Analyzing tweets and news to predict stock price movement for Tesla

Group 7

Jacqueline Huang, Jing Li, Paridhi Agal

Table of Contents

Executive Summary

Stock market movement prediction is an important issue in the financial world, as it can influence investment strategies and contribute towards making various decisions related to stock exchange transactions. We are a team of analysts working for an investment firm to analyze the stock price movement for Tesla Inc. We want to help them gauge the market sentiment in order to make better investment decisions.

Historically news about the industry and companies has been shaping the investors' opinion. We decided that as these days social media is widely used by people it can represent the public opinion, mood and sentiment. Also we saw specifically Twitter has been a topic of interest to study public sentiment for a lot of researchers. So, we chose two different sources of information Twitter and Nasdaq news to gauge the market sentiment.

The present analysis has used two different textual representations, Word2vec for tweets and Doc2Vec for news articles for analyzing the public sentiments. We have applied sentiment analysis and supervised machine learning models to the text extracted from Twitter and Nasdaq news website. We also did principal component reduction on the Word2Vec embedding for tweets. We tried to predict the direction of stock market movements for the next day, will it rise or fall and predict the percent change bracket for the next day. The idea is that 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. Our prediction models based on tweets and news separately have an accuracy of 60%, based on combination of both news and tweets has an accuracy of 54%. We made time series models for stock price movement which had 90%+ accuracy and treated them as a benchmark. We have also shared how the analysis can be extended for other use cases.

Background, Context, and Domain Knowledge

"The most important thing to know to accurately forecast future stock prices is what mood investors will be in in the future," writes Housel.

Based on this we aim to gauge the market sentiment for an investment firm who wants to invest in Tesla.

Who is Tesla? What is the scenario?

Tesla is an American electric vehicle and clean energy company specializing in electric vehicle manufacturing and battery energy storage. Cybertruck is its star product, an all-electric, battery-powered vehicle. Despite one of its major selling points being the ultimate durability, its armor glass was smashed on Tesla news conference. However, such an embarrassing scenario didn't prevent the stock price from surging and stock price steadily doubled after the launch.

What is sentiment analysis? How sentiment impacts stock market?

Sentiment Analysis is the most common text classification tool that analyses an incoming message and tells whether the underlying sentiment is positive, negative or neutral.
Investors typically describe market sentiment as bearish or bullish. Emotion often drives the stock market, so market sentiment is not always synonymous with fundamental value which is about business performance.

What is Word2Vec? What is Doc2Vec?

The word2vec intends to give a numeric representation for each word and to capture relations between words. The Doc2Vec builds on the concept of word2vec but it intends to represent the concept of a document rather than just words, regardless of the document length. A document-unique feature vector is added to achieve this.
There are two common models - Distributed Bag of Words (DBOW) and Distributed Memory (DM). DBOW is the doc2vec model analogous to Skip-gram model in word2vec. The paragraph

vectors are obtained by training a neural network on the task of predicting a probability distribution of words given a randomly sampled word from the paragraph. Distributed Memory (DM) acts as a memory that remembers what is missing from the current context.

Traditional Approach to solve the problem

The term "investment firm" can refer to different kinds of companies. Investment banking is the closest to our role, which acts as intermediaries between investors and corporations. Generally investment firms make money by buying and selling various financial products. Every broker has their own strategy but networking is a main method to get information for trading.

The nonlinear and nonstationary data of the stock prices trend makes forecasting stock prices a challenging and difficult task in the financial market. Conventional time series models have been used to forecast stock prices, and many researchers are still devoted to the development and improvement of time-series forecasting models. The most well-known conventional time series forecasting approach is autoregressive integrated moving average (ARIMA)[1], which is employed when the time-series data is linear and there are no missing values [2]. Statistical methods, such as traditional time series models, usually address linear forecasting models and variables must obey statistical normal distribution [3]. Therefore, conventional time series methods are not suitable for forecasting stock prices, because stock price fluctuation is usually nonlinear and nonstationary. Previous studies have concluded that the aggregate public mood collected from twitter may well be correlated with Dow Jones Industrial Average Index(DJIA).

