Analyzing Trends from Twitter Tweets
Final Report

Team Members
1. Fagun Raithatha
2. Vinod Kumar
3. Aanchal Kumar Rohira

Introduction
Twitter is a social media platform where people share their thoughts by posting tweets. It is a popular platform for getting insights into trending topics, people's opinions about various issues, and the latest news and announcements from popular influencers. A tweet from an influencer can influence many things, including but not limited to the stock market. Tweets are a valuable measure of understanding people's opinions, ideas, and responses to the world. The project aims to analyze the trend from tweets on Twitter from April 2020 to July 2020. This project helps to understand users' significant and minor interests using their tweets. We can explore the popularity of stocks and cryptocurrencies and which one to invest in. We aim to use big data tools and techniques to extract relevant information from a continuous stream of data originating from Twitter.

Related work
There is related work on this dataset. People have tried to perform sentiment analysis on this dataset. They used text mining and natural language processing techniques to classify a tweet as positive or negative. We extended this idea by analyzing a large fragment of tweets to discover the patterns and trends. We are following an unsupervised learning approach.

Tools
- Python
- Pandas and NumPy
- Natural Language Toolkit
- TextBlob
- Matplotlib
- word cloud
- Scikit-learn - Kmeans, a bag of words
- MlXtend (machine learning extensions) - frequent patterns
    - FP Growth algorithm
    - Association rules

Approach

We used a dataset for Kaggle that contains a million tweets gathered between April 2020 to July 2020. The dataset contains a million rows where each entity represents the tweet's date and a text field containing the raw text of the tweet. We only extracted the tweet information. Hence, the type of input data for this project is entirely textual data. Following are the steps to approach this problem.

1. Text cleaning - removing unwanted characters, removing/retaining stopwords, breaking the attached word, lemmatization/stemming, and spell and grammar correction
2. Using Natural Language Processing in Python, Tokenization - breaks each sentence into separate tokens (words).
3. Convert tokens (words) to a standard tabular format. Columns represent words, and rows represent a tweet and the frequency of each expression in that tweet.
4. Ran machine learning algorithms to cluster the data.
5. Used Association rule mining to understand correlation among topics.
6. Visualize the data and identify trends, such as the most popular topic or stock.
7. Generated a report on the findings

Challenges

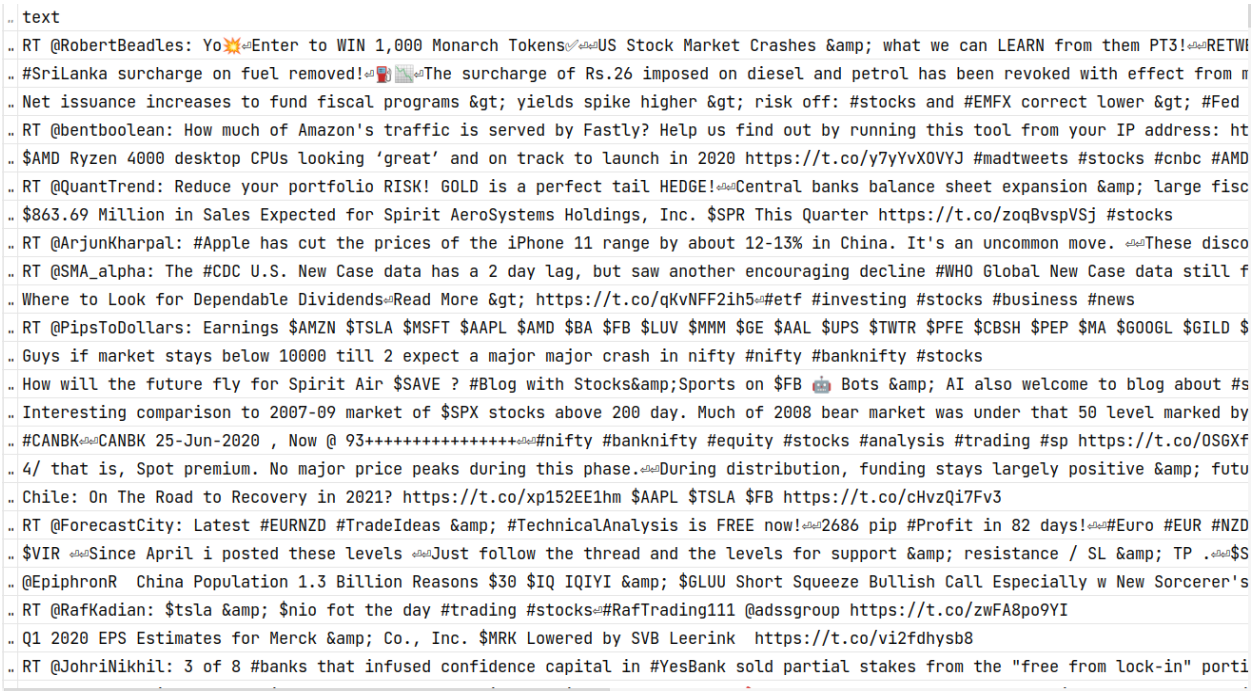
1. Storing and working on big data
2. Working on high-dimensional data (approximately 15000 columns)
3. Handling the large sparse matrix
4. Visualization of results
5. Finding rules through association rule mining on extensive data
6. Using a variety of machine learning algorithms to discover hidden information

