# CS688 – Term Project

# Twitter stock market sentiment analysis:

Gaurav Tungare

Facilitator: Dr. Teresa

Apr 19, 2021

1.Objective

## To gather into Twitter stock market tweets and performing sentiment analysis

2. Background of Dataset: identify the

Gainers & loser stock

**Identification of Gainer & looser stock**

I was looking for ways to pick up the gainer stock, lot of interesting stock did come my way but I had some clear identification criteria for picking up the gainers

The three-stocks picked as gainer stock are Tesla, Conibase and Moderna

The three-stocks picked as looser stock are AMC Entertainment, Peloton and J&J

**Snapshots of Choices – Identification Criteria – Gainer Stock**

**Snapshots of Choices – Looser Stock**

I am searching for 100 tweets per country, retweets in search result is set to False to make ensure not to include any retweets

The following additional steps were taken to achieve it

* I have used saveRDS an readRDS so that the tweets which I have use can be replicated by facilitator incase it needs to be replicated.

3) Process Use of the pre-processing transformations

I have used multiple pre-preprocessing transformation to clean up the text in the tweets.

The following are the pre-preprocessing step performed to achieve a clean text tweet

* Remove twitters #tags
* Remove emojis AND a lot of other stuff (other languages, etc.)
* Remove http elements manually
* Remove special characters “[\\()\*&^%$#@.?<>=+;'!,\_/~|:-]” manually
* Convert to lower case
* Remove non alpha numeric elements manually
* Remove all the numeric characters

Please find example of one tweet to see the preprocesing steps.

The order of processing is maintained below, it is a 7 step preprocessing

|  |  |  |
| --- | --- | --- |
| Original gainer tweet no 3 | @TMobile @Tesla I’m using resumes or bags when going grocery shopping and using rechargeable batteries for my things. :] | Comments  As we can see we have lots of here in one tweet  We have Twitter # tags - @TMobile @Tesla  We have lots of unwanted characters like : , ], - etc |
| Step 1: Remove the twitter # tags names | I’m using resumes or bags when going grocery shopping and using rechargeable batteries for my things. :] | Twitter # tags @TMobile @Tesla are removed |
| Step 2: Remove other UTFs | Im using resumes or bags when going grocery shopping and using rechargeable batteries for my things. :] | No change as there was no other UFT character |
| Step 3: Remove http elements manually | Im using resumes or bags when going grocery shopping and using rechargeable batteries for my things. :] | No change as there was not http character |
| Step 4: Remove all "()\*&^%$#@.?<>;'!," characters | Im using resumes or bags when going grocery shopping and using rechargeable batteries for my things ] | Character “:” is removed  As you can see character “]" is still there but we have a safety net further to catch that |
| Step 5: Convert to lower case | im using resumes or bags when going grocery shopping and using rechargeable batteries for my things ] | Converting to lower case |
| Step 6: Remove all the non-alphanumeric | im using resumes or bags when going grocery shopping and using rechargeable batteries for my things | Character “]” is removed, it is a clean version |
| Step 7: Remove all the numeric characters | im using resumes or bags when going grocery shopping and using rechargeable batteries for my things | There are no numbers so there is no change here |

4) Process each set of tweets into tidy text or corpus objects.

So, to achieve this functionality unnest\_tokens was used.

* 300 tweets for Gainer resulted in total of “2934” word
* 100 tweets for looser resulted in total of “3299” word

5 ) List of the most frequent terms for Gainer and Looser tweets.

Dplyr function was used to get the most frequent words.

