feature extraction and model development using pyspark libraries like Pyspark API and also other libraries.This makes our project very relevant to this course.

2.4 Existing approaches

We have referenced a lot of papers who have done research on stock recommendations which we are explaining in this section.Some of the researchers have used Efficient Market Hypothesis and sentiment analysis.They have proposed a learning model that analyzes the stock trend based on tweets and historical data.There are lot of papers and researches done in order to predict the stock market to achieve the defined metrics. Those models used if-then-else rules, Artificial Neural Network (ANN), Fuzzy systems, Bayesian algorithms and so on to create a model to predict the stocks.

Our work is based on Bollen et al's strategy which had widespread media coverage. They also predicted the behavior of the stock market by the mood of people on Twitter. Those authors considered the tweet data of the users in 2008 and used the OpinionFinder and Google Profile of Mood States (GPOMS) algorithm to classify public sentiment into 6 categories, namely, Calm, Alert, Sure, Vital, Kind and Happy.They then validated the resulting mood time series by detecting the public's response towards the presidential elections and Thanksgiving day in 2008.They also used causality analysis to investigate the public mood states, which was measured by the OpinionFinder and GPOMS mood time series.The authors used Self Organizing Fuzzy Neural Networks to predict DJIA values using previous values.

2.5 Why our project is a better approach

We will follow a computation procedure by applying tokenization (based on relatively simple regular expressions) to the tweets collected through Twitter API.In this project we

aim to provide a comparative study of different models for predicting the stock and provide a consolidated report on why a particular model would be better and why another model would not which was not identified in any previous paper.

2.6 Area of scope and investigation

Decisions should be made before investing in stocks.There are many features that affect the stock prices like economic trends,expectations and so on.For our future work we would really like to build model that will consider different features and predict the values and not just depending on tweets or the emotions of the people.Also once our model is done successfully,based on the output we will work to implement more features and also try different machine learning techniques to get the most accurate model for recommending the stock prices.

We also have plans to consider the situation like covid,or other economical issues and the companies performance that has the impact on the stock prices.For eg the price of the stock decreases for some companies when their production decreases.We would like to retrieve data based on these features also so that we can identify all the factors that affect our goal as our existing dataset does not have all this information.We would also work for a model in which we can compare the trends both in newspapers and tweets find the most truthful statements and thus make a more perfect predicting model than the one which are creating now. For making recommendations we are planning to use textblob which checks the global polarity and increases the count accordingly. Instead of Textblob there are various other libraries for natural language processing which we can use to make the recommendations more precisely. Our dataset doesn’t

map the exact sentiment of the public,it just considers the english speaking people who use twitter. We in future try to obtain a higher correlation if the actual mood is studied.

3 Theoretical Bases and Literature Review

3.1 Problem Definition

Sentiment analysis is the most important concept of the research area in various different fields of study. Data collection is the initial part carried out for a significant period of time by different authors. It can be analyzed that a period of 1 year's data is sufficient for predicting the stock market with the help of web financial data available online daytoday. One way to find out the public mood from tweets is to use tools which compute the mood time series values from the given time series data such as Opinion finder and GPOMS (Google Profile of Mood States) results into 2 dimensional and 6 dimensional mood time series. These Mood time series can be further correlated with stock price time series to judge relevance. Sentiment analysis can additionally be performed using various machine learning algorithms, lexicon based or hybrid methods of approaches. Sentiment analysis is a very difficult task due to sarcasm. The words or text data implied in a sarcastic sentence come with a different sense of meaning in the tweets depending on the senders or situations. Sarcasm is remarking someone opposite of what you want to say on Twitter. The objective is to learn and investigate how machine learning techniques can be used to identify trends.

3.2 Theoretical background

The high volume of user-generated content that is created on social media sites every day is an important factor. we would consider this trend and continue with exponentially more content in the future. The challenge would be critical to address management and utility of massive user-generated data.The source of data is from Twitter and consists of relevant tweets, along with their sentiment derived. The data were collected by Twitter Search API, where a search query consists of the stock cash-tag (e.g., “$NKE” for Nike). All the available tweets with cash-tags are acquired. The Twitter restriction of 1% (or 10%) of tweets applies to the Twitter Streaming API platform, and only in the case when the specified filter (query) is general enough to account for more than 1% (or 10%) of all public tweets extracted.

Text mining techniques are used to extract semantic characteristics from review texts. Semantic characteristics are more influential and important than other characteristics in affecting how helpful and supported vote reviews are received from twitter. Despite the high quality of the datasets used, the level of empirical correlation between stock price derived financial time series, yahoo finance and web derived time series remains limited, especially when a textual analysis of web messages is applied. This observation suggests that the relation between these two systems is more complex and a machine learning algorithm that a simple measure of correlation is not enough to capture the dynamics of the interaction between the two systems. It is possible that the two systems are dependent only at some moments of their evolution, and not over the entire time period using linear regression. A sentiment analysis task is usually modeled as a classification and grouping problem, whereby a classifier is fed a text and returns a category, e.g. positive, negative, or neutral from the tweets.

3.3 Related Research

There are many researchers that study and aim to identify a method to predict sentiment analysis on different area fields of study. Social media is a popularized source of data which collects useful data such as blogs, micro-blogs, Facebook, Twitter etc. Many studies have used text mining approaches to study the impact of news on market behavior for analyzing the stock market. Schumaker et al. (2012) [1] applied positive and negative sentiment analysis to subjective news articles to predict price direction and trading return of the stocks.

Yu et al. (2013) proposed a contextual entropy model to create a thesaurus of emotion words that express the public emotions and their corresponding intensities from online stock market news articles. An entropy measure was used to calculate the similarity between the seed words and candidate words on twitter and then used to classify the sentiment collected from the platform of the news articles.

Hagenau et al. (2013) used a combination of advanced feature extraction methods and a feedback-based feature selection to boost classification accuracy percentage and improve sentiment analytics of twitter. According to them, feature selection significantly improves classification accuracy by reducing the number of less-explanatory features, i.e. noise, and may limit negative effects of over-fitting when applying machine learning approaches to classify text messages extracted from tweets.

3.4 Advantages and Disadvantage of previous research

Most of the literature uses a lexicon based approach in their paper, but this requires a good and powerful dictionary which is not always available. When performing sentiment

analysis from twitter, considering semantics of the text also plays a major role, which is usually not taken care of in most of the cases. It has been noted that if Semantics are incorporated with sentiments, prediction accuracy increases and the machine learning methodology also decides the accuracy of prediction.They observed that specific emotions and opinions of tweets (calm) were more correlated with stock movements in the market than generalized ‘positive’ tweets, confirming a superiority of this method over volume and types of sentiment classification methods of prediction analysis. The ability to use Twitter data to predict stock market movements is used by the previous research including correlation.