Result

Following are our step-by-step results.

1. **Text preprocessing**

   We first preprocessed the textual data. We cleaned the text using the natural language toolkit in Python in preprocessing. Following is the raw text that we received from Kaggle.



text
RT @RobertBeadles: Yo💥⏎Enter to WIN 1,000 Monarch Tokens🦋⏎⏎US Stock Market Crashes & what we can LEARN from them PT3!⏎⏎RETWI
#SriLanka surcharge on fuel removed!⏎⛽📉⏎The surcharge of Rs.26 imposed on diesel and petrol has been revoked with effect from m
Net issuance increases to fund fiscal programs > yields spike higher > risk off: #stocks and #EMFX correct lower > #Fed
RT @bentboolean: How much of Amazon's traffic is served by Fastly? Help us find out by running this tool from your IP address: ht
$AMD Ryzen 4000 desktop CPUs looking 'great' and on track to launch in 2020 https://t.co/y7yYvXOVYJ #madtweets #stocks #cnbc #AMD
RT @QuantTrend: Reduce your portfolio RISK! GOLD is a perfect tail HEDGE!⏎⏎Central banks balance sheet expansion & large fisc
$863.69 Million in Sales Expected for Spirit AeroSystems Holdings, Inc. $SPR This Quarter https://t.co/zoqBvspVSj #stocks
RT @ArjunKharpal: #Apple has cut the prices of the iPhone 11 range by about 12-13% in China. It's an uncommon move. ⏎⏎These disco
RT @SMA_alpha: The #CDC U.S. New Case data has a 2 day lag, but saw another encouraging decline #WHO Global New Case data still f
Where to Look for Dependable Dividends⏎Read More > https://t.co/qKvNFF2ih5⏎#etf #investing #stocks #business #news
RT @PipsToDollars: Earnings $AMZN $TSLA $MSFT $AAPL $AMD $BA $FB $LUV $MMM $GE $AAL $UPS $TWTR $PFE $CBSH $PEP $MA $GOOGL $GILD $
Guys if market stays below 10000 till 2 expect a major major crash in nifty #nifty #banknifty #stocks
How will the future fly for Spirit Air $SAVE ? #Blog with Stocks&Sports on $FB 🤖 Bots & AI also welcome to blog about #s
Interesting comparison to 2007-09 market of $SPX stocks above 200 day. Much of 2008 bear market was under that 50 level marked by
#CANBK⏎⏎CANBK 25-Jun-2020 , Now @ 93++++++++++++++++⏎⏎#nifty #banknifty #equity #stocks #analysis #trading #sp https://t.co/OSGXf
4/ that is, Spot premium. No major price peaks during this phase.⏎⏎During distribution, funding stays largely positive & futu
Chile: On The Road to Recovery in 2021? https://t.co/xp152EE1hm $AAPL $TSLA $FB https://t.co/cHvzQi7Fv3
RT @ForecastCity: Latest #EURNZD #TradeIdeas & #TechnicalAnalysis is FREE now!⏎⏎2686 pip #Profit in 82 days!⏎⏎#Euro #EUR #NZD
$VIR ⏎⏎Since April i posted these levels ⏎⏎Just follow the thread and the levels for support & resistance / SL & TP .⏎⏎$S
@EpiphronR China Population 1.3 Billion Reasons $30 $IQ IQIYI & $GLUU Short Squeeze Bullish Call Especially w New Sorcerer's
RT @RafKadian: $tsla & $nio fot the day #trading #stocks⏎#RafTrading111 @adssgroup https://t.co/zwFA8po9YI
Q1 2020 EPS Estimates for Merck & Co., Inc. $MRK Lowered by SVB Leerink https://t.co/vi2fdhysb8
RT @JohriNikhil: 3 of 8 #banks that infused confidence capital in #YesBank sold partial stakes from the "free from lock-in" porti

   We applied several cleaning methods to this raw text to clean it. These methods are the following.