Analyses and Model Building

Exploratory Data Analysis

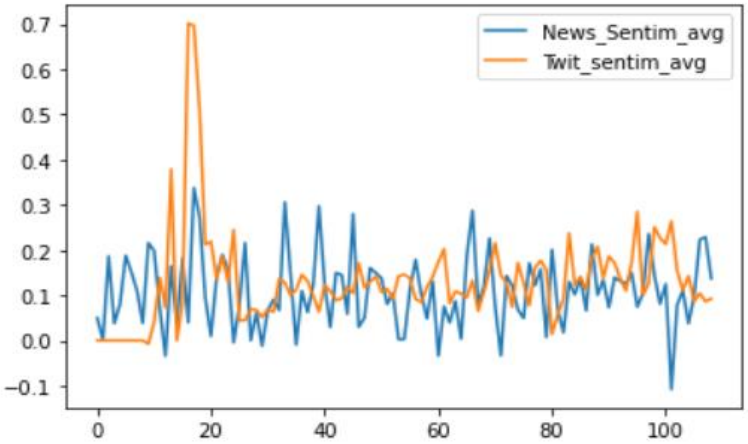
Summary of the data snapshot

1. Twitter data - 9675 tweets from 160 days

2. Nasdaq New - 806 news from 10000 pages for Oct 1, 2019 to Mar 10, 2020
3. Yahoo Finance Stock Prices- 111 day's stock price details

We saw that the volume of tweets had a big spike on October 24th 2019. Mainly because Tesla shares soared as much as 20% on that day after the company said it achieved profitability for the first time since the fourth quarter of 2018. It marked the stock's biggest intraday spike since May 2013.It defied consensus Wall Street analyst forecasts of a quarterly loss.Tesla also beat third-quarter earnings-per-share estimates and CEO Elon Musk said that the company's new Shanghai factory had already reached trial production and that the upcoming Model Y crossover would be released sooner than expected.

Also the tweets started to rise in November as we moved closer to the cybertruck launch.

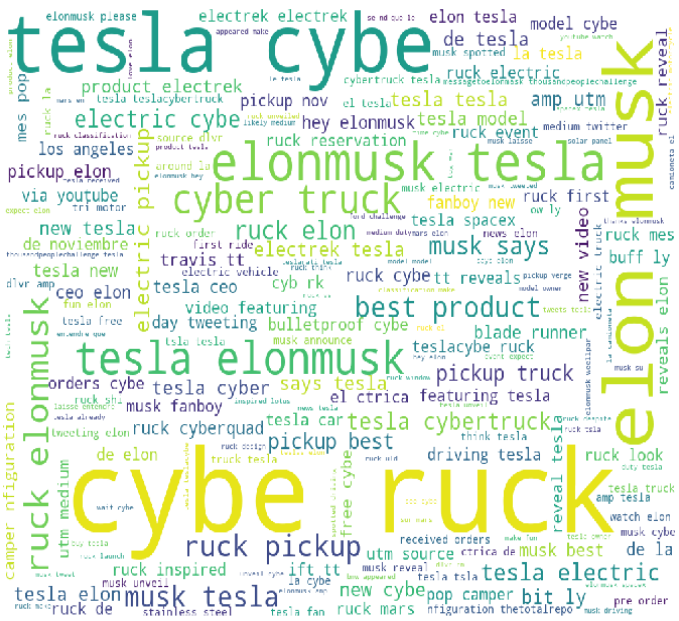


We plotted the trend of avg sentiment of news and tweets as seen in the graph below.

We cleaned the tweets and news and tried to see the distribution of overall sentiment for them:

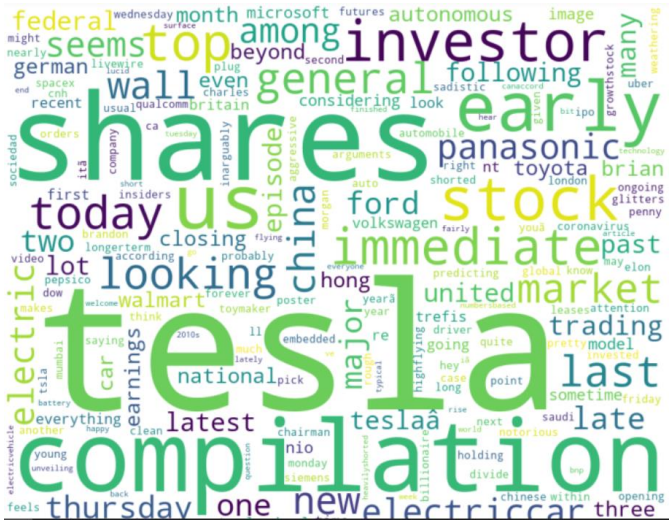
Sentiment/Data source	Tweets	News
Positive	47.06%	92.66%
Negative	10.79%	6.42%
Neutral	42.14%	0.92%

This showed that most of the tweets and news were positive and also the stock price has been rising from Oct 1st, 2019 to Feb 19th, 2020.



