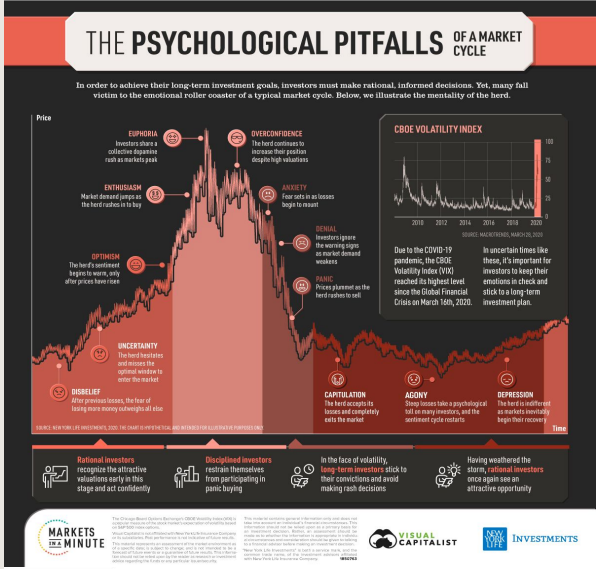
DSI
Capstone Project:
Investor Sentiment
&
S&P 500

Benjamin Lee



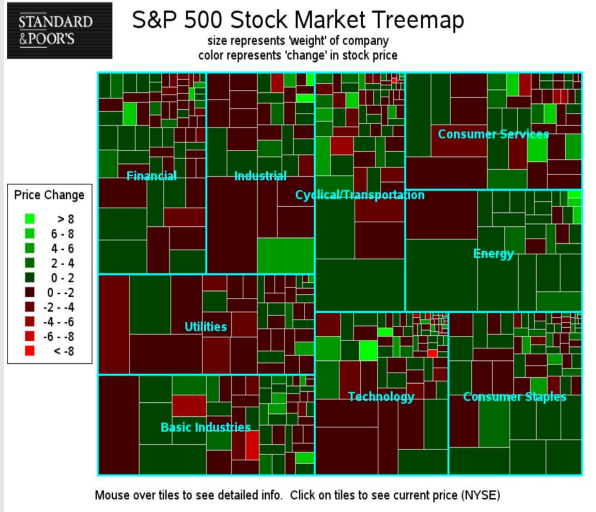


TABLE OF CONTENTS

Introduction

Introduction

Introduction (Reddit & S&P500)

Reddit

- American social news aggregation, web content rating, and discussion website
- Members submit content to the site
- Posts are organized by subject into user-created boards called "subreddits"

S&P 500

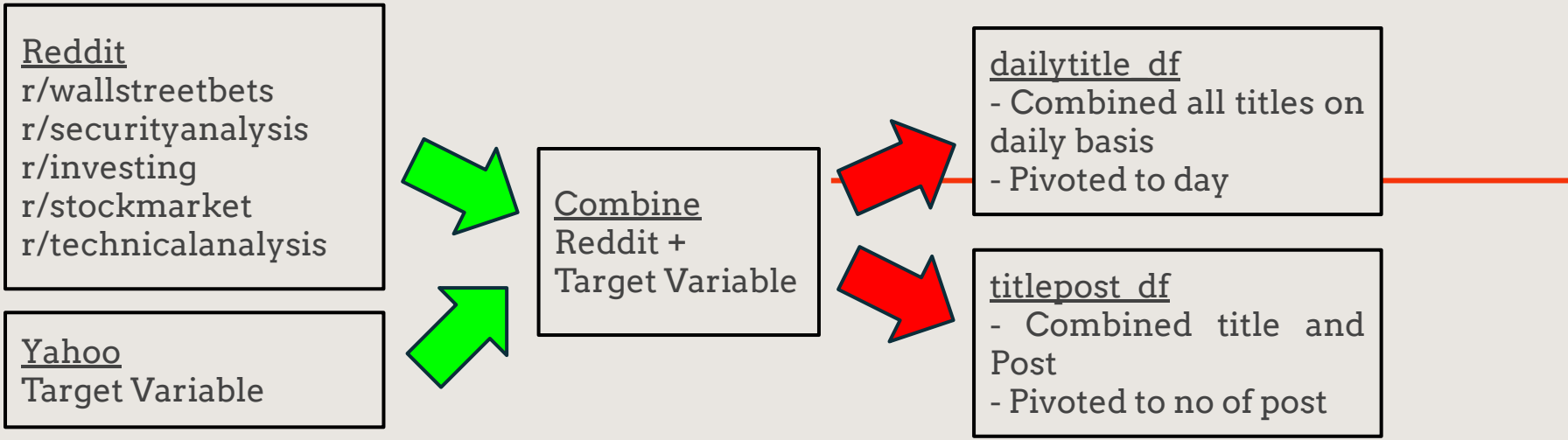
- Stock market index that measures the stock performance of 500 large companies listed on stock exchanges in the United States
- One of the most commonly followed equity indices
- The 10 largest companies
  - Apple Inc.
  - Microsoft
  - Amazon.com
  - Facebook
  - Tesla, Inc.
  - Alphabet Inc. (class A & C)
  - Berkshire Hathaway
  - Johnson & Johnson
  - JPMorgan Chase & Co

