Stock Market Prediction
Big Data Analysis Final Project

Jashaina Thomas & Sarah Lu

May 12, 2020

Contents

1 Executive Summary

Exec summary This report focuses on stock market prediction through mathematical modeling and big data analysis. The models discussed are mainly based on Natural Language Processing (NLP) algorithms as this topic was a very relevant one covered in this Big Data Analysis course. These models include Linear Regression, Support Vector Regression, Recurrent Neural Network, and Latent Dirichlet Allocation (LDA) models. The overall conclusions in this report are drawn using LDA topic modeling and sentiment analysis based on recent stock related data collected from twitter.

2 Introduction

Apple Inc. is an American international technology company which creates and sells electronics, software, and online services. It is considered one of the Big Five technology companies, alongside Microsoft, Amazon, Google, and Facebook. Apple (APPL) stock trades on the NASDAQ and is pretty highly valued, with a closing price of about $311.41 per share as of May 12, 2020. However, due to current the current COVID-19 pandemic, the stock market has been volatile over the past few months and APPL stock has not been much of an exception.

In this work not only we out to build a sufficient model to predict APPL stock price but more importantly draw conclusions on global sentiment about this stock during this uncertain time. The following section includes an exploration of the data used to build our models and perform our analysis. The succeeding section is dedicated to the theory methodology used to complete this project, followed by the analysis and results, and conclusion sections.

3 Data

3.1 Description

There are two data sets included in this project. One data set was used to build the Apple stock price prediction models while the other was used for the topic modeling and sentiment analysis. Although it took some time to gain access to the Twitter developer API, we were more successful in collecting quality data to perform relevant sentiment analysis on Apple stock than were were in collecting actual current stock data due to access and time constraints.

The dataset used to build our Apple stock price prediction models is from kaggle.com; kaggle.com/dgawlik/nyse provided us with financial data for SP 500 companies. We used two of four files available to us. The prices.csv file gives daily trading prices from 2010 to the end 2016 (date range is shorter for companies new on stock market). The prices-split-adjusted.csv file is similar to the prices.csv, except it includes adjustments for stock splits. The securities.csv file gives general description of each company with division based on sectors. The fundamentals.csv file provides popular fundamental indicators extracted from annual SEC 10K fill- ings to derive popular fundamental indicators. We chose to work with the prices.csv and the securities.csv files.

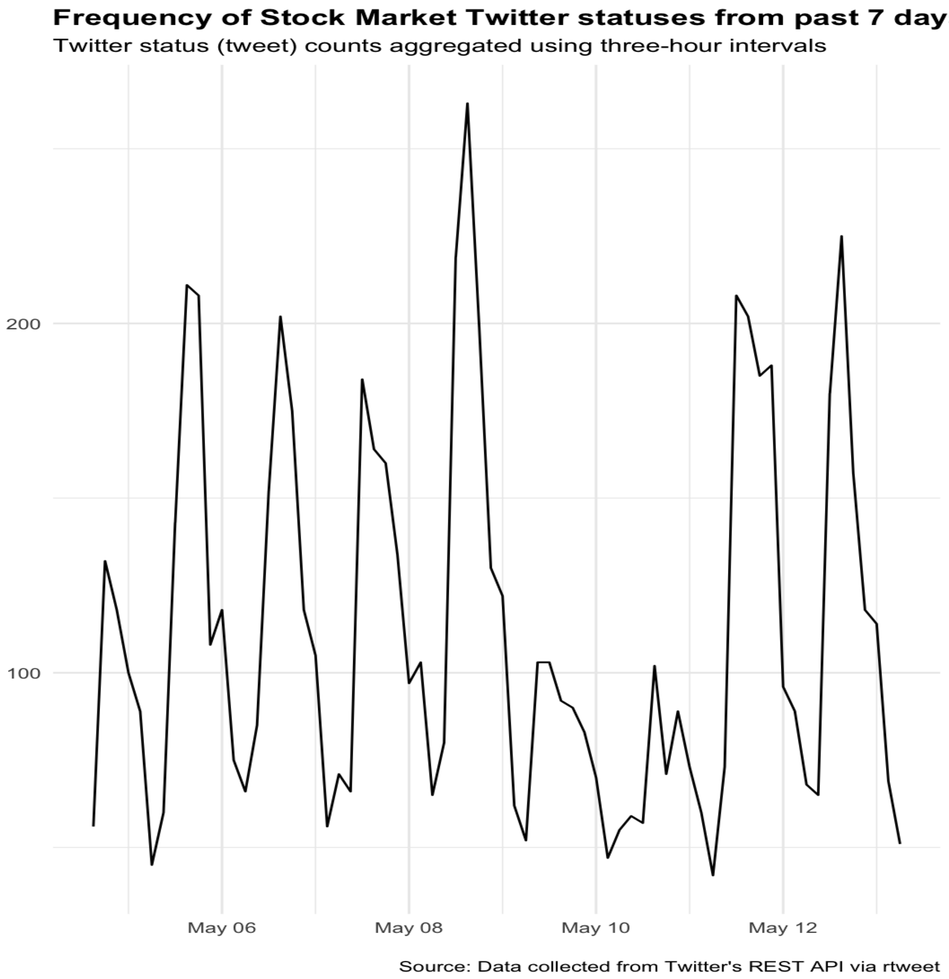
The price.csv file contains 851,246 stock data points and the securities.csv file contains 505 data points about company information. There are 7 variables used in our study, described as follows:

Name	Description
Date	Date and time of the trading information
Symbol	Unique stock symbol that identify each stock
Open	Last price anyone paid for a stock during the business day
Close	Price from the first transaction of a business day
Low	Lowest traded price for a stock during the business day
High	Highest traded price for a stock during the business day
Volume	Total quantity of shares traded for a stock
GICS Sector	Industry in which the company operates.