Despite the possible positive outcomes shown in previuos papers, there are some disadvantages in applying automatic analysis due to the difficulty to implement it because of the ambiguity of natural language, machine learning and also the characteristics of the posted content. The analysis of tweets is an example of this, for they are usually coupled with hashtags, emoticons and links, creating difficulties in determining the expressed sentiment of the tweets. In addition, there is a need for automatic techniques and deployments that require large datasets of annotated posts or lexical databases approach where emotional words are associated with sentiment values. Another important aspect is that analyses are suitable for the English language, in which there is a limitation for other languages used.

3.5 Solution to solve the problem

When preprocessing tweets, we will be removing URLs because they normally do not represent relevant content and contain extra characters that require processing. We

also removed cash-tags (e.g., “$FRE”) and username mentions (e.g., “@jamesdoe”) to make a tweet independent of a specific stock (company) in the stock market and users involved in the discussion, and thus make the first step towards generalizing our model and processing the data. We collapsed letter repetitions (e.g., “greatttt” becomes “great”). This step is relatively easy to implement and has proven useful for sentiment classification tasks and text processing. After these steps, we followed a typical bag-of-words computation procedure by applying tokenization (based on relatively simple regular expressions) to the tweets collected through Twitter API.

Considering Twitter is updated hundreds of millions times a day this seems an insignificantly big-sized sample to be a fair representation of the population’s mood and opinion. Having said that, the sample of this study should aim to be a good representation of the financial discussions on Twitter by the UK population, and as the total number of them is unknown, it is difficult to say whether the sample of Tweets is a fair representation of the general mood of the public. The second and more detrimental issue to the study is the size of the specified sample period. Although the 6-day period was specifically chosen due to the aim of the study to measure the existence of a short-term relationship only, based on the study of an event (Stock Market). The size of the sample was not big enough for the relationships to be classed as significant at the 95% certainty, due to the fact that relationship trends are not prominent enough in a smaller sample. However, just because a relationship is found as not statistically significant, it does not mean that the relationship does not exist. Finally, it should be noted that the mood expressed in this study is reflective of the discussions specifically related to the issue of public opinion. General day-to-day financial sentiment was not measured as a

part of this study and the relationship with the stock market may vary to the observations as per this paper and our study.

We find that two things matter in tweet content: sentiment and subject matter. We will be showing that tweets with positive or negative sentiment were consistently linked with a reduction in permanent price impact and an increase in temporary price impact, as measured by the variance in the stock price market. The results show that: 1) the fluctuation of stock prices is more sensitively to the intraday sentiment of individuals investing in the stock market; 2) There is no significant correlation between general sentiment of individual investors and the evaluation of consulting institutions, though the two factors both surely influence the change of stock price in the market.

Our design algorithm will look at only the important words like “pretty” , “impressed”, “good” etc.(and not words like “I”, “am” etc., also our Algorithm may not know spaceX , Elon, Musk so it will probably just ignore those tweets). Considering that the Algorithm has previously seen how positive and negative text looks(while training it), it has already figured out that words like “pretty”, “impressed”, “great” are mostly associated with positive emotions. Hence it is likely to label the text as positive and increments the positive counter**.**

4 Hypothesis

Hypothesisis basically is based on precision score, recall and f1. Basically, precision score is the fraction between truly positive and negative observations. Next is the recall, is the ratio of correctly predicted positive observations to the all observations in actual

class. Finally, F1 score is the weighted average of Precision and Recall. Therefore, the result takes both false positives and false negatives into account.

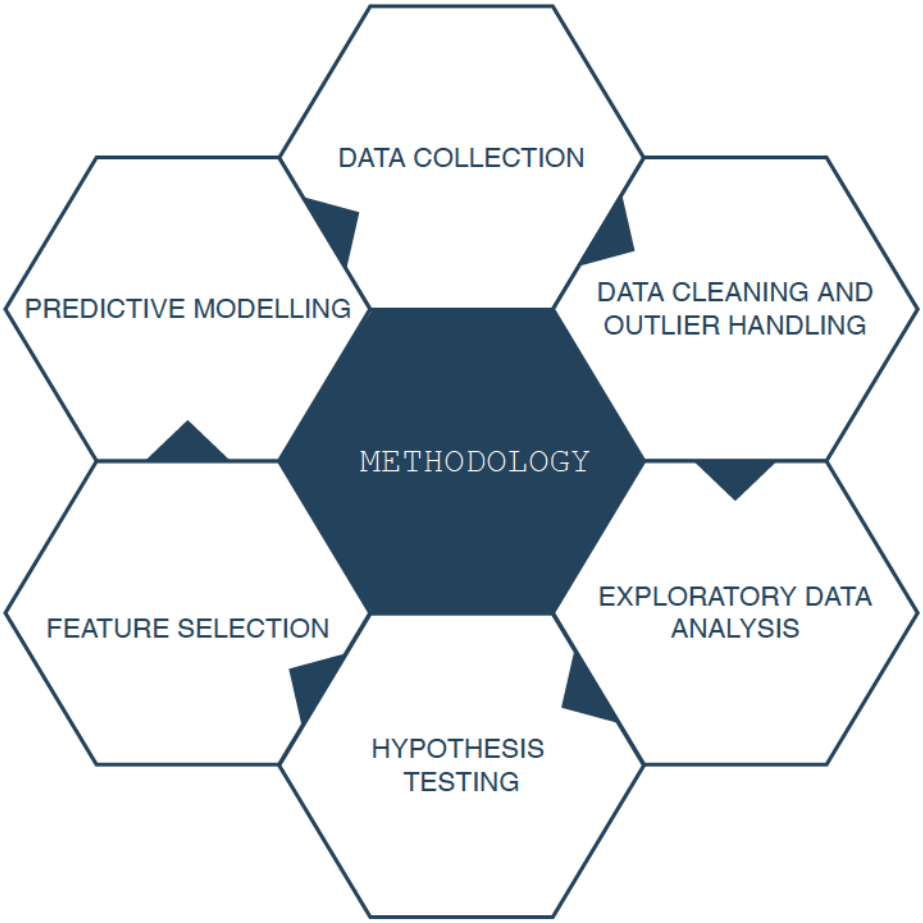
4.1 Single/Multiple Hypothesis

We want to create a good model to classify whether the tweets are positive or negative . It is hard to classify only based on user tweets. So, we consider both user profile description (word embedding) and user tweets. Precision score, recall, f1 are the factors we use.

4.2 Positive Hypothesis

The daily total number of positive tweets is aggregated as Mt Positive while total number of negative tweets as Mt Negative. LinearRegression is used in this application to do prediction for stocks and once the stock predictions are done,we are using the tweets that are retrieved from twitter with the help of twitter development account which matches the symbol of the companies.