**Most frequent terms – Gainer Stock**

|  |  |  |
| --- | --- | --- |
|  | word | n |
| 1 | moderna | 79 |
| 2 | coinbase | 34 |
| 3 | pfizer | 33 |
| 4 | tesla | 28 |
| 5 | vaccine | 19 |
| 6 | water | 14 |
| 7 | recycle | 13 |
| 8 | recycling | 13 |
| 9 | covid | 10 |
| 10 | dont | 10 |
| 11 | dosis | 10 |
| 12 | reusable | 10 |

**Most frequent terms – Looser Stock**

|  |  |  |
| --- | --- | --- |
|  | word | n |
| 1 | johnson | 93 |
| 2 | peloton | 72 |
| 3 | boris | 33 |
| 4 | covid | 13 |
| 5 | football | 12 |
| 6 | treadmill | 12 |
| 7 | dont | 11 |
| 8 | pelotn | 11 |
| 9 | people | 11 |
| 10 | time | 9 |
| 11 | tsla | 9 |
| 12 | vaccine | 9 |
| 13 | week | 9 |

**Most frequent terms – Gainer Stock**

Chart

Description automatically generated

**Most frequent terms – Looser Stock**

A picture containing chart

Description automatically generated

There are couple of interesting **observations**

**For Gainer Stock**

* Vaccine related words are quite prevalent
* Water, recycling, recyle & reusable are appearing frequency indicating a relationship between electric vehicles , Tesla and environment
* Vaccine companies are appearing frequency

**For Looser stock**

* Surprisingly, johnson word is appearing frequently which have interesting correlation to British prime minister Boris Johnson so we you can the search is not ideal and it has both Johnson&Johnson company and the British prime minister
* Also interesting is the word treadmill related to peloton which was in the news for infant accident

6) Top word pairs (bigrams) for stocks

The following was done to achieve this as described in the class

* The original cleaned text was taken and given as input to unnest\_tokens with token = "ngrams"
* The word pair comes as a complete string but where separated in two separate words
* Any words in the pair which are part of stop words were filtered/ removed
* Additional filter criteria was applied to remove any word pairs containing ‘NA’ as well
* Once we had clean pair word , it was given as input to ggplot geom\_edge\_link to plot the ngram

**Top Word Pair from Country – Gainer Stock**

|  |  |  |  |
| --- | --- | --- | --- |
|  | word | word | n |
| 1 | dosis | de | 9 |
| 2 | de | pfizer | 8 |
| 3 | de | las | 7 |
| 4 | moderna | vaccine | 7 |
| 5 | de | la | 6 |
| 6 | elon | musk | 6 |
| 7 | fatal | tesla | 6 |
| 8 | la | segunda | 6 |
| 9 | las | vacunas | 6 |
| 10 | lost | billion | 6 |

**Top Word Pair from Country – Looser Stock**

|  |  |  |  |
| --- | --- | --- | --- |
|  | word | word | n |
| 1 | boris | johnson | 28 |
| 2 | amp | johnson | 14 |
| 3 | johnson | amp | 13 |
| 4 | gme | amc | 9 |
| 5 | amc | gme | 7 |
| 6 | de | la | 7 |
| 7 | de | los | 4 |
| 8 | johnson | vaccine | 4 |
| 9 | le | peloton | 4 |
| 10 | peloton | tread | 4 |

**Graphical Representation**

**Word Network – Gainer Stock**

A picture containing text, indoor, map, several

Description automatically generated

There are couple of interesting **observations**

**For Gainer Stock**

* Bicoin-buy-sell-hold-avg is appearing which is interesting as those indicate the patters which people are tweeting
* Fatal-tesla-crash again a pattern which is appearing which is related to recent event where tesla car was involved in a crash
* Grocery-bags-shopping-reusable are appearing frequency indicating a relationship between electric vehicles , Tesla and environment
* Moderna-shot-vaccine are appearing frequency relating to how the pars are related
* Also I see some Spanish wards appearing in the word pairs

**Word Network – Looser Stock**

A screenshot of a computer

Description automatically generated with low confidence

There are couple of interesting **observations**

**For Looser stock**

* The word pair peleton-workout-recall are appearing which makes sense since there is incident which has happened with infant getting injured in the news
* Words like blood-clots are there everywhere because of the concerns around the J&J vaccine
* Words like consumer-protection-product-safety appears indicating the sentiments around the peleton incident
* Also we see something for hethrow-expansion which indicate that the search is not perfect since it is picking that from johnson word search