The dataset used to perform topic modeling and sentiment analysis consists of 7877 tweets from the past 7 days as of may 13, 2020. We have preferred to have more tweets to work with but this is data we have access to at this time. This data was collected directly from the Twitter developer API. We mined tweets related to Apple stock for the purpose of our analysis.

3.2 Data Exploration

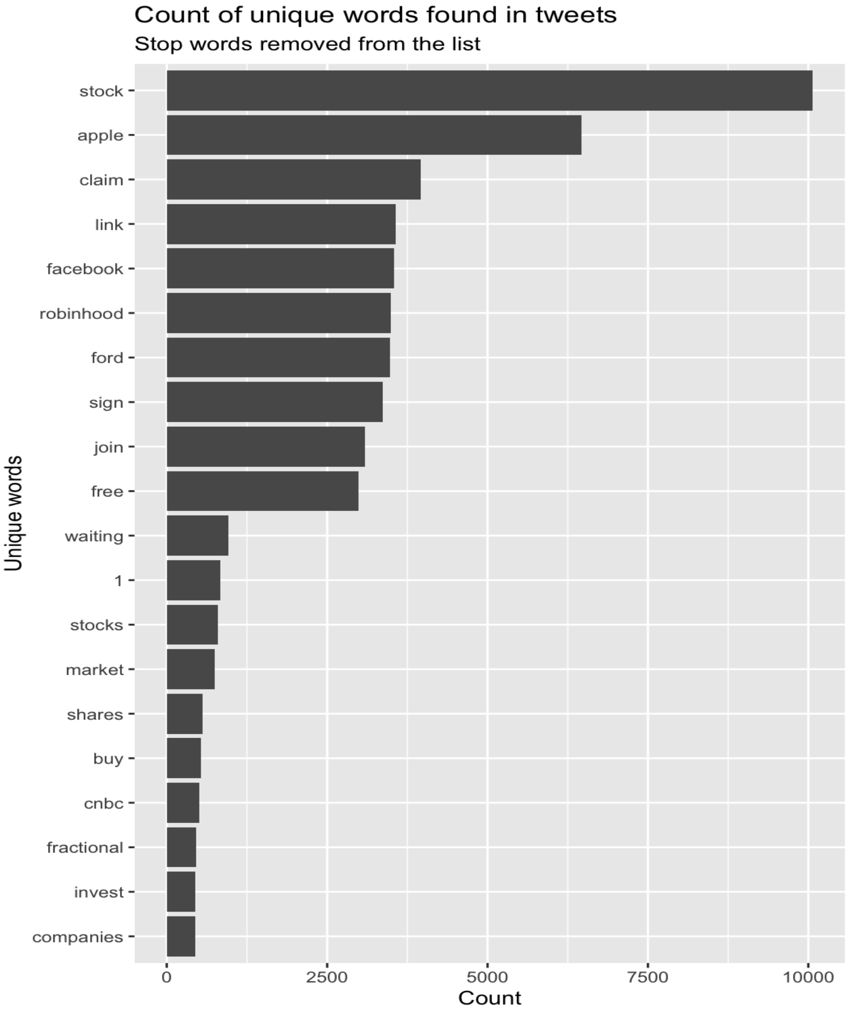
We will take a more thorough look at the twitter data used for our analysis here.


Frequency of Stock Market Twitter statuses from past 7 days

Twitter status (tweet) counts aggregated using three-hour intervals

Source: Data collected from Twitter's REST API via rtweet

We can visualize the frequency of tweets over time by aggregating the number of tweets over three-hour intervals as shown in figure 1. We can see there was a fair amount of daily posting activity related to Apple stocks. The number of tweets related to Apple stocks posted within a three hour time frame was well over 200 several times this past week. One interval had over 300 tweets while a couple had very few.



To get an idea of what individuals are tweeting about we can take a look at the frequency of unique words found in the tweets collected, we see that ”stock”, ”apple”, and ”facebook” are among the top five. This is not surprising given that we have mined tweets about specifically apple stocks. Words insignificant words such as ”the” are ignored here.

Next we are interested in visualizing all of the relationships among all words simultaneously. As one common visualization, we can arrange the words into a network graph. We found the top paired words that occur together in tweets and included them in a network showing their relationship to one another. Words included appeared in pairs at least 20 times. This visualization is included on the next page. We can see that terms such as ”appl”,“stock”, “price”, “rise” and “disney” are close to each other in this network.