We did qualitative analysis for cleaned tweets and represented them in a form of a wordcloud (Fig 4.1). We searched for tweets with works tesla/elon musk/cyber truck. We saw words like best product, electric pickup, bullet proof, youtube had relatively high frequency alluding to the cybertruck launch.

<- Fig 4.1 Tweets' Word Cloud



We did qualitative analysis for cleaned news articles and represented them in a form of a word cloud(Fig 4.2). Though we searched only about Tesla we see the words like shares, stock, investor have high frequency in the news.

<- Fig 4.2- News' Word Cloud

Word chart from Word2Vec embedding for the tweets show which words are closer to each other. We used tweets data to train the Word2Vec model. There are 2792 words in the model vocabulary. We choose the top 500 frequent cleaned words and plot them using the first 2 principal components from principal component analyses (see appendix). The closer they are,

the more often they appear together in tweets. We see words like youtube and video, cyber truck and pick-up, unveil and announce are closer together as the Neural Network says they are similar which also supports the intuition. They indicate people discussed cybertruck launch.

Advanced Analyses on Tweets

Time Series

We did time series analysis to create a benchmark for predicting the stock price. We created two models. They are-

1. To predict the stock price by using a 17 day old stock price - We used linear regression which had an accuracy of 74%

$$Stock\ Price_{\ t} = \ intercept + \beta * Stock\ Price_{\ t-17} + \ Error$$

2. To predict the direction of movement of stock prices for the next day by using the stock prices till today - We used multinomial logistic regression it gave an accuracy of 92%

$$Stock\ Price\ Movement_t = \ intercept + \beta * \ Stock\ Price_{t-1} + \ Error$$

Sentiment Analysis

We used Textblob for natural language processing, which is a python library built on the shoulders of NLTK and Pattern. The sentiment function of textblob returns two properties, polarity, and subjectivity. Polarity is float which lies in the range of [-1,1] where 1 means positive statement and -1 means a negative statement. Subjective sentences generally refer to personal opinion, emotion or judgment whereas objective refers to factual information. Subjectivity is also a float which lies in the range of [0,1]

Word2Vec Training

By using Word2Vec, words or phrases from the vocabulary are mapped to vectors of real numbers, which can be used to apply mathematical models. We cleaned the twitter data and split it into sentences as the input of Word2Vec training model. We used the default size(the

number of output word dimensions) 100 to train the model and got the 100 dimension vector for every word in the model vocabulary. Then, every tweet is represented by the sum of word vectors appearing in the tweet. In this way, we converted tweets into vectors.

PCA Dimension Reduction

In order to predict stock price movement we grouped the twitter data by date. We used PCA to reduce dimension because 100 dimensions will be too much for 160 days data. PCA helps reduce the dimension by extracting linearly uncorrelated features from the 100 dimensions. Also, it helps simplify our final model. After applying PCA, we got the first 2 principal components which contains the information of tweets every day.

Prediction Model

Predicting the direction/movement in the stock price for the next day by using the information around the tweets till today.

Dependent variables: Stock price movement after one day

For the purpose of gaining more precise prediction, we divided stock price movement into 6 levels according to the relative percentage change:

$$relative\ percentage\ change = \frac{stock\ price_t - stock\ price_{t-1}}{stock\ price_{t-1}}$$

We defined movement by defining the label for it as per the below mentioned table -

Condition	Label- Movement
Relative percentage change <= -0.15	0
-0.15 < Relative percentage change <= -0.05	1
-0.05 < Relative percentage change <= 0	2
0 < Relative percentage change <= 0.05	3

0.05 < Relative percentage change <= 0.15	4
Relative percentage change > 0.15	5

Independent variables

- Principal components PC1 and PC2 from PCA which represent the tweets content.
- Average sentiment. The average sentiment from sentiment analysis of tweets every day.
- Number of tweets. Number of relevant tweets about Tesla every day.
- Retweet_count. Number of retweets for each tweet which indicates the influence of that tweet in twitter.
- Interaction term: Retweet_count * Average_sentiment. We assume that there is an interaction between retweet and sentiment in affecting the stock price movement. The more number of retweets, the more likely the tweet will inspire others agreement or disagreement, which may influence the overall sentiment in twitter.