   1. Convert each word to lowercase letters
   2. Remove stopwords
   3. Spelling correction using TextBlob
   4. Stemming
   5. Lemmatization
   6. Store the clean result in a new column of the data frame
   7. Filter list of hashtags from each tweet
   8. Store hashtags in a recent column of the dataset

After applying these methods, we got the following cleaned text.

text_clean
rt robertbeadles yo enter win monarch tokens us stock market crash amp learn retweet watch video
srilanka surcharge fuel remove surcharge impose diesel petrol revoke effect midnight june say power energy transport minister m
net issuance increase fund fiscal program gt yield spike higher gt risk stock emfx correct lower gt feed come ycc gt stock new
rt bentboolean much amazons traffic serve fastly help us find run tool ip address httpstco
amd ryzen desktop cpus look great track launch madtweets stock cnbc amd
rt quanttrend reduce portfolio risk gold perfect tail hedge central bank balance sheet expansion amp large fiscal deficits amp
million sales expect spirit aerosystems hold inc spr quarter httpstcozoqbvspvsj stock
rt arjunkharpal apple cut price iphone range china uncommon move discount n
rt cdc us new case data day lag saw another encourage decline global new case data still flat
look dependable dividends read gt etf invest stock business news
rt pipstodollars earn amzn tsla msft aapl amd ba fb luv mmm ge aal up twtr pfe cbsh pep googl gild sbux ual
guy market stay till expect major major crash nifty nifty banknifty stock
future fly spirit air save blog stocksampsports fb bots amp ai also welcome blog salesforce crm
interest comparison market spx stock day much bear market level mark red line get level summer smoke clear weak something watch
canbk canbk nifty banknifty equity stock analysis trade sp httpstcoosgxfbjkqf
spot premium major price peak phase distribution fund stay largely positive amp futures trade premium constantly constant spot
chile road recovery aapl tsla fb
rt forecastcity latest eurnzd tradeideas amp technicalanalysis free pip profit days euro eur nzd money forex
vir since april post level follow thread level support amp resistance sl amp tp spx nasdaq technicalanalysis stockmarket daytra
epiphronr china population billion reason iq iqiyi amp gluu short squeeze bullish call especially w new sorcerers arena game ne
rt rafkadian tsla amp nio fot day trade stock adssgroup
eps estimate merck amp co inc mrk lower svb leerink
rt johrinikhil bank infuse confidence capital yesbank sell partial stake free lockin portion btw

Following are the extracted hashtags from each tweet into a new column of hashtags.

tags
<null>
#srilanka #lka #fuelprices #taxes #economy #stocks #stockmarket
#stocks #emfx #fed #ycc
<null>
#madtweets #stocks #cnbc #amd
<null>
#stocks
#apple
#cdc #who
#etf #investing #stocks #business #news
<null>
#nifty #banknifty #stocks
#blog #salesforce
<null>
#canbk #nifty #banknifty #equity #stocks #analysis #trading #sp
#btc #bitcoin #ta #sp500
<null>
#eurnzd #tradeideas #technicalanalysis #profit #euro #eur #nzd #money #forex…
#nasdaq #technicalanalysis #stockmarket #daytrading
<null>
#trading #stocks #raftrading111
<null>
#banks #yesbank
<null>

2. **Bag of Words model**

   After cleaning the text, we created a matrix of words on which we performed the analysis. We created unigrams, bigrams, and trigrams using tokenization and used TFIDF and frequency for word cloud and trending hashtags.

   1. Sparse Matrix

This is the sparse matrix we have created from the tweets. This represents the bag of words model. Each column is a unique word, and each row is a tweet. Each cell represents the TFIDF score of the particular word.

