Twitter Sentiment Analysis

Likhitha Vanga
Computer Science
Portland State University
Portland, Oregon, USA
vanga@pdx.edu

Shashank Shekhar
Computer Science
Portland State University
Portland, Oregon, USA
sshekhar@pdx.edu

ABSTRACT

In the modern era, vast amounts of data are transmitted online through different social media channels. Data contains information about virtually every topic. Twitter is a microblogging platform where 100 million users login daily and more than 500 million tweets are sent per day. Each tweet has a maximum of 140 characters hence more than 70 billion characters generated per day only on twitter. Though each tweet may not be valuable, we can extract important data that provide valuable insight into public mood and sentiment score on certain topics.

In Data science, it is an interesting topic to collect vast amounts of data from social websites, process it, and run data analysis on retrieved data to extract relevant information. In the case of twitter, we have collected large amounts of live data, processed tweets, and generated a feature matrix to apply machine learning algorithms.

KEYWORDS
[1]-Data from Wikipedia.
[2]-Elon Musk

ACM Reference format:
Likhitha Vanga and Shashank Shekhar. 2019.Twitter sentiment Analysis. In *Proceedings of CS 410/510: Explorations of Data Science. Portland, OR, USA, 10 pages.*

1 Introduction

One of the main reasons for choosing the stock market is the vast amount of data available and secondly, we would like to see how we could use machine learning to predict the fluctuation of the entire market or a particular stock based on tweets made by influential personalities. Not too long ago during the rise of cryptocurrencies Charlie Lee creator of Litecoin tweeted to all Litecoin holders that these were good times and one should be prepared for and to expect the coin value to fall to as low as $20, this tweet caused a stir in the market causing the coin to drop hundreds of dollars in a matter of no time, however when he tweeted about some of the new security features and robustness of the software we saw a spike in the coin price. Another example is President Trump tweeting on increasing import duty on Chinese products, we saw a frenzy in the markets. In today's world where information is readily available and at your fingertips, we think and believe that the paradigm of how stocks rise and fall based on information has changed in a way where we can use machine learning to use available data to decide if the outcome will cause a positive or negative impact.

2 Twitter Sentiment Analysis!

Sentiment Analysis is the automated process of analyzing text data and sorting it into sentiments positive, negative, or neutral. Performing Sentiment Analysis on data from Twitter using machine learning can help companies understand how people are talking about their brand.

With more than 321 million active users, sending a daily average of 500 million tweets, Twitter allows businesses to reach a broad audience and connect with customers without intermediaries. On the downside, it's harder for brands to quickly detect negative content, and if it goes viral you might end up with an unexpected PR crisis on your hands. This is one of the reasons why social listening — monitoring conversation and feedback in social media — has become a crucial process in social media marketing.

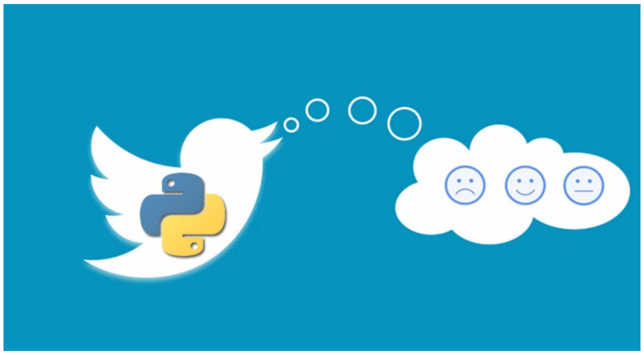


Figure 1: Python + Tweets = Sentiment Analysis!

Monitoring Twitter allows companies to understand their audience, keep on top of what's being said about their brand and their competitors, and discover new trends in the industry. Are users talking positively or negatively about a product? Well, that's exactly what sentiment analysis determines.