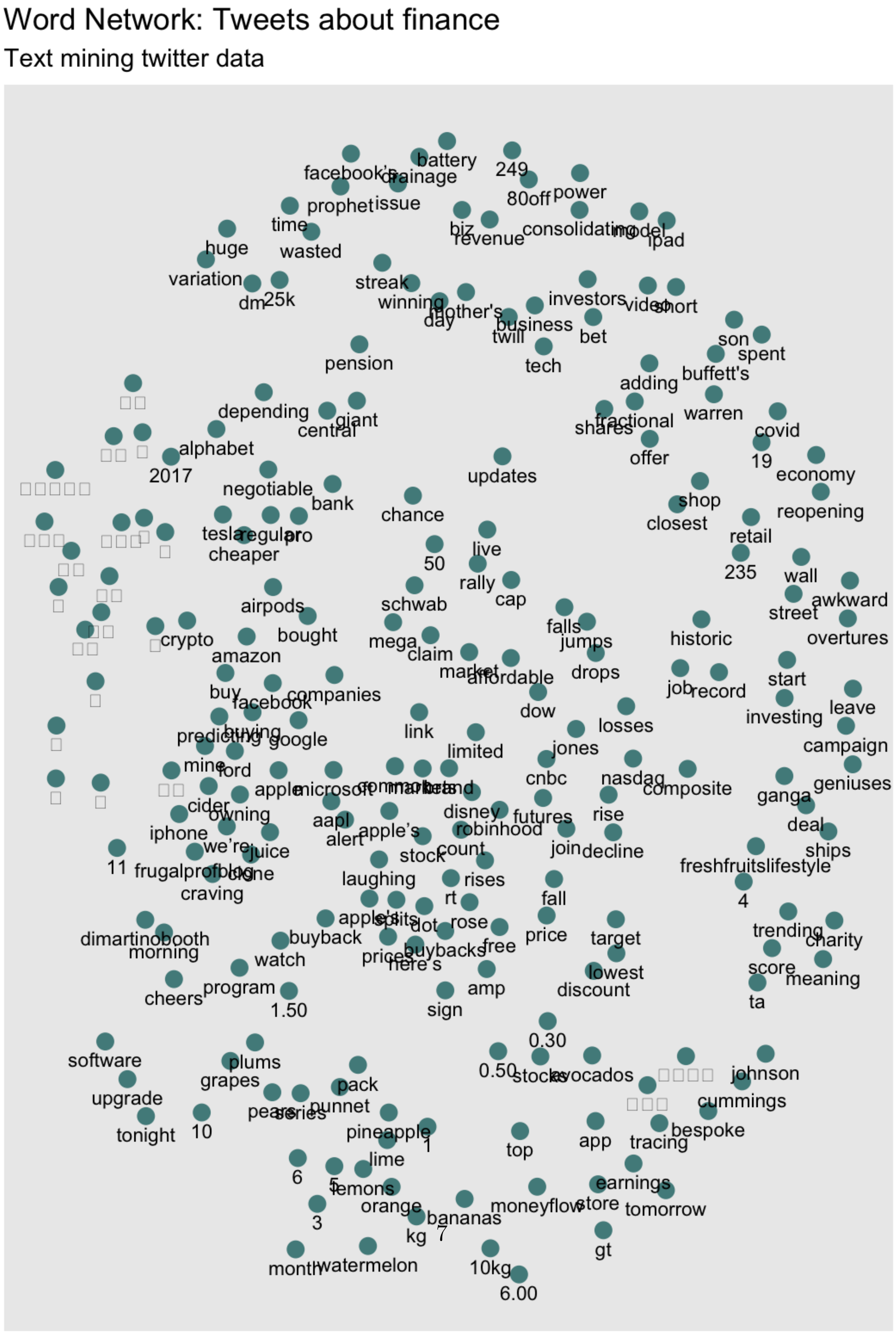

Word Network: Tweets about finance
Text mining twitter data
battery
facebook's
drainage
249
80off
power
prophet
issue
time
biz
revenue
consolidating
model
ipad
huge
wasted
variation
dm
25k
streak
winning
day
mother's
investors
video
short
business
twill
bet
son
spent
tech
pension
adding
buffett's
depending
giant
central
fractional
shares
warren
covid
offer
19
alphabet
2017
economy
updates
negotiable
bank
shop
closest
reopening
chance
tesla
regular
pro
cheaper
retail
live
50
rally
235
wall
cap
awkward
airpods
schwab
street
falls
bought
crypto
mega
jumps
historic
overtures
amazon
claim
market
affordable
drops
job
record
start
buy
facebook
companies
dow
leave
investing
predicting
buying
google
link
limited
losses
campaign
jones
mine
ford
cnbc
nasdaq
composite
apple
microsoft
common
markets
brand
ganga
geniuses
cider
owning
aapl
disney
futures
rise
deal
iphone
alert
apple's
robinhood
join
decline
ships
we're
juice
stock
count
11
frugalprofblog
clone
freshfruitslifestyle
craving
laughing
rises
4
rt
apple's
splits
dot
rose
fall
trending
charity
dimartinobooth
morning
buyback
watch
prices
buybacks
here's
free
price
target
score
meaning
lowest
program
cheers
amp
discount
ta
1.50
sign
0.30
software
plums
grapes
0.50
stocks
avocados
johnson
pack
upgrade
cummings
pears
series
punnet
tonight
10
pineapple
1
top
app
tracing
bespoke
lime
6
5
lemons
earnings
orange
3
bananas
moneyflow
store
tomorrow
kg
7
gt
month
watermelon
10kg
6.00

Figure 1: ...