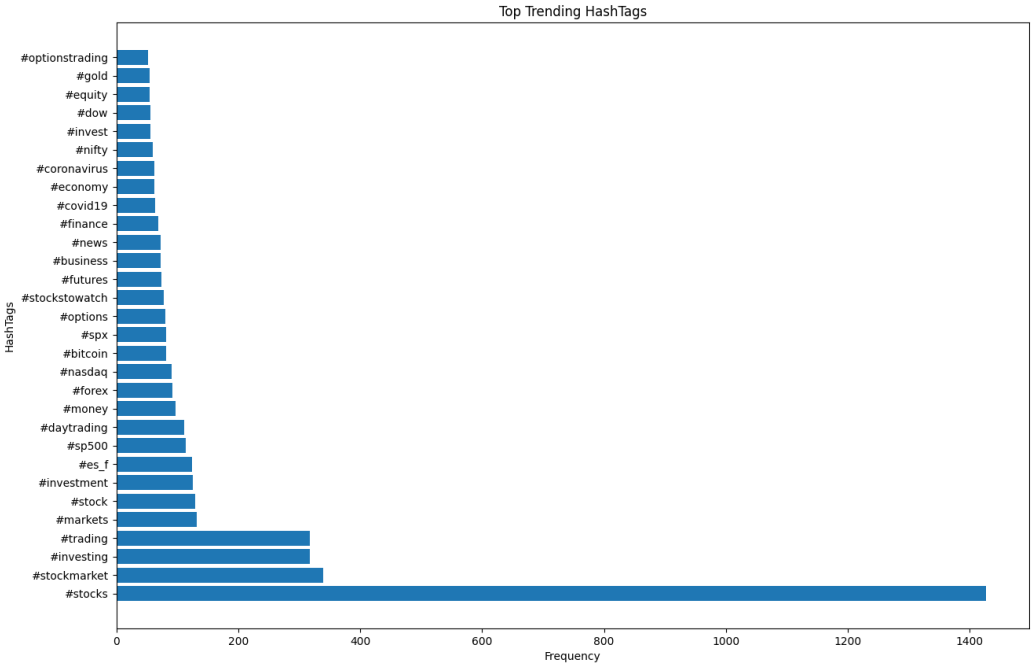
3. **Word cloud**

We tried to analyze the most frequently used words used by the people in tweets. We used WordCloud visualization for this task. Following is the illustration.

As can be seen, people extensively wrote about stocks. In stocks, they mainly showed interest in the stocks of RT, SPX, Amazon, Facebook, Microsoft, and other big tech companies.

4. **HashTag frequency**

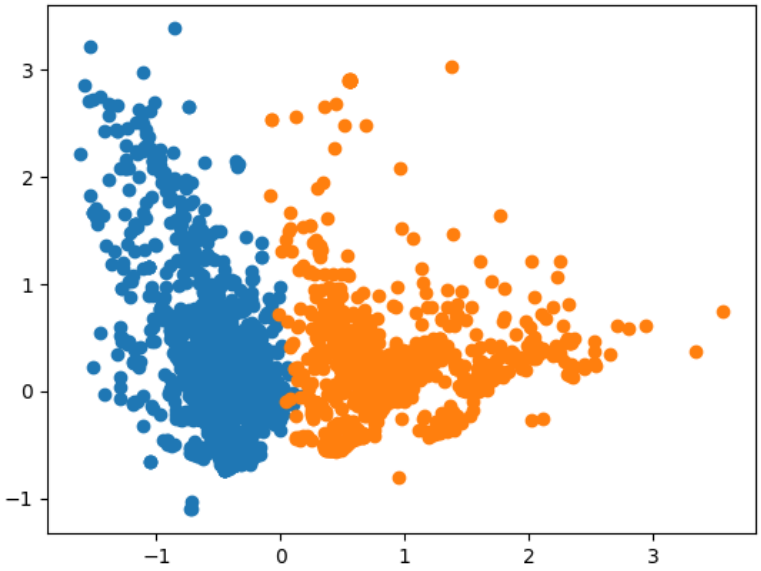
Hashtags carry actual weight in the tweet. They provide information about the trending topics. People generally use hashtags to stress an important issue, trend, or word. We analyzed the top hashtags used in the given period using a frequency bar chart.



This bar chart shows that stock is the most frequently used hashtag followed by stockmarket. The X-axis shows the frequency, and the y-axis displays the hashtags.

5. **Clustering**

After getting a general idea of the tweets trend, we were interested in finding any groups or clusters of tweets to understand the different tweets segments better. We used unsupervised learning to cluster the data. We used K-means clustering for this task and found that there are two main clusters with minimal overlapping. Following is the scatter plot of groups after applying PCA to reduce the data dimension to two dimensions.

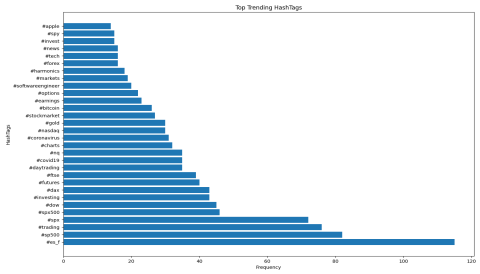


As can be seen, there are two different segments of tweets. There is a slight overlap between them, but they are mostly well separated.