Nearly 80% of the world's digital data is unstructured, and data obtained from social media sources is no exception to that. Since the information is not organized in any predefined way, it's difficult to sort and analyze. Fortunately, thanks to the developments in Machine Learning and NLP, it is now possible to create models that learn from examples and can be used to process and organize text data.

Twitter sentiment analysis systems allow you to sort large sets of tweets and detect the polarity of each statement automatically. And the best part, it's fast and simple, saving teams valuable hours and allowing them to focus on tasks where they can make a bigger impact.

2.1 Advantages of Twitter Sentiment Analysis

Scalability: If we need to analyze hundreds of tweets mentioning a brand manually, it would take hours and hours of manual processing and would end up being inconsistent and impossible to scale. By performing Twitter sentiment analysis, we can automate this task and obtain cost-effective results in a very short time.

Real-Time Analysis: Twitter sentiment analysis is critical to notice sudden shifts in customer moods, detect if critics and complaints are increasing, and take action before the problem escalates. You can be monitoring your brand in real-time and get valuable insights that allow you to make changes or improvements when needed.

Consistent Criteria: Analyzing sentiment in a text is a subjective task. When done manually, the same tweet may be perceived differently by two members of the same team, and the results will probably be biased. By training a machine learning model to perform sentiment analysis on Twitter, you can set the parameters to analyze all your data and obtain more consistent and accurate results.

3 Implementation and Approach

Performing sentiment analysis on Twitter data involves several steps:

1. Gather Twitter Data
2. Prepare Your Data
3. Creating a Sentiment Analysis Model
4. Visualizing Results

We fetched the tweets from Kaggle which are pre-processed twitter data and then perform sentiment analysis on the same. Similarly, get the stock prices for the respective company for the dates we fetched the tweets. We considered the percentage change in stock values.

Preprocessing of the data:

For Twitter Data:

- Delete any blank tweets from CSV file
- Remove the special characters and keep plain text alone.

For Stock prices:

- Add the previous day's stock price in case the stock price value is blank such as on Saturday, Sunday, and national holidays.
- So, for Saturday and Sunday, the stock of a given company would be the closing price on Friday.

Merge the data:

- For the text blob sentiment analyzer, we merge the polarity, confidence with percent off (the price change for the previous day) which are the results from twitter analysis and analyzing stock prices.

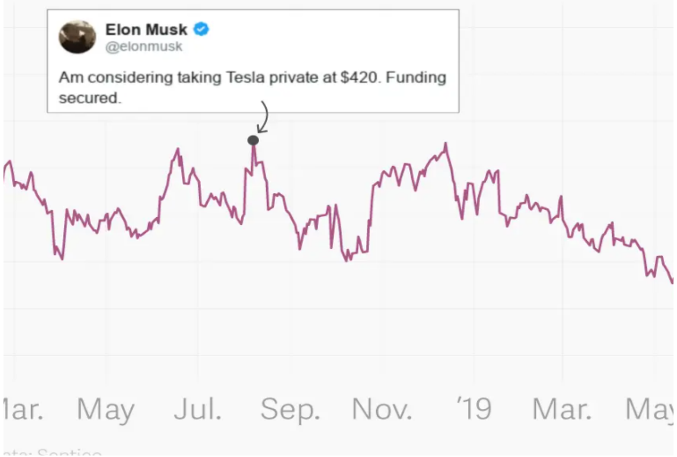
Polarity: Defines the positivity or negativity of the text; it returns a float value in the range of "-1.0 to 1.0", where '0.0' indicates neutral, '+1' indicates a very positive sentiment and '-1' represents a very negative sentiment.

Stock Prices for Tesla:

Tesla, Inc. is an American electric vehicle and clean energy company based in Palo Alto, California. The company specializes in electric vehicle manufacturing, battery energy storage from home to grid scale and, through its acquisition of SolarCity, solar panel and solar roof tile manufacturing.

Outside of the courts, Tesla has been the subject of other public controversies, ranging from securities fraud allegations to product delays to workers safety complaints [1].