Multinomial Logistic Regression Model at Time t

$$movement = intercept + \boldsymbol{\beta}_1 * PC1 + \boldsymbol{\beta}_2 * PC2 + \boldsymbol{\beta}_3 * AvgSentiment + \boldsymbol{\beta}_4 * NumTweets + \boldsymbol{\beta}_5 * RetweetCount + \boldsymbol{\beta}_6 * (RetweetCount + 1) * AvgSentiment + Error$$

Because the dependent variable has 6 levels, we applied multinomial logistic regression. We used 80% of the data to train the model and the rest 20% to test. The model accuracy is 60%.

Advanced Analyses on News

Doc2Vec training and predicting

We used a different NLP approach for the news - Doc2vec to process the different format of news' paragraphs.

In Word2vec each paragraph has a unique label. We generated it by bracketing the stock price movement for the next day of the news published date.(see Dependent variables section)

Multinomial logistic regression is the algorithm used on training. Both DM model [1] and CBOW[2] models are tried, the latter has a higher accuracy of 60%. The regression equation is:

$$movement = intercept + \boldsymbol{\beta}_1 * taggedParsedWord + Error$$

Besides the Doc2vec, we also put the average sentiment of news into a logistic regression model along with data from Twitter to predict the stock price movement for the next day. The accuracy is around 54%. The regression equation is:

$$movement = intercept + \boldsymbol{\beta}_1 * NewsSentiment + \boldsymbol{\beta}_2 * TweetsSentiment + \boldsymbol{\beta}_3 * NumTweets + \boldsymbol{\beta}_4 * RetweetCount + Error$$

Recommendations and Business Value provided

Our machine learning models will help the investment firm plan their investments in Tesla stocks by predicting the direction of stock price movement on the basis of a new news article, predicting the direction of stock price movement basis tweets and by predicting the direction/ range of percentage change in the stock price by combining both tweets and news. This will make them take a comprehensive decision by combining the insights from these models with their qualitative study of financial reports and facts about the industry.

We recommend that this analysis can be extended to find similar words that coexist in news/tweets, to see if something else is impacting the sentiment and thus the stock price- like health/environment . This will help the investors do analysis on impact of a particular sentiment. Also the analysis can be replicated for other companies by fetching data for any symbol from yahoo finance (Eg - AAPL), searching the news for them (search ‘Apple’) and search the tweets with keywords(‘apple’,’iphone 11’).

[1] DM model: Distributed Memory version of Paragraph Vector
[2] CBOW model : Distributed Bag of Words

We can also extend the analysis by collecting news from websites like Finviz.com that consolidate news from various platforms to get a stronger sense of sentiment.

The model can be improved if we collect hourly data as then we will have more data points to train the model on. We can also do the analysis on a year's data instead of over 5 months' data and then do random forest instead of logistic regression which can give higher accuracy.

Summary and Conclusions

In summary, we first used word cloud and word chart from the Word2Vec model to explore the context of relevant tweets and news about Tesla. Then we developed two models based on these two data sources to forecast the stock price movement of Tesla. We achieved forecast accuracy of 60%. In addition, we had a time-series predicting models which are commonly used by the industry to benchmark our models which had accuracy of 74% and 92%.

However, there are several limitations of our approach.

Firstly, merging the news/tweets with stock prices makes us lose some tweets/news as stock prices are not there for weekends/holidays. To expand this approach, we recommend trying to find a reasonable way of imputing the stock price on weekends or attributing the news/tweets on weekends to the next Monday as they most likely influence the stock price on the next Monday. Secondly, as most of the price movement is around -0.15 ~ 0 which is one of the levels in our dependent variable, it is not a very balanced dataset.Thirdly, there can be other confounding variables/influencers like coronavirus which can influence the whole stock market significantly. It limits our models in predicting the movement correctly. We suggested future models take these confounding variables into account.Fourthly, we only collected 160 days data from twitter and news. As neural network models need large amounts of training data to achieve accuracy, we recommend expanding the data time range or data source to gain more training data.Lastly, because there can be slang/hinglish, we can clean the data more exhaustively.