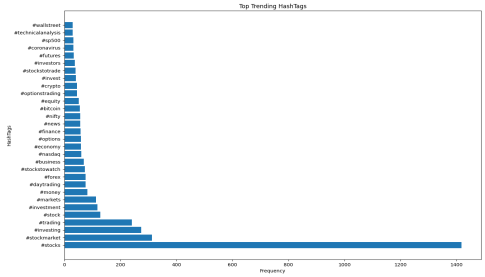
6. **Discovering information from clusters**

We used a word cloud and frequency bar chart to understand each cluster further. Following are the results of the analytics using visualization.

Cluster 1



Cluster 2



As can be seen, there is minimal overlapping. Overlapping is talking about the stocks. However, the first cluster mainly talks about companies, especially top tech companies in the united states. In contrast, the second cluster is only concerned with stocks, the stock market, trade, investment, and earnings.

7. **Association rule mining**

We further tried to understand the correlation among words. We were interested to know which things people were primarily interested in together. For example, if they tweet about stocks, do they always mention particular stock or a company? Such association can be found via association rule mining. We used the FP growth algorithm to extract the frequent itemsets and mine association rules in association rule mining. After mining, we found the following top 25 rules from the dataset.

	antecedents	consequents	support	confidence	lift
1	stockmarket, trade	stock	0.0214	0.981651376146789	2.7451101122673074
2	stockmarket, invest	stock	0.0302	0.9805194805194806	2.741944856038816
3	amazon, microsoft, facebook	apple	0.02	0.9174311926605505	6.826124945391001
4	stockmarket	stock	0.0622	0.9174041297935103	2.5654477902503086
5	inc	stock	0.0268	0.8933333333333334	2.4981357196122302
6	qqq, apple	spy	0.021	0.875	6.782945736434108
7	investment	stock	0.0262	0.8733333333333334	2.4422073079791202
8	amazon, microsoft	apple	0.027	0.8653846153846154	6.438873626373627
9	apple, microsoft, facebook	amazon	0.02	0.8620689655172414	7.483237547892721
10	microsoft, facebook	apple	0.0232	0.8592592592592592	6.393298059964726
11	amazon, facebook	apple	0.0302	0.8483146067415731	6.3118646334938475
12	spy, facebook	apple	0.0218	0.8384615384615385	6.238553113553114
13	qqq, spx	spy	0.0224	0.835820895522388	6.479231748235566
14	microsoft, facebook	amazon	0.0218	0.8074074074074075	7.008744855967079
15	invest	stock	0.0624	0.8041237113402061	2.2486680965889434
16	qqq	spy	0.0394	0.8008130081300813	6.20785277620218
17	tesla, facebook	apple	0.0232	0.7891156462585034	5.871396177518627
18	apple, facebook	amazon	0.0302	0.7704081632653061	6.6875708616780045
19	amazon, spy	apple	0.0252	0.7544910179640719	5.613772455089821
20	apple, microsoft	amazon	0.027	0.75	6.510416666666667
21	microsoft, facebook	amazon, apple	0.02	0.7407407407407408	15.561780267662622
22	amazon, apple, microsoft	facebook	0.02	0.7407407407407408	7.558578987150416
23	amazon, spy	tesla	0.0242	0.7245508982035929	10.176276660162822
24	tesla, facebook	amazon	0.0212	0.7210884353741497	6.259448223733939

The above diagram shows the top 25 association rules mined from the dataset. Each association rule has its support, confidence, and lift value. Analyzing this data implies that Microsoft, Facebook, and Amazon are correlated. Whenever users mentioned Microsoft and Facebook in the tweet, they said amazon. Similarly, Facebook was mentioned 72% of the time with amazon, apple, and Microsoft.

Summary

From the raw text to association rule mining, we analyzed all the tweets and presented the extracted information. In summary, tweets and hashtags are a great way to understand people's thoughts and trendy topics. Using tweet analytics, we analyzed the tweets from April 2021 to July 2021. We found the stock market dramatically influences people. They showed significant interest in the stocks, stock market, trading, and investment. In particular, top tech companies, such as Facebook, Amazon, Microsoft, Apple, Google, Tesla, and Netflix, are hot stocks in the stock market. Many tweets were inclined toward these companies.