4 Models & Methodology

4.1 Prediction Using Support Regression Model

The first task we set out to do is predict stock price. Because there are a lot of factors that can influence the stock market, the stock price prediction problem has always been very complicated. Support Vector Regression is a tool from machine learning that can build a regression model based on the historical time series data in the purpose of predicting the future trend of the stock price. We did this by running a local support vector regression on every five days of data to develop a model to fit our data. Apple stocks price is predicted based on its daily opening, closing, high, and low prices along with its volume and sector.

4.2 Prediction Using Recurrent Neural Network Model

A Recurrent Neural Network (RNN) is a type of neural network well-suited to time series data and have proved to be one of the most powerful models for processing sequential data. RNNs process a time series step-by-step, maintaining an internal state summarizing the information they've seen so far. In this project, we construct RNN with the Long Short-Term Memory (LSTM) approach to predict Apple stock prices. Our LSTM model consists of a sequential input layer followed by 2 LSTM layers and a dense output layer with linear activation function.

4.3 Latent Dirichlet Allocation Topic Modeling

Latent Dirichlet allocation (LDA) is a topic model which allows sets of observations to be explained by unobserved groups showing how some parts of the data are similar. In this case, the data points are words collected into documents which are tweets. Each document is a mixture of 5 topics and each word's presence is attributable to one at least of the document's topics. In LDA the topic distribution is assumed to have a sparse Dirichlet prior. The sparse Dirichlet priors encode the intuition that documents cover only a small set of topics and that topics use only a small set of words frequently. This results in a better rehashing of words and a more precise assignment of documents to topics.

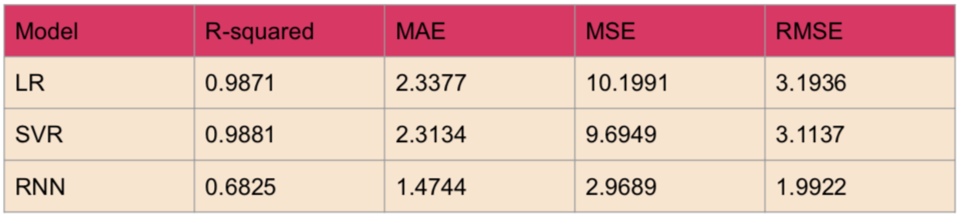
4.4 Sentiment Analysis

Sentiment analysis is the interpretation and classification of emotions (positive, negative and neutral) within text data using text analysis techniques. This analysis allows one to identify sentiment toward products, brands or services in online conversations and feedback. We focus on sentiments towards Apple stock found in online conversation on Twitter, in this example.

5 Analysis & Results

5.1 Prediction Model Evaluation Metrics

We will focus on the performance of our prediction models rather than the output as the input data available to us was not ideal for our purposes. The following indicators evaluate the performance of our local support vector regression and recurrent neural network prediction models:



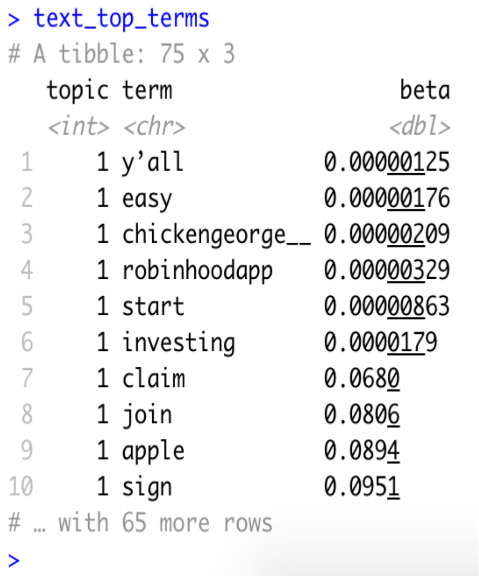
Model	R-squared	MAE	MSE	RMSE
LR	0.9871	2.3377	10.1991	3.1936
SVR	0.9881	2.3134	9.6949	3.1137
RNN	0.6825	1.4744	2.9689	1.9922

In a previous work we constructed a local linear regression (LR) model to predict stock price in a fashion similar to our support vector regression (SVR) model discussed in this report so we have included the results of that model just for comparison. We can see by the R-squared metric that the SVR model performs the best. However based on the error metrics, the recurrent neural network performed the best. An R-squared score of 0.9881 suggests that approximately 98.81% of the variation in the dependent variable, which is closing stock price, is explained by the variables we included in our SVR model. These indicators suggest that our model gives an accurate prediction of the 5-day forward stock prices we set out to predict. These are very good results. Comparing the RNN and SVR models, one may prefer higher R-squared to a lower mean squared error if they are more concerned with the meaningfulness of model. In this case, the SVR model is preferable. On the other hand, one might prefer the lower MSE if accuracy is more of a concern in which they would prefer the RNN model.