5 Methodology



5.1 Collect Input Data

Each tweet is a summary of a person's mood or opinion about a certain subject, then the aggregate of tweets about the subject should express the collective mood. Twitter puts a limit of 280 characters for users to update any post which is always easier to analyze. Tweets need to be filtered down according to our requirements to get the useful data. Filtering of tweets is based on the different keywords which may represent the explicit moods of the public.Tweets are in the form of unstructured data, which needs to be narrowed down and pre-processed.

The first step of stock prediction is **data collection**, which includes tweets from the past for getting reliable predictions related to the stock market. Currently more than 250 million messages are posted on Twitter everyday.

Relevant and focused tweets are collected by using Twitter's Search API. Commodity prices, referred to as economic indicators are also extracted from source. These economic indicators include Gold, Silver, Copper, Aluminum, Crude Oil and Natural Gas.

The system is based on Twitter sentiment analysis retrieved from tweets containing the symbol as text inside it. The API keys from Twitter are collected from Twitter Developer Platform by creating a twitter account. The GET /tweets endpoint provides developers with public Tweet data for requested available tweets in twitter platform. The number of tweets to be retrieved for the sentiment analysis is collected. A Tweet object contains public Tweet metadata such as id, text, created_at, lang, source, public_metrics. To get data from the Twitter API, we first use the predefined credentials to authenticate the connection to the API. After the authentication, we stream the tweet data objects that contain a selected keyword and language**.**The response for the Tweets request will look like the JSON payload.

A CSV file is a comma-separated contains values, which allows data to be saved in a tabular format. CSVs look like a garden-variety spreadsheet but with a . csv extension. CSV files can be used with most any spreadsheet program project, such as Microsoft Excel or Google Spreadsheets applications. The input data to filter the tweets for the Stock recommendation system is done by creating a CSV file, containing the list of companies with their stock symbols. The CSV file contains 8 columns indicating the

symbol, Name of the Company, Lastsale, Market capitalization, ADR TSO whether it contains or considered to be not applicable, IPO (initial public offering), Sector the company belongs to and the Industry classification along with the summary quote. The symbols of the companies listed are unique which is identified as a primary key for each company.

To check if the stock symbol introduced is valid or if it exists, confirm if it appears in the csv file that contains every stock symbol available. This csv file is read and loaded into the data frame.

	Symbol	Name	LastSale	MarketCap	ADR TSO	IPOyear	Sector	Industry	Summary Quote
2	PIH	1347 Property Insurance Holdings, Inc.	7.3962	44264526.29	n/a	2014	Finance	Property-Ca	https://www.nasdaq.com/symbol/pih
3	PIHPP	1347 Property Insurance Holdings, Inc.	26.67	0	n/a	n/a	Finance	Property-Ca	https://www.nasdaq.com/symbol/pihpp
4	TURN	180 Degree Capital Corp.	2.38	74069317.56	n/a	n/a	Finance	Finance/Inv	https://www.nasdaq.com/symbol/turn
5	FLWS	1-800 FLOWERS.COM, Inc.	14.65	946448937	n/a	1999	Consumer	Other Speci	https://www.nasdaq.com/symbol/flws
6	FCCY	1st Constitution Bancorp (NJ)	22.3	186859326.6	n/a	n/a	Finance	Savings Ins	https://www.nasdaq.com/symbol/fccy
7	SRCE	1st Source Corporation	55.93	1452226757	n/a	n/a	Finance	Major Banks	https://www.nasdaq.com/symbol/srce
8	VNET	21Vianet Group, Inc.	8.94	550748288.8	61604954	2011	Technology	Computer S	https://www.nasdaq.com/symbol/vnet
9	TWOU	2U, Inc.	72.13	4092384198	n/a	2014	Technology	Computer S	https://www.nasdaq.com/symbol/twou
10	TPNL	3PEA International, Inc.	3.65	0	n/a	n/a	n/a	n/a	https://www.nasdaq.com/symbol/tpnl
11	JOBS	51job, Inc.	72.49	2641258036	36436171	2004	Technology	Diversified C	https://www.nasdaq.com/symbol/jobs
12	EGHT	8x8 Inc	22	2049371544	n/a	n/a	Public Utilit	Telecommu	https://www.nasdaq.com/symbol/eght
13	AVHI	A V Homes, Inc.	21.3	476537253.3	n/a	n/a	Capital Goo	Homebuildi	https://www.nasdaq.com/symbol/avhi
14	SHLM	A. Schulman, Inc.	43.7	1289549942	n/a	1972	Basic Indus	Major Chem	https://www.nasdaq.com/symbol/shlm
15	AAON	AAON, Inc.	39	2042127009	n/a	n/a	Capital Goo	Industrial Ma	https://www.nasdaq.com/symbol/aaon
16	ABEO	Abeona Therapeutics Inc.	12.9	610238176.5	n/a	n/a	Health Care	Major Pharn	https://www.nasdaq.com/symbol/abeo
17	ABEOW	Abeona Therapeutics Inc.	8	0	n/a	n/a	Health Care	Major Pharn	https://www.nasdaq.com/symbol/abeow
18	ABIL	Ability Inc.	5.99	15432725.85	n/a	2014	Consumer	Telecommu	https://www.nasdaq.com/symbol/abil
19	ABMD	ABIOMED, Inc.	368	16514467728	n/a	n/a	Health Care	Medical/Der	https://www.nasdaq.com/symbol/abmd
20	ABP	Abpro Corporation	n/a	0	n/a	n/a	n/a	n/a	https://www.nasdaq.com/symbol/abp
21	AXAS	Abraxas Petroleum Corporation	2.08	346470086.6	n/a	n/a	Energy	Oil & Gas Pr	https://www.nasdaq.com/symbol/axas
22	ACIU	AC Immune SA	9.1	522345141.5	n/a	2016	Health Care	Major Pharn	https://www.nasdaq.com/symbol/aciu
23	ACIA	Acacia Communications, Inc.	39.23	1575066493	n/a	2016	Technology	Semiconduc	https://www.nasdaq.com/symbol/acia
24	ACTG	Acacia Research Corporation	3.9	199866739.8	n/a	n/a	Miscellaneo	Multi-Sector	https://www.nasdaq.com/symbol/actg
25	ACHC	Acadia Healthcare Company, Inc.	38.55	3401459982	n/a	n/a	Health Care	Medical Spe	https://www.nasdaq.com/symbol/achc
26	ACAD	ACADIA Pharmaceuticals Inc.	14.29	1783763097	n/a	2004	Health Care	Major Pharn	https://www.nasdaq.com/symbol/acad
27	ACST	Acasti Pharma, Inc.	0.468	16928408.48	n/a	n/a	Health Care	Major Pharn	https://www.nasdaq.com/symbol/acst
28	AXDX	Accelerate Diagnostics, Inc.	22.4	1209209098	n/a	n/a	Capital Goo	Biotechnolo	https://www.nasdaq.com/symbol/axdx
29	ACCP	Accelerated Pharma, Inc.	n/a	0	n/a	n/a	n/a	n/a	https://www.nasdaq.com/symbol/accp
30	XLRN	Acceleron Pharma Inc.	47.62	2179820977	n/a	2013	Health Care	Biotechnolo	https://www.nasdaq.com/symbol/xlrn