Introduction (Problem Statement)

We, a consultation firm (Data Insights Pte Ltd), was recently engaged by Investors Club to predict whether if the **S&P 500** will be **Bullish or Bearish** from a particular **post from reddit** through their sentiment. A classification model will be devised and evaluated based on accuracy score.

This will help the Investors Club better understand if the trend is based on sentiments of the general public or not.

Introduction (Dataset)



Data Collection

Scraping Scraping Scraping... ...

```
In [54]:  1  df_final.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 508254 entries, 0 to 991
Data columns (total 8 columns):
 #   Column      Non-Null Count   Dtype  
---  ------      --------------   -----  
 0   date        508254 non-null  object 
 1   title       508254 non-null  object 
 2   selftext    349645 non-null  object 
 3   is_self     508254 non-null  float64
 4   upvotes     508254 non-null  float64
 5   n_comments  508254 non-null  float64
 6   permalink   508254 non-null  object 
 7   author      508254 non-null  object 
dtypes: float64(3), object(5)
memory usage: 34.9+ MB
```

```
1  sp500_df['Percent_Change_Class'].value_counts()
```

```
up      294
down    210
n/a       1
Name: Percent_Change_Class, dtype: int64
```

Preprocessing + Feature Engineering

Preprocessing
Stopwords()
TweetTokenizer()
WordNetLemmatizer()

Feature Engineering (Length)
- Title_len
- Post_len
- TitlePost_len

Feature Engineering (titlepost)
- fresh_meat
(Initial preprocessing)
- lean_meat
(Updated preprocessing)

Feature Engineering (Length)
- Title_len

Feature Engineering (dailytitle)
- Title on daily basis

Modelling

Modelling & Benchmark

```
1  dailytitle_df['percent_change_class'].value_counts(normalize=True)
```

```
1    0.583665
0    0.416335
Name: percent_change_class, dtype: float64
```

Vec
Count Vectorizer
Tfid Vectorizer

Mod
Logistic Regression
Multinomial NB
Random Forest
Ada Boost

Model Insights on dailytitle_df

```
1  dailytitle_df['percent_change_class'].value_counts(normalize=True)
```

```
1    0.583665
0    0.416335
Name: percent_change_class, dtype: float64
```

	model	vectorizer	train	test	roc	precision	recall	f_score
0	lr	cvec	0.779841	0.488	0.422287	0.560000	0.575342	0.567568
1	lr	tvec	0.710875	0.568	0.507640	0.584071	0.904110	0.709677
2	nb	cvec	0.665782	0.504	0.448630	0.563218	0.671233	0.612500
3	nb	tvec	0.633952	0.600	0.516860	0.593496	1.000000	0.744898
4	rf	cvec	0.583554	0.584	0.500000	0.584000	1.000000	0.737374
5	rf	tvec	0.583554	0.584	0.500000	0.584000	1.000000	0.737374
6	ada	cvec	0.777188	0.560	0.531744	0.588235	0.821918	0.685714
7	ada	tvec	0.854111	0.528	0.502107	0.574468	0.739726	0.646707