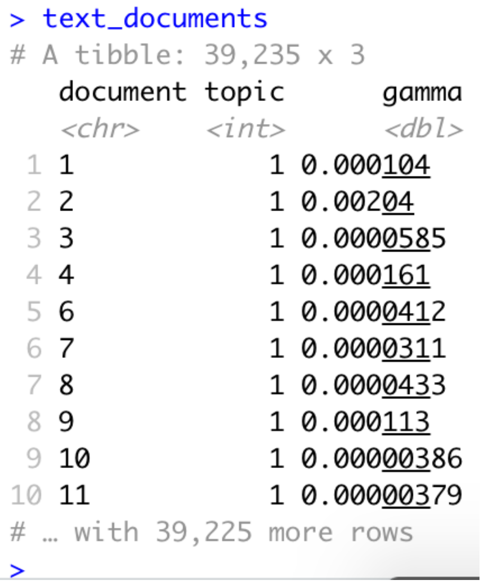
5.2 Topic Modeling & Sentiment Analysis Results

Next, we will take a look at the Latent Dirichlet Allocation topic modeling results. To analyze the twitter data we conctructed a five-topic LDA model. We can interpret the model by extracting the per-topic-per-word probabilities, called β ”beta”. The per-topic-per-word probabilities are shown below for some of the words in topic 1. For example, we can observe that the term “apple” has 0.0894 probability of being generated from topic 1.

We can also extract the per-document-per-topic probabilities, called γ ”gamma”. This is the estimated proportion of words from that document that are generated from that topic. As mentioned earlier, each document is a tweet in this case and there are many so the probabilities can be pretty small. Typically topic modeling will have far more words than documents so our γ values are not as interesting or useful for this case.

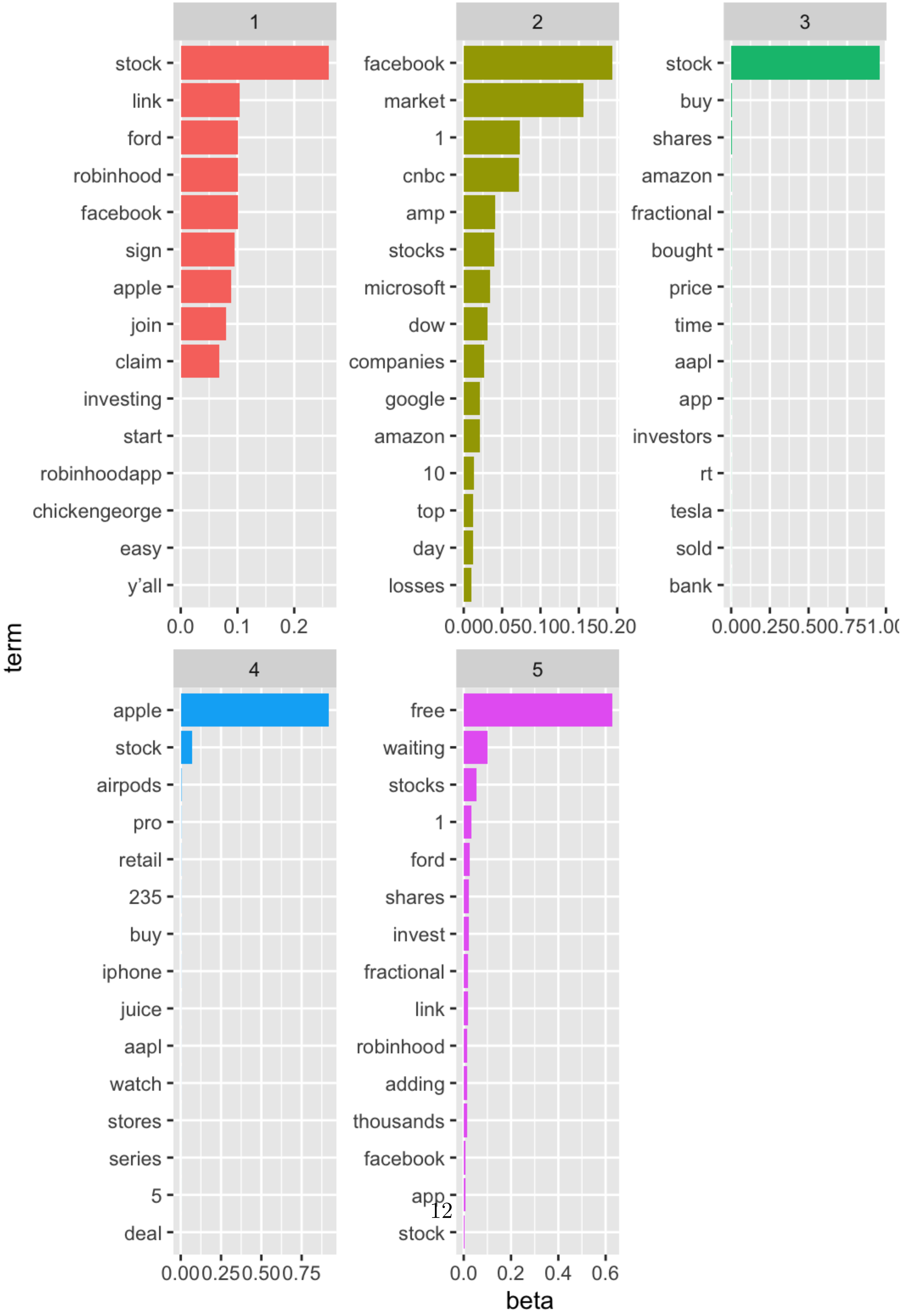


```
> text_top_terms
# A tibble: 75 x 3
   topic term                 beta
   <int> <chr>               <dbl>
 1     1 y’all          0.00000125
 2     1 easy           0.00000176
 3     1 chickengeorge__ 0.00000209
 4     1 robinhoodapp   0.00000329
 5     1 start          0.00000863
 6     1 investing      0.0000179
 7     1 claim          0.0680
 8     1 join           0.0806
 9     1 apple          0.0894
10     1 sign           0.0951
# … with 65 more rows
>
```