Also, Elon musk tweets have a pretty large effect on stock prices of Tesla.

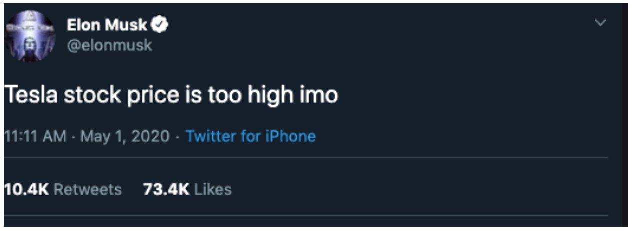


Elon Musk
@elonmusk

Am considering taking Tesla private at $420. Funding secured.

Figure 2: Elon Musk Tweet and stock price of tesla on respective day

The main reason to choose tesla stock prices is because we analyzed how relative the stock prices are to his[2] tweets.



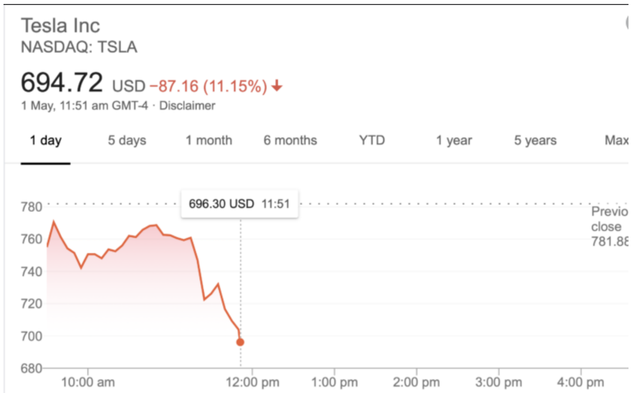
Elon Musk
@elonmusk

Tesla stock price is too high imo

11:11 AM · May 1, 2020 · Twitter for iPhone

10.4K Retweets 73.4K Likes

Figure 3: Elon Musk Tweet on May 1



Tesla Inc
NASDAQ: TSLA

694.72 USD −87.16 (11.15%)↓

1 May, 11:51 am GMT-4 · Disclaimer

Figure 4: Fall in stock prices within minutes after the tweet.

3.1 Implementation Details

- We did twitter sentiment analysis using NLTK Vader and text blob.
- After preprocessing data, we are doing sentiment analysis of tweets.
- We calculated the percent change for price.
- Partition data into train and test data set. Train the model with the past 4 years of data and predict the price change for next year.
- To predict next year's data, we used multiple learners: Linear regressor, MLP regressor, Random forest regressor.

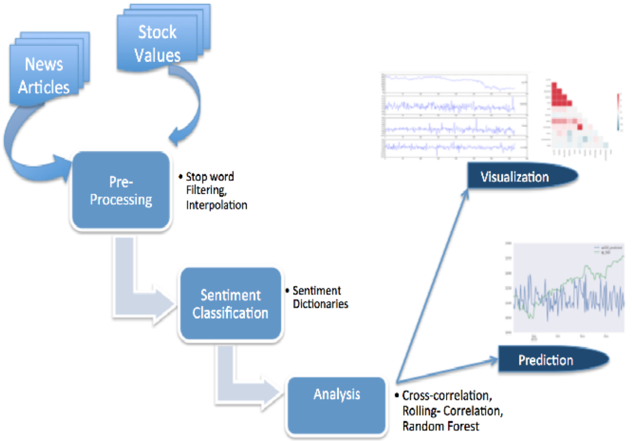


Figure 6: Abstract model of the project

3.2 Algorithms

Linear Regression Algorithm: Linear regression is perhaps one of the most well-known and well understood algorithms in statistics and machine learning. It is based on supervised learning.

It performs a regression task. Regression models a target prediction value based on independent variables. ... When training the model – it fits the best line to predict the value of y for a given value of x.