Fig.1 CSV file with first 30 rows

Stock Prices obtained using Yahoo! Finance API. This dataset consists of the Open, Close, High and Low values for each day. Yahoo! Finance is a media property that is

part of the Yahoo! network. It provides financial news, data and commentary including stock quotes, press releases, financial reports, and original content. It also offers some online tools for personal finance management. It includes premium data and charting, advanced portfolio analytics, research reports and investment ideas, and company profiles. The next step is to create the Pandas DataFrame of the introduced symbol stock market values from last year from now. The information is retrieved from Yahoo! Finance using yahoo finance fix.

The main functionalities of the yfinance library includes retrieving both, company financial information (e.g. financial ratios), as well as historical market data. To be able to use the library, we will need to install it. Then to use the package, we need to import it in our Python script. We pass as an argument of Ticker i.e the ticker of the company (symbol). We can also download historical prices for more than one stock simultaneously.

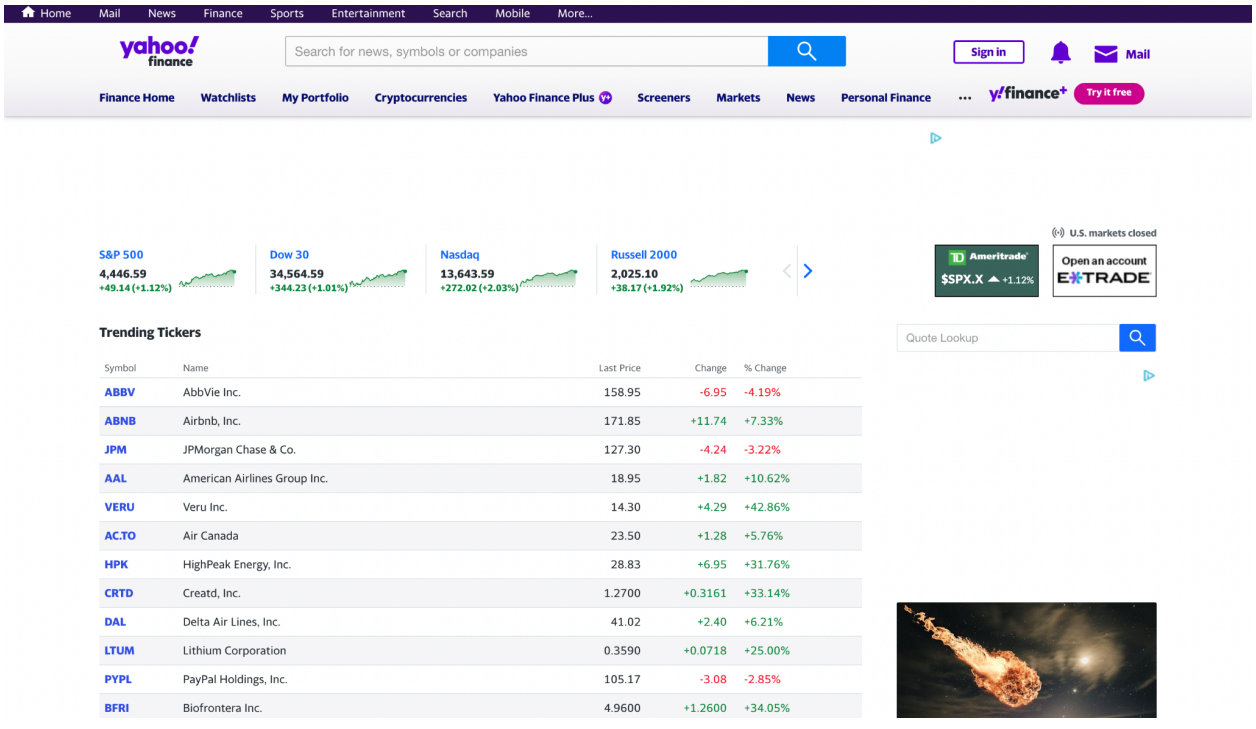


Fig.2 Yahoo Finance home with companies listed

The steps involved in the Yahoo Finance platform,

1. Sign in to Yahoo Finance [4]
2. Click My Portfolio.
3. On the right, click Create Portfolio.
4. Enter a name and select a currency for the new list.
5. Click Submit to get the data of the companies

5.2 Solving the Problem

5.2.1 Algorithm Design

The influence of social media like Twitter on the stock market is undeniable. In many ways, it's a double-edged sword case where while social media can promote stocks and make them go viral, it can also lead to their downfall when investors start shorting them or betting against them. Once a company goes public, its shares start trading on a stock exchange, its share price is determined by supply and demand in the market. If there is a high demand for its shares due to favorable factors, the price will increase and henceforth it is recommended to the investor. We are using machine learning algorithms in order to gauge sentiment from tweets to predict whether the stock increases and provides profit or not and whether a stock will move a certain way. For example, if there are more positive words than negative words in a tweet, our algorithm labels it as a higher score and predicts the stock price to move upwards.The recommendations explicitly mention whether to ‘Buy’ or to ‘Sell’ a particular stock, which helps users to take correct course of action and the investors in decision making.

The company symbol is captured using text mining techniques to fetch the actual stock prices from the website. Pre-processing is then followed by sentiment analysis. Lexicon-based approach is used for sentiment analysis. A dictionary has been created with two lists of Positive and Negative words. This dictionary focuses on the words related to trading.

Upon matching with the positive word, the positive counter moves and negative counter moves, upon matching with the negative word. Positive and negative counters are compared, and the greatest is labeled to the tweet as the Sentiment. In case of the tie, the tweet is treated as neutral. Time series data with the number of positive, negative and neutral tweets along with polarities is generated.

Polarity of tweet calculated using formula, Polarity = (P-N) / (P+N), where P= No. of Positive tweets, N= No. of Negative tweets collected from the twitter. Using the polarity we calculate the stock results.

