

FLUCTUATIONS IN STOCK PRICES BASED ON TWEETS USING SENTIMENT ANALYSIS

*Anand (anand6@buffalo.edu), Rahul Patil (rpatil@buffalo.edu)
Supreeth Konanur Suresh (ssuresh7@buffalo.edu)*

OVERVIEW

This is 21st century, also known as a digital book, also known as an era of data overload. Data is generated and used in every single aspect of our life, be it as simple as cooking, or as complex as determining a satellite trajectory. In such a data-driven world, social media plays a major role in shaping this modern generation. One such powerful tool of social media is Twitter. Twitter is a micro-blogging and social networking service on which users post and interact with messages known as “Tweets”. Tweets have in the past, and continue to influence policies, and overturn markets. In this project, we wish to understand the nature of these tweets and predict, whether such a tweet, when coming from a global icon, can influence the stock prices of their organization.

WHAT IS SENTIMENT ANALYSIS? WHY DO WE NEED IT?

Sentiment Analysis is the interpretation and classification of emotions (positive, negative and neutral) within text data using text analysis techniques. Sentiment Analysis allows businesses to identify customer sentiment towards products, brands or services in online conversations. Sentiment Analysis are of various types viz. Rule-based sentiment analysis, Automatic and Hybrid sentiment-analysis. In this project, we use rule-based sentiment analysis that is a set of human-crafted rules to help identify subjectivity, polarity, or the subject of an opinion. Rule-based systems are very naive since they don't consider how words are combined in a sequence.

Stock analysis is among the trending jobs of the decade. In the long run, valuations may drive stock prices, but in the short term it is market sentiment that moves prices. This can create investment opportunities for long term investors to find attractive entry points, and for active traders to both enter and exit positions.

WHAT ARE THE PACKAGES USED?

Python has various packages built around the concept of sentiment analysis, each using its own method of calculating the polarity and scores of the given tweet. The result of these tweets is usually in the form of Positive, Negative, Compound and Neutral, each in the range of 0 to 1.

In this project, we use several packages, each of which is used as stated below:

SQLITE3	Extract data from the given dataset file, normalize and load in the form of Tables of a SQLite Database
PANDAS	Used for Data Manipulation in Python
REGEX	Used to remove noise from the given tweet, such as hashtags, etc.
NTLK	Used for data preprocessing, by converting given tweet into a list of strings and then remove morphological affixes from the word, leaving only the stem words.
VADER SENTIMENT	Acronym for Valence Aware Dictionary and sEntiment Reasoner, is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media.
MATPLOTLIB	It is a plotting library for python programming. It provides an object-oriented API for embedding plots into applications.
PLOTLY	Plotly is python's graphing library that makes interactive, publication-quality graphs.

WHICH ARE THE DATASETS USED ?

For Sentiment Analysis, we used the following datasets

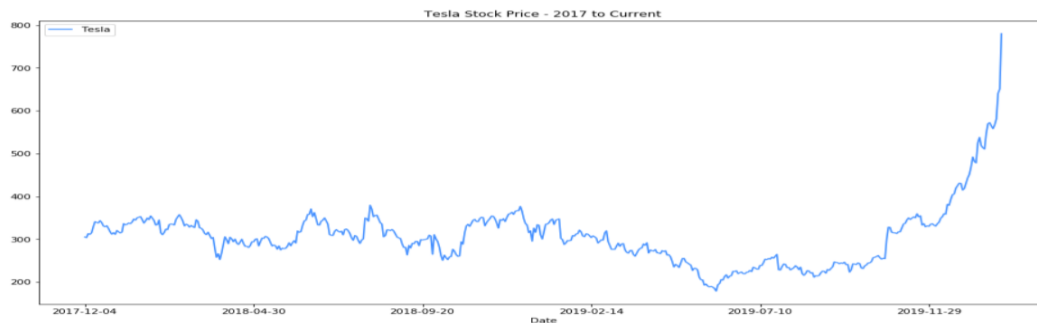
1. Elon Musk's tweets in the past two years i.e. 2018-2020 <https://data.world/barbaramaseda/elon-musk-tweets/workspace/file?filename=user-tweets.jsonl>
2. Tesla Stock data <https://www.kaggle.com/timoboz/tesla-stock-data-from-2010-to-2020>

WHAT IS THE CODE FLOW?