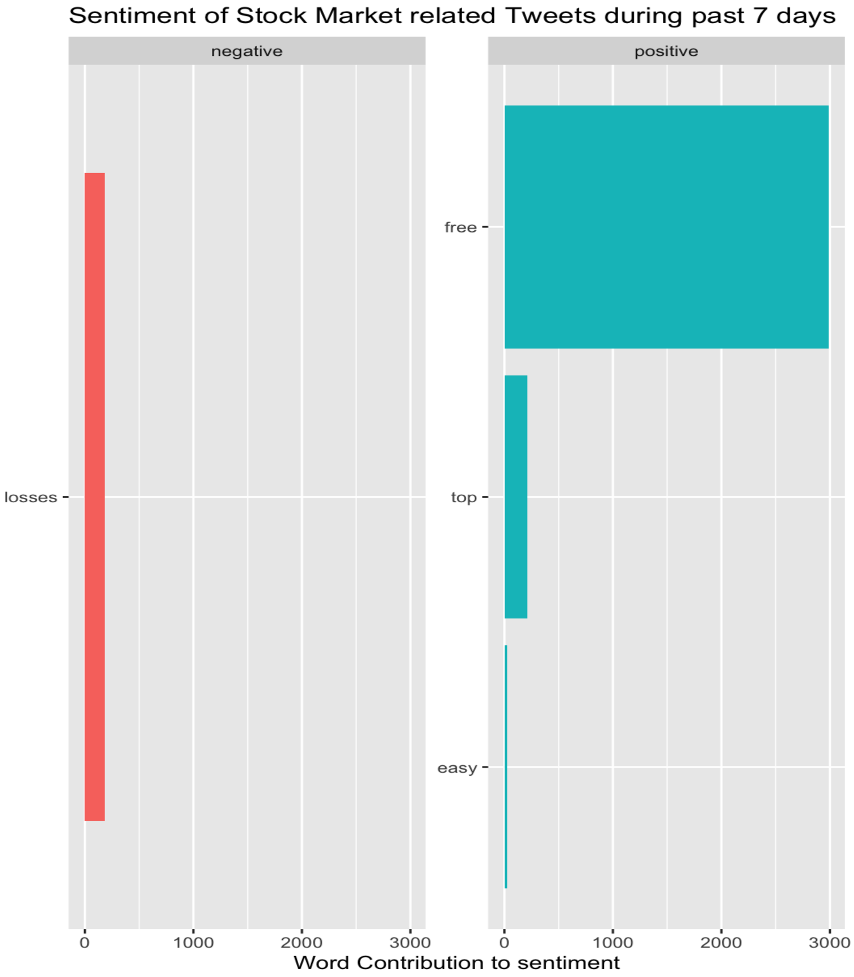


```
> text_documents
# A tibble: 39,235 x 3
   document topic      gamma
   <chr>    <int>      <dbl>
 1 1            1 0.000104
 2 2            1 0.00204
 3 3            1 0.0000585
 4 4            1 0.000161
 5 6            1 0.0000412
 6 7            1 0.0000311
 7 8            1 0.0000433
 8 9            1 0.000113
 9 10           1 0.00000386
10 11           1 0.00000379
# … with 39,225 more rows
>
```