Model Insights on titlepost_df

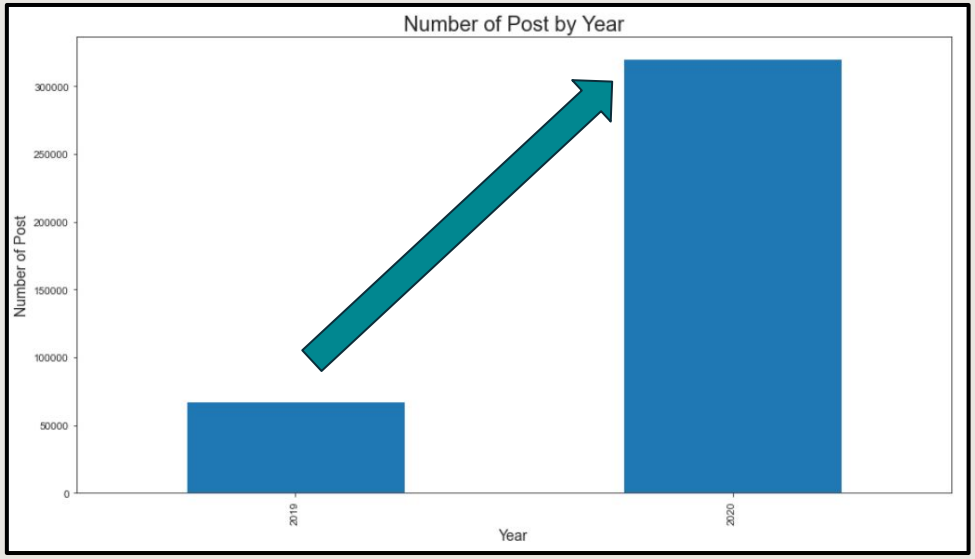
```
1  titlepost_df['percent_change_class'].value_counts(normalize=True)
```

```
1    0.592558
0    0.407442
Name: percent_change_class, dtype: float64
```

	model	vectorizer	train	test	roc	precision	recall	f_score
0	lr	cvec	0.779841	0.488	0.422287	0.560000	0.575342	0.567568
1	lr	tvec	0.710875	0.568	0.507640	0.584071	0.904110	0.709677
2	nb	cvec	0.665782	0.504	0.448630	0.563218	0.671233	0.612500
3	nb	tvec	0.633952	0.600	0.516860	0.593496	1.000000	0.744898
4	rf	cvec	0.583554	0.584	0.500000	0.584000	1.000000	0.737374
5	rf	tvec	0.583554	0.584	0.500000	0.584000	1.000000	0.737374
6	ada	cvec	0.777188	0.560	0.531744	0.588235	0.821918	0.685714
7	ada	tvec	0.854111	0.528	0.502107	0.574468	0.739726	0.646707

EDA

Surge in Post due to Covid




```
print(dailytitle_df_2019['title_cleaned'].str.contains('covid',
print()
print(dailytitle_df_2020['title_cleaned'].str.contains('covid',
```

```
False    251
Name: title_cleaned, dtype: int64

True     219
False     32
Name: title_cleaned, dtype: int64
```

```
print(dailytitle_df_2019['title_cleaned'].str.contains('corona',
print()
print(dailytitle_df_2020['title_cleaned'].str.contains('corona',
```

```
False    249
True       2
Name: title_cleaned, dtype: int64

True     211
False     40
Name: title_cleaned, dtype: int64
```