Linear regression is been studied at great length, and there is a lot of literature on how your data must be structured to make best use of the model. As such, there is a lot of sophistication when talking about these requirements and expectations which can be intimidating. In practice, we can use these rules more as rules of thumb when using Ordinary Least Squares Regression, the most common implementation of linear regression.

It is a python library in sklearn:

from sklearn.linear_model import LinearRegression

To prepare data for linear regression one can try different heuristics like Linear Assumption, Remove Noise from data, remove collinearity from data, Gaussian Distributions, Rescale Inputs, etc. In our project, we have removed noise from tweeter data by deleting blank tweets and special characters from tweets. For the stock price, we are updating the holiday stock prices with the previous day's closing price. And then we are calculating the percentage change in price and applying linear regression learners to train data and predict the stock price for next year.

MLP Regressor Algorithm: Multilayer Neural Networks has one input layer and one or more hidden layers and an output layer. It has an activation function and a cost function for the layers. In this project, we experimented with different numbers of hidden layers and regression functions such as ReLU, logistic regression. The reason for using a multilayer neural network is to try for a more powerful model than linear regression.

Random Forest Algorithm: Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean prediction of the individual trees.

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction.

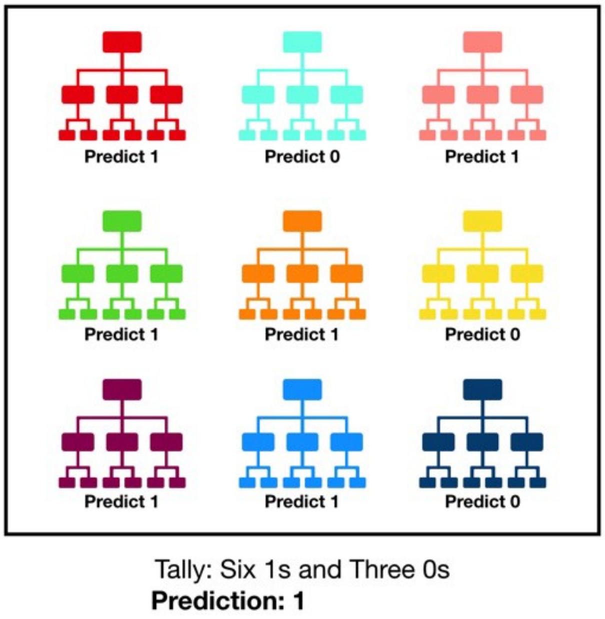


Figure 7: Random Forest representation

It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.An important thing to remember is While the algorithm itself via feature randomness tries to engineer these low correlations for us, the features we select and the hyper-parameters we choose will impact the ultimate correlations as well.

4 Conclusion

Twitter Sentiment Analysis allows a deeper understanding of how your customers feel. It adds an extra layer to the traditional metrics used to analyze the performance of brands on social media and provides businesses with powerful opportunities.

There are a lot of techniques and tricks used by people to improve their results. Different classification tools and analytical methods like SOFNN or Self Organizing Fuzzy Neural Network model as described in Bollen at el paper can be explored to further improve this project.

We are using several measures to predict the stock prices using tweets such as polarity, sentiment confidence but there are things like company product launch dates, press conferences, financial reviews, and commercials that also affect the nature of the stock value. Indeed, if predicting stocks were that simple then life would be different.

ACKNOWLEDGMENTS

We would like to thank Professor Kristin Tufte, the instructor for this course, for providing advice, ideas, resources and constant support for this project. Also, we would like to thank Jevin West and Carl T. Bergstrom from Calling B.S who truly inspired us from their lectures.

REFERENCES

[1] https://monkeylearn.com/blog/sentiment-analysis-of-twitter/#Why-is-Twitter-Sentiment-Analysis-important

[2] https://www.sciencedirect.com/science/article/pii/S2405918817300247 Tahir M.Nisar ManYeung