Stock Price Prediction using machine learning helps to discover the future value of company stock price and other financial assets traded on an exchange. The entire idea of predicting stock prices is to gain significant profits. The main factor of the system is achieved by implementing a machine learning algorithm on this data is linear regression. The linear regression model returns an equation that determines the relationship between the independent variables and the dependent variable. Tweepy is an open source Python package that provides a very convenient and efficient way to access the Twitter API with Python. Tweepy includes a set of classes and methods that represent Twitter's models and API endpoints, and it transparently handles various implementation details, such as: Data encoding and decoding. We have used a linear

regression approach to predict the polarity of the tweets. This data analytic approach performs better than the support vector machine and naïve bayes approach. Hence will use LinearRegression because it has a nice accuracy score overall. We also apply the cross-validation algorithm to establish the X and y values for training (70%) and testing(30%).

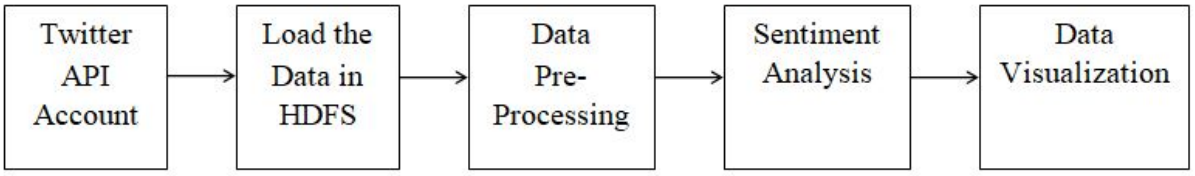


Fig.3 Flow diagram for Twitter data

5.2.3 Languages Used

In this project, we will be using PySpark to perform data analysis on a large Twitter dataset. **Spark** is fast and a general engine that is used for large-scale data processing. It's way faster than the previous approaches like MapReduce because it runs on memory. The data is stored in the RAM of the servers so the data can be accessed quickly and in turn increases the speed of the analytics. Spark has its own machine learning module called MLlib.

PySpark is the interface for Apache Spark in Python. It helps to work with RDDs(Resilient Distributed Datasets) in a python context.PySpark is the most preferred because it can create more scalable analyses and pipelines. The packages that are used as a part of implementation are:Python, PySpark, NumPy, Tweepy, Pandas, Scikit-learn, fix_yahoo_finance, Textblob.NumPy is an open source python library that is used in various field of science and engineering.It’s library contains multidimensional

array and matrix data structures.NumPy is used to perform a various mathematical operations on arrays. Tweepy is a Python package that gives a very convenient way to access the Twitter API. Tweepy consists of a set of classes and methods that represent Twitter's models and API endpoints. It is used for simple automation and creating twitter bots. Pandas is a python library that is used to analyze the data. It provides various data structures and operations for time series and also manipulates other numerical data. Scikit learn is part of the Python machine learning toolkit that is widely used for classification, predictive analytics, and many other machine learning tasks.TextBlob is a library for processing textual data. It is the best opt for a sentiment analyzer.

5.2.4 Tools Used

In this project the primarily used tools are Jupyter Notebook and Databricks. **Jupyter Notebooks** is a web application and an interactive computing platform that allows people to create and share documents called notebooks that contain live codes,texts,visualizations,graphs and plots.Both the code and results and can been together.Jupyter Notebooks is the best when it comes to security because the data that are sensitive and should be protected, are not stored in the local machines. **Databricks** is a cloud based data engineering tool that processes and transforms large quantities of data and further explores the data through machine learning models.It processes both structured and unstructured data and reduces the batch processing time.Databricks uses more data science to support decision-making.

5.3 Generating output

Once the predictive models are built using appropriate machine learning techniques, we should validate the performance of the models by using test data.The data that is retrieved is basically split into train and test data.The training will be done on the train data and the validation will be done using the test data to evaluate the accuracy of the models. The performance of the models is compared by using few performance metrics like accuracy and precision. To make recommendations we will check if the prediction is favorable compared to the last non-predicted value and the global polarity is checked and appropriate recommendations are made.

5.4 Testing against Hypotheses

The prediction models which we use are for both the linear regression and the Naive Bayesian classifier. We will be using the same dataset to compare the performance. The performance results of the two classifier models would show whether there is any difference or not.

6 Implementation

6.1 Code

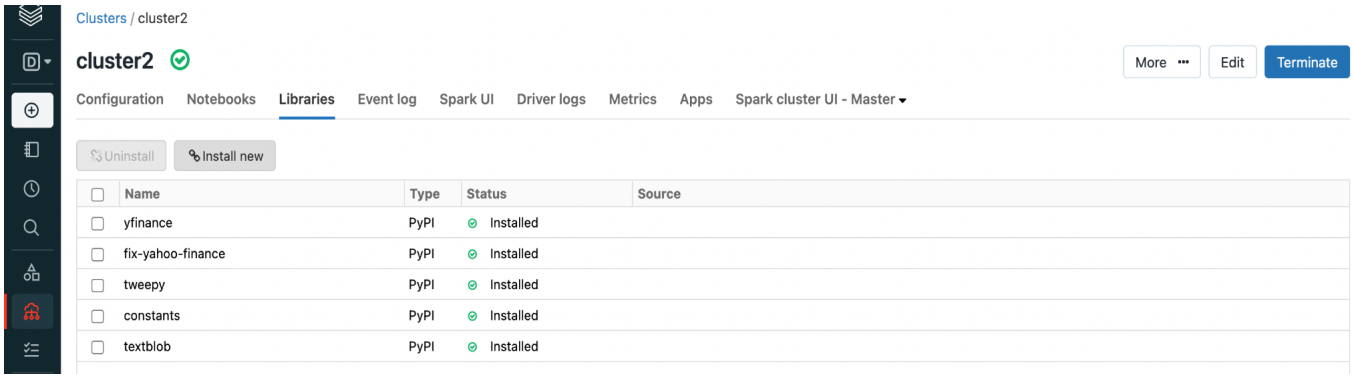


Fig.4 Installing libraries

Fig. 5. Importing required packages and libraries

The necessary packages and libraries are imported to be used throughout the program execution. Importing VectorAssembler from ML library. It is a feature transformer that merges multiple columns into a vector column. We are importing the VectorAssembler class and passing in a list of the feature column names. Matplotlib is imported for graph representation of the resultant data. Pandas is imported to convert the input as pandas dataframes. We are creating a predictive model using ML Regression model LinearRegression by importing the module from pyspark ML library. The dataset is split using the method train_test_split from sklearn model selection that splits arrays or matrices into random train and test subsets. TextBlob is a Lexicon-based sentiment analyzer used in this project. It has some predefined rules or word and weight dictionary, where it has some scores that help to calculate a sentence's polarity for the given input. Hence Lexicon-based sentiment analyzers are also called “Rule-based sentiment analyzers”.