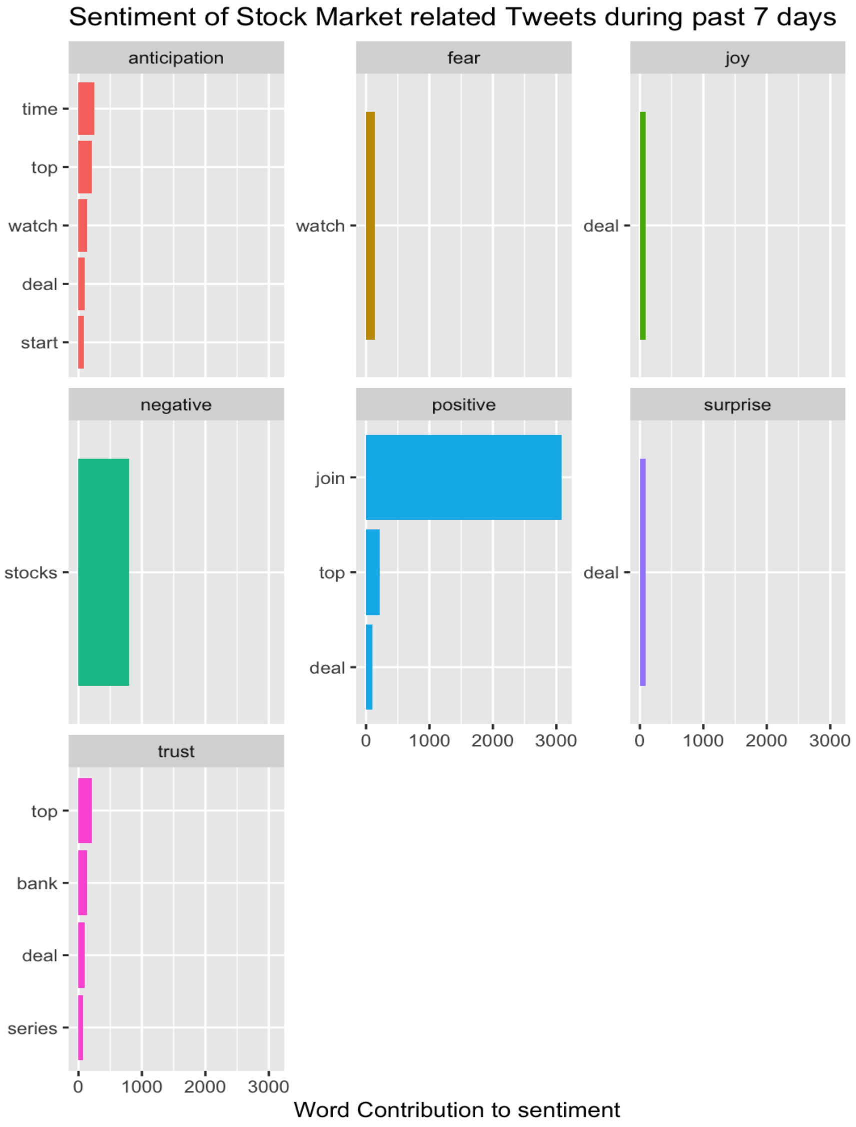
Now that we have the per-topic-per-word probabilities we can take a look at the top terms among each topic. The next page includes plots corresponding to the top 15 words that are most common within each of the five topics based on their β value. All topics appear relevant to Apple stock accept topic five based the top words included in the plot (Does not include ”apple”, ”appl”).


1
stock
link
ford
robinhood
facebook
sign
apple
join
claim
investing
start
robinhoodapp
chickengeorge
easy
y'all
0.0
0.1
0.2
2
facebook
market
1
cnbc
amp
stocks
microsoft
dow
companies
google
amazon
10
top
day
losses
0.000.050.100.150.20
3
stock
buy
shares
amazon
fractional
bought
price
time
aapl
app
investors
rt
tesla
sold
bank
0.000.250.500.751.0(
term
4
apple
stock
airpods
pro
retail
235
buy
iphone
juice
aapl
watch
stores
series
5
deal
0.000.250.500.75
5
free
waiting
stocks
1
ford
shares
invest
fractional
link
robinhood
adding
thousands
facebook
app
stock
0.0
0.2
0.4
0.6
beta

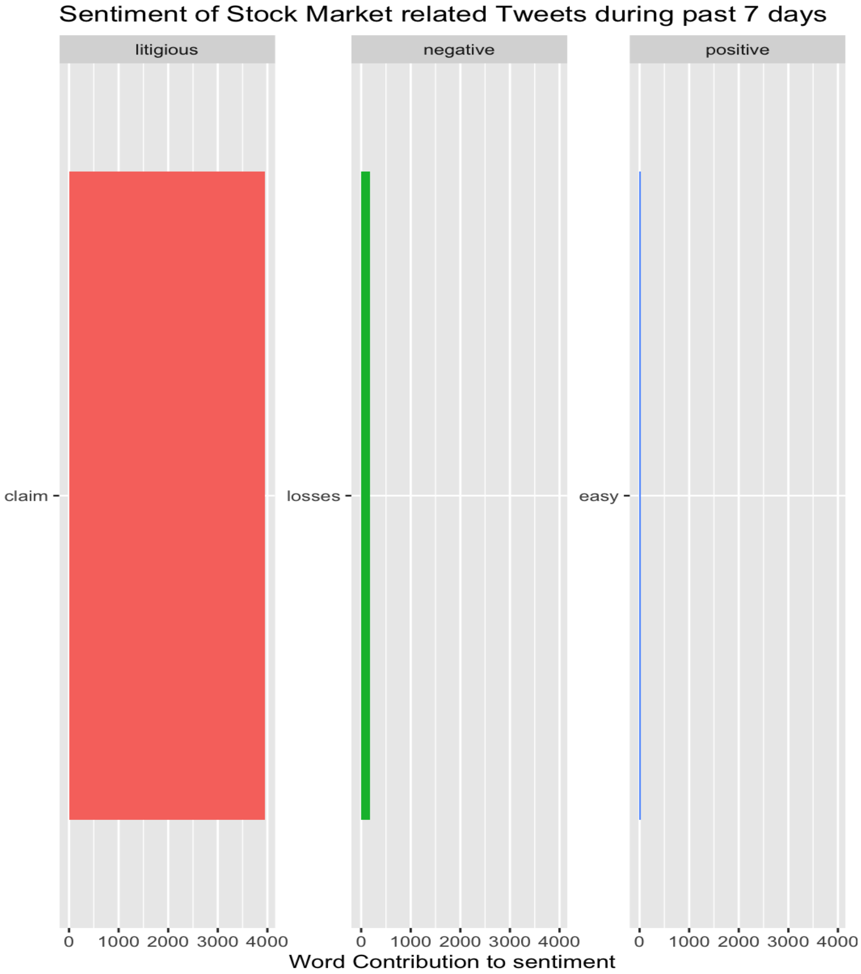
We can now perform sentiment analysis based on top the terms found in each of the five topics using different sentiment lexicons.