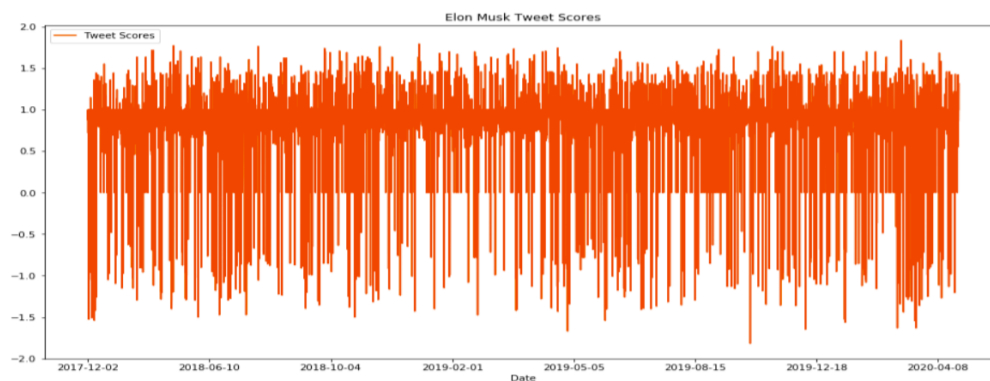
The code-flow of the project is as follows:

1. Using generic database methods to create table, drop table or execute SQL statements under 'SentimentAnalysis.db'.
2. Read tweet data from the dataset and extract user tweet and tweet date-time to a SQL table 'Tweet_data'.
3. Read Tesla stock data from the Tesla stock dataset and load it onto SQL table 'Tesla_stock'.
4. Preprocess tweets by removing the stop words using the package from Natural Language Toolkit called 'stopwords', remove retweet tags and hashtags using regular expressions, convert the given string into a list using nltk.tokenizer and then finally, remove affixes using nltk.stemmer, such that the text passed onto the sentiment analysis is a composition of only words, numbers and emoticons such as happy, sad, angry, etc.
5. Perform sentiment analysis on each of the tweets and return their polarity and scores in the form of a pandas dataframe. Further, manipulate the dataframe such that the negative compound values get appended to negative scores and positive compound values to positive scores.
6. Create a table 'Tweet_Scores' that has the sentiment scores of all the tweets.
7. Manipulate data from tables such that the tweet_scores and the stock prices are matched by date.
8. Plot graphs that depict the total positive scores, negative scores and neutral scores.
9. Plot graphs visualizing the rise and fall of stock prices over a period of time. And sentiment scores for dates.
10. Plot an interactive graph that shows the rise and fall of stock prices for any given data and the corresponding tweet_polarity and tweet score. This graph helps in predicting the effect of sentiment scores on the stock prices.

WHICH ARE THE GRAPHS AND WHAT ARE THE INFERENCES FROM IT ?

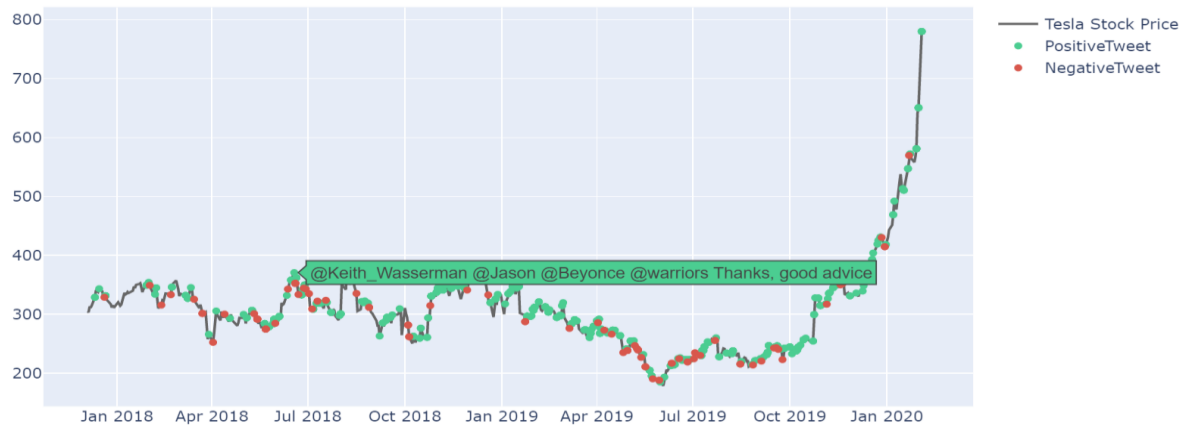


Plot of rise and fall of Twitter Stock Price over a period of 2 years



Plot for the positive and negative tweet scores centered at 0 over a period of 2 years

Tesla Stock Prices v/s Elon Musk Tweets Sentiments



Interactive plot showing the rise and fall of stocks with datapoints being the positive and negative tweets with green points for positive and red points for negative tweets

Tesla Stock Prices v/s Elon Musk Tweets Sentiments



Plot showing the fall in stock price and a negative tweet at that data point

From these plots, we can infer that the tweets from Elon Musk have a direct impact on the rise and fall of stock prices of Tesla. Therefore, Sentiment Analysis can be seen as a way ahead in predicting stock values.

WHAT DO WE CONCLUDE?

The insights from the project are as follows:

- Whenever there is a positive tweet, we see a spike in the stock price and vice versa.
- With further research, there is a possibility to precisely predict the fluctuations in stock prices based on the sentiment analysis scores.
- Sentiment Analysis using VADER is not 100% accurate. However, there is a huge scope for research in this domain using concepts of Machine Learning.