Fig. 6. check_stock_symbol method to compare the stock symbols

The check_stock_symbol method has the input parameter company_list.csv which contains the list of companies with their symbols, Last sale, Market capital, Sector and Industry. This fetches the company's stock symbol and stores it in a data frame using df = pd.read_csv(companies_file, usecols=[0])

Fig. 7. get_stock_data method to retrieve the input data frame

The get_stock_data function module takes the input parameters as the company symbol, from_date, and to_date. The financial data of the particular company is downloaded from yfinance using the symbol, start which is the from_date i.e, past date from when the tweets have to be fetched. The end date will be the current date to fetch all the tweets until current. The data collected is stored in a dataframe with columns 'Open', 'High', 'Low', 'Close', and 'Volume'.

Fig. 8. Twitter Authorization details

To access Twitter data using tweepy API, create a twitter developer platform to get the API_key, Access token, API_secret key and Access token secret. Number of tweets to be fetched is given implicitly to the attribute num_of_tweets

Fig. 9. Retrieve tweet polarity

The function retrieving_tweets_polarity takes input parameter as symbol as check for polarity values.

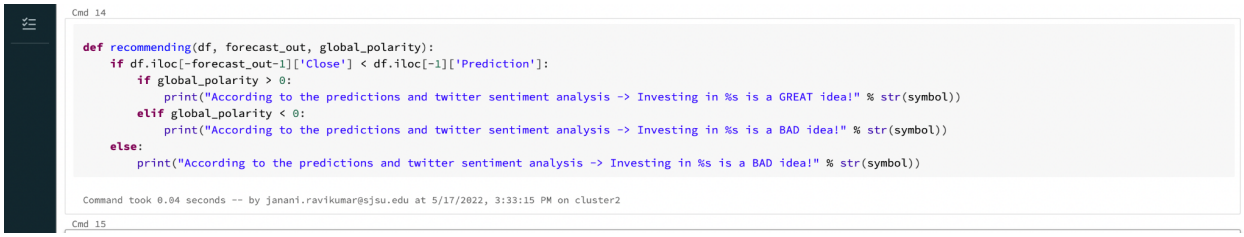


Fig.10. Function module to print recommendations for a particular company

Based on the polarity stock recommendation is suggested to the investor on whether to invest in a particular company's stock. If the polarity is greater than 0 it is considered to be positive and is recommended to the investor as a good idea to invest. Otherwise it is considered as a negative tweet.

Plotting existing and forecasted values

```
def forecast_plot(df):
    df['Close'].plot(color='black')
    df['Prediction'].plot(color='green')
    plt.legend(loc=4)
```

```
    plt.xlabel('Date')
    plt.ylabel('Price')
    plt.show()
```

Generating recommendation based on prediction & polarity

```
actual_date = dt.date.today()
past_date = actual_date - dt.timedelta(days=365 * 3)
actual_date = actual_date.strftime("%Y-%m-%d")
past_date = past_date.strftime("%Y-%m-%d")
symbol = 'ATVI'
print ("Retrieving Stock Data from introduced symbol...")
dataframe = get_stock_data(symbol, past_date, actual_date)
print ("Forecasting stock DataFrame...")
(dataframe, forecast_out) = stock_forecasting(dataframe)
print ("Plotting existing and forecasted values...")
forecast_plot(dataframe)
polarity = retrieving_tweets_polarity(symbol)
print ("Generating recommendation based on prediction & polarity...")
recommending(dataframe, forecast_out, polarity)
```

6.2 Design Document and Flowchart

Spark API is available in multiple programming models and languages such as Scala, Java, Python, and R. Spark has three different data structures available through its APIs: RDD, Dataframe, and Datasets. The corresponding machine learning library is Sparkml. We are using RMSE scores to validate the efficiency of our models and to

analyze which model works better for the used dataset. Some circumstances determine which of the API has to be used. The rows which have missing values such as price values are further processed. The data along with the sentiment scores is divided into train and test data and is fed to the model. This system is based on Pyspark programming that offers stock investment recommendations based on Machine Learning predictions from last 3 year's values of any market symbol and also based on Twitter sentiment analysis from retrieved tweets containing the symbol as text inside of it.