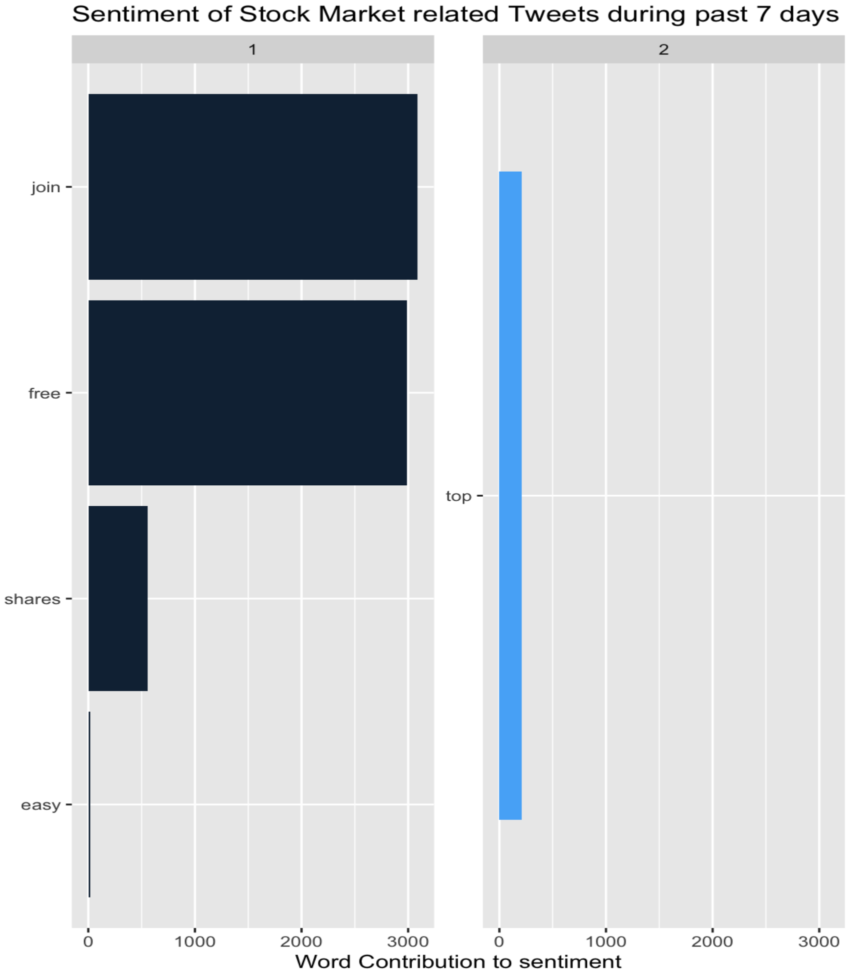
This graph shows the sentiments of key words included in the top tweets. These terms have been classifies as either negative or positive sentiments and are plotted against their contribution to that sentiment. Ignoring ”free” as it was included in topic 5 which did not seem related, we see that ”losses” contributed to less negative sentiment while ”top”and ”easy” contributed to more positive sentiment.



This graph also shows the sentiments of key words included in the top tweets. These terms have been classifies as either negative, positive, and other sentiments and are plotted against their contribution to that sentiment. We can see that many terms here contribute to the anticipation, positive and trust sentiments. Most of the sentiments found among the top words here are positive sentiments except for ”negative” associated with the term ”stocks” and fear associated with the term ”watch”.



This graph also shows the sentiments of key words included in the top tweets. These terms have been classifies as either negative, positive, and litigious, and are plotted against their contribution to that sentiment. We can see that the few terms here contribute to all three sentiments. The negative and litigious sentiments here can be viewed as negative sentiments for our purposes.



This graph also shows the sentiments of key words included in the top tweets but in a different format. The sentiment lexicon used rates terms for valence with an integer between minus five (negative) and plus five (positive). These terms have been classifies as either negative, positive on a sliding scale, and are plotted against their contribution to that sentiment. We can see that all of the terms here contribute to positive sentiment as they are all valued above zero.

6 Conclusion

We have built several sufficient models to predict Apple stock price and drawn some basic conclusions on current global sentiment towards this stock. Based on our analysis and results we find that although there was some negative sentiment, the overall sentiment towards Apple stock is positive over the past seven days according to twitter data. There also appears to be some degree of uncertainly accompanied with emotions such as fear and litigious. This is expected as the entire stock market is currently displaying a decrease in value and an increase volatility. It would be valuable to further derive and plot sentiment towards Apple stock over time to see how the current COVID-19 crisis may has had an affect.

Since only a certain amount of data was accessible for this project due to time and access constraints, it would be a major improvement to be able to collect more tweets at a time from a longer time period since they are extremely short documents and there are so many available. Finding a way to include sentiment analysis output in stock price prediction models like the ones described in this work is a complex but could also be extremely beneficial. Lastly, given more time we have liked to attempt combining documents from data sources other than just Twitter such as Bloomberg or Yahoo Finance news to further improve the quality of our data.

7 Appendix

7.1 References

https://www.kaggle.com/dgawlik/nyse