References

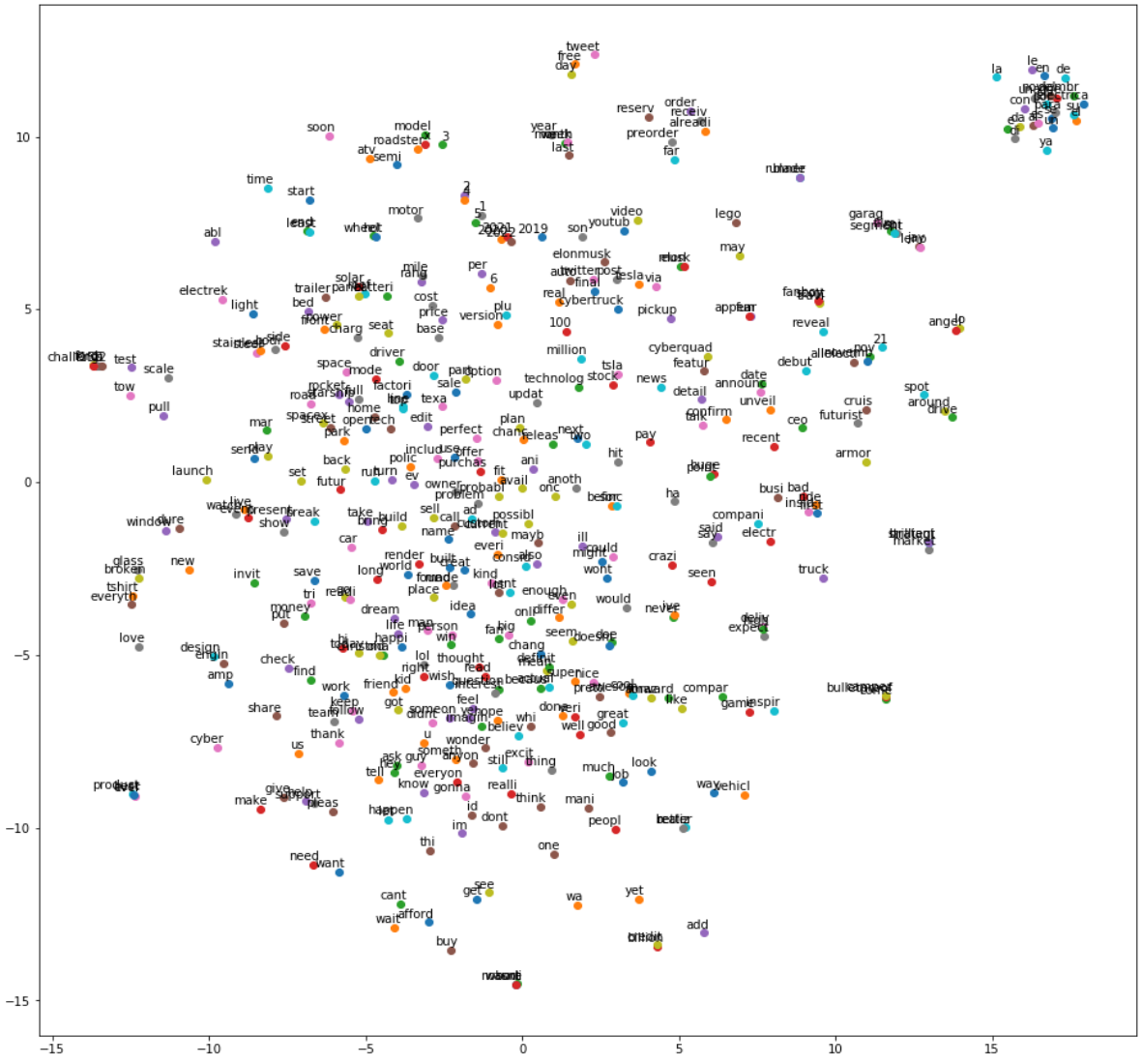
Sentiment Analysis of Twitter Data for Predicting Stock Market Movements- https://arxiv.org/pdf/1610.09225.pdf

Forecasting leading industry stock prices based on a hybrid time-series forecast model https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0209922&type=printable

Tesla spikes 20% — the most in 6 years — after shocking Wall Street by turning a profit last quarter (TSLA) https://markets.businessinsider.com/news/stocks/tesla-stock-price-surges-on-q3-earnings-beat-surprise-profit-2019-10-102862762

Appendix

Word2Vec word chart



Python notebooks and necessary files attached

1. ML_Twitter+StockpriceTime Series.pynb - It has the analysis to generate overall sentiment distribution for tweets and word cloud for the cleaned tweets. It also has the code for the two time series models for predicting stock price and the percentage change bracket in the movement of stock price

2. ML_News_Analysis - It has the Doc2Vec ,wordcloud , and models for news. It also has the model that combines the sentiments from both twitter and news.

3. Stock_Price_Movement_Prediction_Twitter.pynb - It has the word2vec code for tweets and the model for prediction using tweets.

4. Tesla_Twitter_Database.sql- It has dump of the SQL table where the tweets were saved

5. TSLA.csv- It has the stock price details from yahoo finance

6. Avg_sentiment.csv - It hast the daily average sentiment scores for tweets