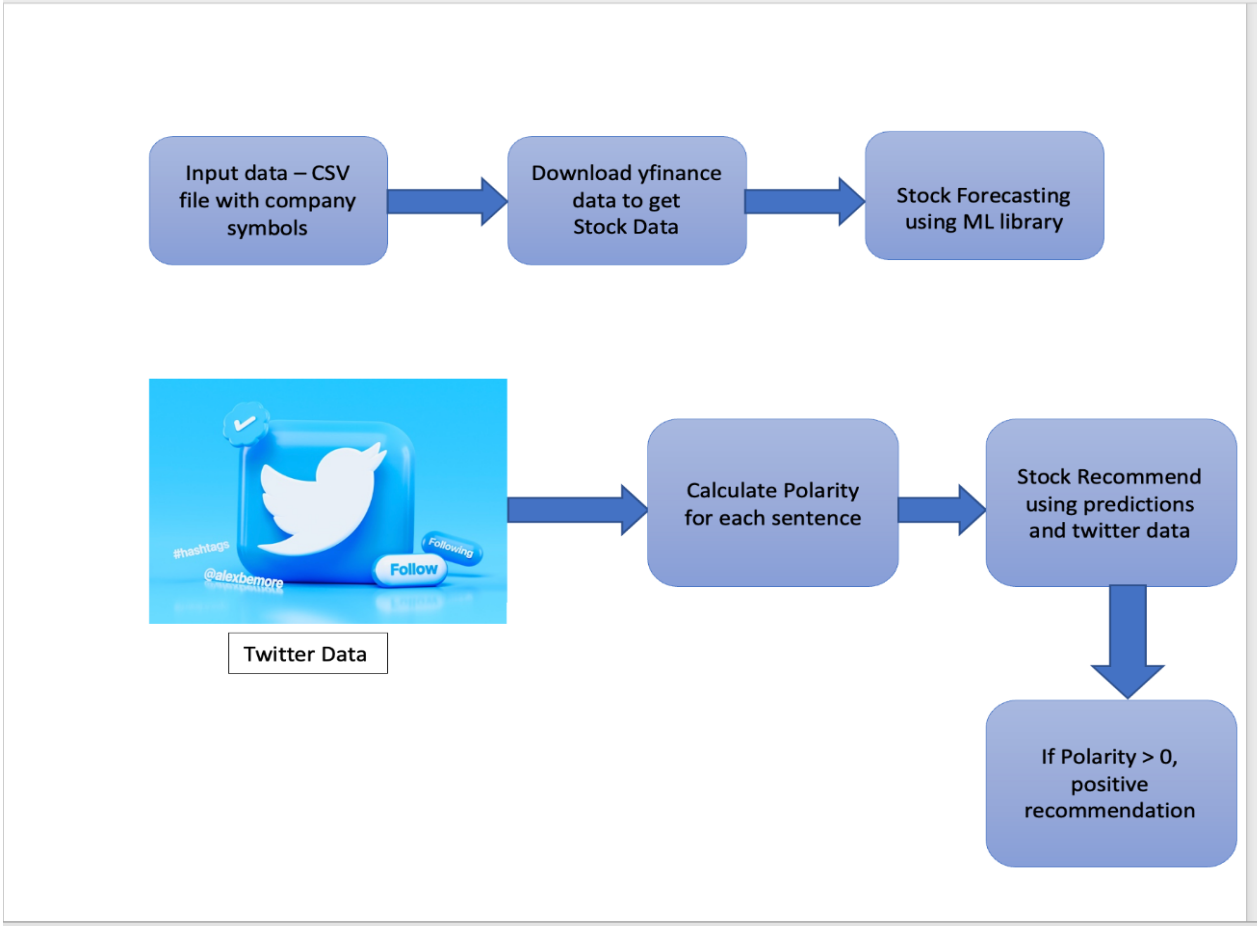


Fig.11. Flowchart for Stock recommendation using Twitter Sentiments

7 Data Analysis and Discussion

7.1 Output Generation

The data collected is stored in a dataframe with few named columns.The data is split into train and test, in order to train our models and objectively measure their performance**.** A complex formula based on the model that we have trained on the sentiments of the Tweets is used to predict the stock price. Using LinearRegression the prediction for stocks is done and once the stock prediction is done,we are using the tweets that are retrieved from twitter using the twitter development account that matches the symbol of the companies. The results from the prediction are used to make the recommendation for the stocks.We have checked whether the prediction is same as the last non predicted value.If it is same as the last non predicted value then the global polarity is checked.The value of global polarity is positive then that tells us that we can invest in that symbol is we can invest in that company. If the global polarity holds a negative value or zero then it's a bad decision to invest in those companies' stock.

7.2 Output Analysis

The dataframe was created to store the companies list csv file.The function check_stock_symbol will check if the company we request is present in the csv file .If the company is not found in the csv list it will throw an error(fig1).

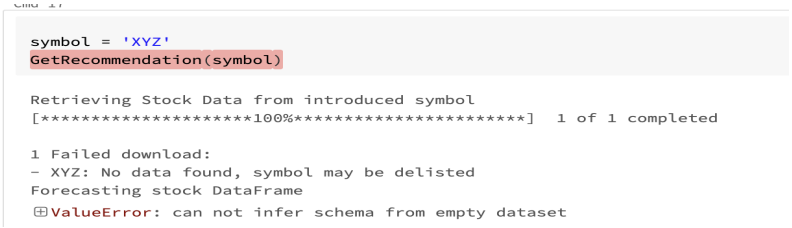


Fig1 Error when the company we request is not present in the Company csv list

The stock_forecasting function module is used to start modelling the data frame using Vector Assembler and used linear regression algorithm for predicting the values.Using LinearRegression algorithm the data is split into training and testing and the accuracy is found.We have used linear regression because it gives us most appropriate accuracy when compared with the other.The linear regression algorithm has been applied and then we have plotted the close and forecast value.The ‘close’ is represented in black lines and ‘prediction’ values are represented as green in the plot diagram.

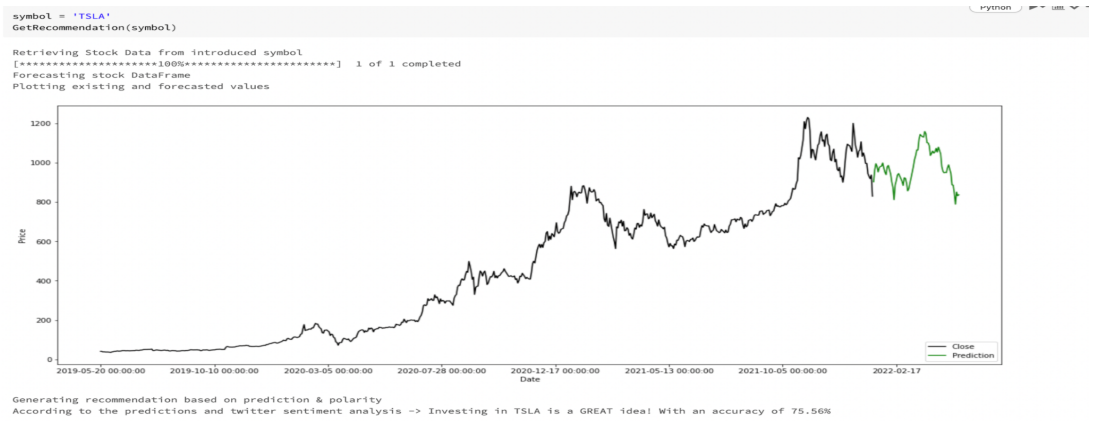


Fig 2 The accuracy is displayed and the recommendation is showed as ‘GREAT idea’ for ‘TSLA’

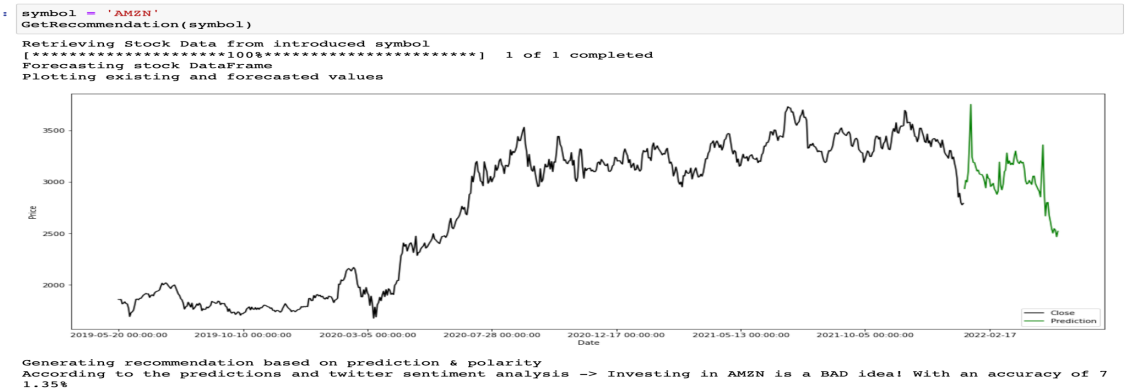


Fig 3 The accuracy is displayed and the recommendation is showed as ‘BAD idea’ and for ‘AMZN’

We followed the same steps by calling the GetRecommendation function and found the accuracy and prediction whether its bad or good for 5 other company stocks like ‘TSLA’(fig2),’AMD’,’AAPL’,’AMZN’,’MSFT’(fig3).Accuracy for each company varies by 1% more or less when the data is trained and tested again.The accuracy that we calculated for ‘AAPL’ is 87.40%,’AMZN’ is 71.75%,’TSLA’ is 74.96%,’AMD’ is 61.30%, MSFT is 84.32%.The global polarity value is checked accordingly with the textblob and if the prediction value is comparatively same to the last non predicted value and if the value of global polarity is less than zero the result will be displayed as “bad idea” which means we should not invest in the company.If the global polarity value is greater than 1 or positive then the result will be displayed as “good idea” which means we should invest in the stocks of that company. The recommendations are also made as ‘bad idea’,’badidea’,’goodidea’,’good idea’,’bad idea’ for ‘AAPL’,’AMZN’,’TSLA’,’AMD’,’MSFT’.