7) Sentiment Analysis

Sentiment score for all the tweets for gainer stock & Loser stock

The following was done to achieve the sentiments analysis score described in the class lecture. Following are the steps done to come up with the sentiment score

* Cleaned text from original tweet were taken and passed through sentiment “bing” function. so it resulted in following observation
  + For Gainer Stock - 125 observation
  + For Looser stock – 186 observation
* Each cleaned text of tween was passed through customer function to give every tweet a score, Type 1 (contains both positive and negative sentiments or tweets with no words in the bing list) or Type 2 with positive/negative sentiments
* **Plot for Sentiment for Gainer Stock**

Chart, bar chart

Description automatically generated

There are couple of interesting **observations**

**For Gainer stock**

* On positive side words coming out are support, protect and help
* We have words like crash, miserable ad fatigues which are common in the covid times
* **Plot for Sentiment for Loser Stock**

Chart, bar chart

Description automatically generated

There are couple of interesting **observations**

**For Loser stock**

* On positive side words coming out are trust, support and free which are natural sentiments
* We have words like risks, warning, dangerous and urgent which correlates to the sentiments

**Combined Plot**

* **Plot for Sentiment Score with Type 1 and Type 2 classification**

Chart, histogram

Description automatically generated

**Observations**

**For Gainer stock**

* As you can see that there are high score for 0 which is expected as we have both positive and negative sentiments expressed. Also we have more sentiments for gainer stock on the positive side

**For Loser stock**

* As you can see that there are high score for 0 which is expected as we have both positive and negative sentiments expressed
* We have loser sentiments on the negative side as well
* **Plot for Sentiment Score with ONLY Type 2 classification**

Chart, histogram

Description automatically generated

**Observations**

**For Gainer stock**

* We have equal score on for the gainer stock and are some sentiments across even for score of 3

**For Loser stock**

* We have loser sentiments on the negative side as well
* **Tabular Format - Sentiment Score with Type 2 classification only**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Stock | Count | Mean | SD | max | min |
| 1 | Gainer | 120 | 0.0167 | 1.47 | 4 | -4 |
| 2 | Loser | 133 | -0.263 | 1.49 | 3 | -4 |

* **Observation & comparison for sentiment score** 
  + The result does make senses so let me explain f
    - For Gainer stock
      * There is higher density of score above zero which indicate that there are positive sentiments for the stock. There are several factor affecting which reflects in the tweet with words like “free”, “clean” appearing often resulting the more density of score above zero
      * Also there are negative sentiments expressed with words containing “fatal”, ”hard” and “issues” being expressed often as well
    - For Looser stock
      * There are high score below zero which indicate that there are negative sentiments there is a high peak for positive sentiments as well. There are several factor affecting negative sentiments which reflects in the tweet with words like “dangerous”, “crisis” appearing often resulting the more density of score below zero
      * On a positive side there are less sentiments expressed with words containing “support”, ”gain” and “support” being expressed often resulting in post covid wave across the country resulting in positive sentiments

8) Data Visualization

* For Gainer stock

Chart, line chart

Description automatically generated

* **Observation**
* For Gainer stock
  + Tesla as a higher price point of the overall stock which reflect in the plot as the price point it around $720
  + For Coinbase closing price of $ 291.60
  + For Moerna closing price of $173.63
* For Loser stock

Chart, line chart

Description automatically generated

* **Observation**
* For Loser stock
  + J&J stock gone through lot of swing ,also AMC stock has seen price fluctuation
  + For AMC closing price of $10.16
  + For J&J closing price of $165.52
  + For Peleton closing price of $101.07

Appendix

* For Gainer stock - variation for entire year is captured
* As you can see coin base was just launched in Apr 21

Chart

Description automatically generated