7.3 Compare Output against Hypothesis

In chapter 4, we proposed a single/multiple hypothesis for the Twitter Sentimental Analysis for Stock Recommendation. The following will compare the output with the hypothesis one by one.

Firstly, the primary goal of this project is to identify attitudes in text data about a subject of stocks by the method Sentimental Analysis. Initially we will be checking the symbol of the stock from the company list, then we will be doing the stock_forecasting in which we will be training and testing the data.

Secondly, we proposed that our project will demonstrate retrieving_tweets by polarity. It is measured by range from 1 to -1.Values closer to 1 indicate more positivity, while values closer to -1 indicate more negativity.

Finally, the stock recommendation is done with the help of global polarity. If it is greater than zero it is marked as “Great Idea” and less than zero is marked as “Bad Idea”.

As we can observe from the above output analysis that we are getting the output as expected so the hypothesis which we have got the hypothesis are the expected as earlier.

7.4 Abnormal Case Explanation

The abnormal cases that we encountered were the stocks which are not that highly successful and not much tweeted such as AutoWeb, Entera Bio Ltd., do not come into the picture unlike the most famous and highly tweeted companies like Tesla, Amazon,etc. So sometimes this leads us to difficulty in predicting the stocks because when we are discussing the stocks that means it must be able to predict all the stocks which are present. This leads to the discussion below of what we feel needs to be explicitly emphasized when talking about stock prediction.

7.5 Discussion

We feel that Stock Market Predictions is a common and well-known problem of interest. So we basically retrieved the data from the famous American social networking site that is Twitter with the help of a “Twitter Developer Account” which we needed to request for the access to the developer account. Once they approve it we can access the tweets,

now we will check the stock company symbol whether it is in our company list or if not present it throws an error. Once that is done then we will fetch the stock data from the yahoo finance we need to import all the required packages for this after fetching it will convert from dataframe to the desired columns. Now this data is set to test and train in the percentages of 70% and 30%. As the latest technologies are emerging at a fast pace we have used one of the machine learning algorithms that is the “Linear Regression” which is used for the prediction. Once the prediction part is done with the help of polarity we will calculate the stocks and we will display the output whether it is good stock or bad stock. Apart from this we found that this model is working well for the highly tweeted companies but not the companies which are less tweeted so which need to to be taken care. Finally this project will help the customers who are new into the stock investments and also to the stock prediction experts who can verify with this such that the stocks can be invested very wisely.

8 Conclusions and Recommendations

8.1 Summary and Conclusions

Machine Learning and pyspark have been used for better stock market analysis. The stock market is often uncertain. With this, we can help the investors from facing significant financial losses.We have imported sklearn preprocessing module that preprocess the input data and removed the null values which gave us better accuracy, all more than 60 percent. We have retrieved data from the last 500 tweets so when the data retrieved is more the accuracy increased.Apart from the linear regression methods, many methods can be implemented in finding better results.We came to know that

better results can be achieved by using the Neural networks.In the Neural networks we can increase the number of nodes to achieve better accuracy. We also found that the Feed-Forward neural network will give a better accuracy prediction for the opening price of the stock.

8.2 Recommendations for future studies

As a future work our team would recommend extracting more historical tweets by purchasing a premium twitter developer subscription.We would really wish to purchase a premium twitter developer subscription that would have helped us to use all-time historical prices together with tweet attributes which would have given us higher accuracy. When we get the premium developer subscription we can retrieve an enormous amount of training data that is 10 times as much as we have used now , which would make a huge difference. We also found that it is interesting to add ARIMA which is a model that uses time series data to understand the dataset and predict stock prices and can also handle the linear part of the data set and use the neural network to handle the non-linear part.

9 Bibliography

[1] Tushar Rao and Saket Srivastava. 2012. Analyzing Stock Market Movements Using Twitter Sentiment Analysis. In Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012) (ASONAM '12). IEEE Computer Society, USA, 119–123.

[2] Man Li, Chi Yang, Jin Zhang, Deepak Puthal, Yun Luo, and Jianxin Li. 2018. Stock market analysis using social networks. In Proceedings of the Australasian Computer

Science Week Multiconference (ACSW '18). Association for Computing Machinery, New York, NY, USA, Article 19, 1–10.

[3] Bhakti G. Deshmukh, Premkumar S. Jain, M. S. Patwardhan, and Viraj Kulkarni. 2016. Spin-offs in Indian Stock Market owing to Twitter Sentiments, Commodity Prices, and Analyst Recommendations. In Proceedings of the International Conference on Advances in Information Communication Technology & Computing (AICTC '16). Association for Computing Machinery, New York, NY, USA, Article 77, 1–7.

[4] Zheni Mincheva, Nikola Vasilev, Ventsislav Nikolov, Anatoliy Antonov.

Extracting Structured Data from Text in Natural Language. August 31, 2021.

10 Appendices

Stock Recommendation System Term Project

https://drive.google.com/drive/folders/1YwrGOsy8J17U-Js0BEc1uPUZJ_ZsrfGn?usp=sharing

Program source code

https://drive.google.com/drive/folders/1iuazyfcgfgVwoy2xU7Xf586ORD9fJb96?usp=sharing

Documentation

https://drive.google.com/drive/folders/1lipbM8ofytJ6pvWAunse1yssDfZhiOgF?usp=sharing

Input/Output listing

https://drive.google.com/file/d/16Y7rxC9rSGaQm76wpX9lONfgBlaiWpPU/view?usp=sharing

README File

https://drive.google.com/file/d/1iQRxyxQLR6_gkVkVE47dkqmDrMWbHeJY/view?usp=sharing

References

1. https://www.emerald.com/insight/content/doi/10.1108/JKM-11-2017-0517/full/html?casa_token=6ZWx-j-2-RoAAAAA:9zEQUvLe7XiEderMrYSgR8LYgvG3oW9HA-wRjvl5XnWqNUTHE7SpglWXRBjdQ5GfEj6tSVe3LRusmo0_0Liiq7uAQXyneQTKkqtbEG6tLWHn3thBchM#ref069
2. https://finance.yahoo.com/