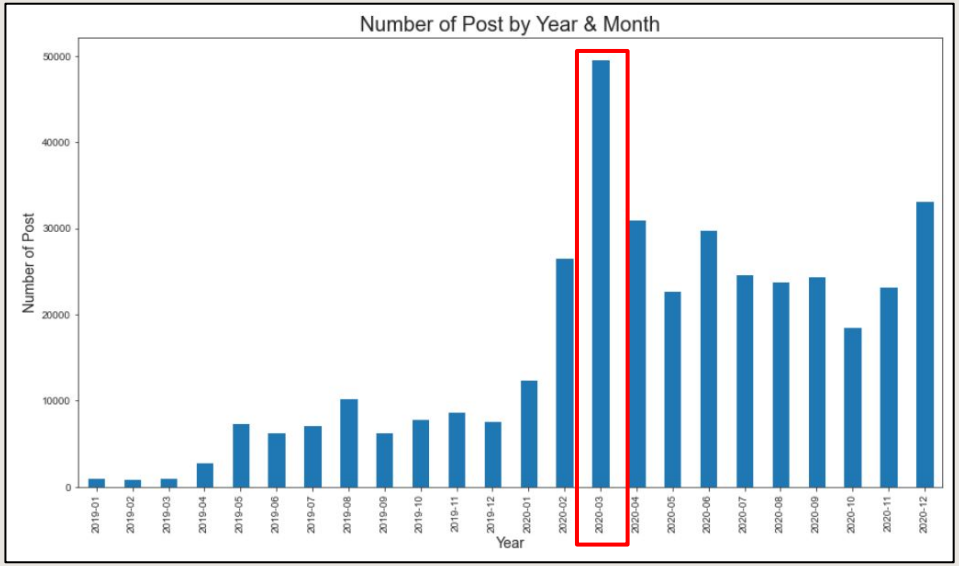
```
print(dailytitle_df_2019['title_cleaned'].str.contains('correction',
print()
print(dailytitle_df_2020['title_cleaned'].str.contains('correction',
```

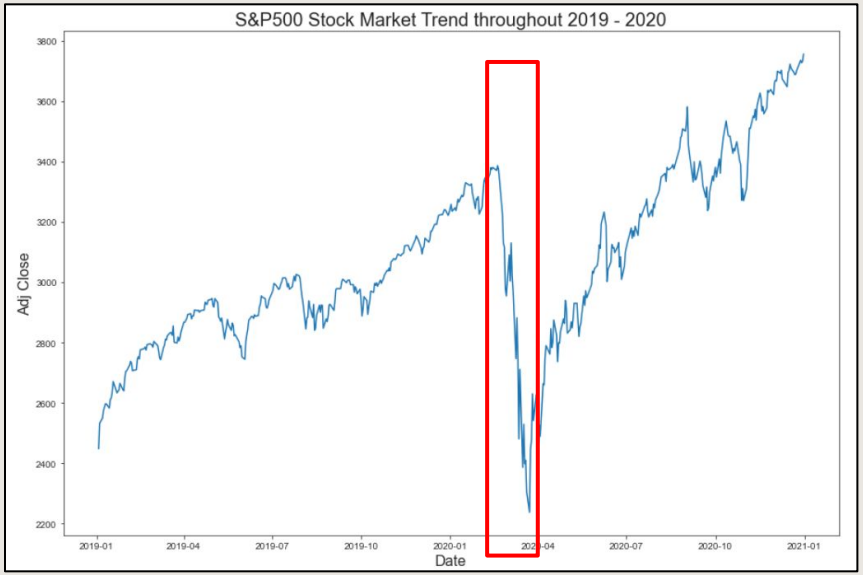
```
False    204
True      47
Name: title_cleaned, dtype: int64

False    130
True     121
Name: title_cleaned, dtype: int64
```

```
print(dailytitle_df_2019['title_cleaned'].str.contains('market crash',
print()
print(dailytitle_df_2020['title_cleaned'].str.contains('market crash',
```

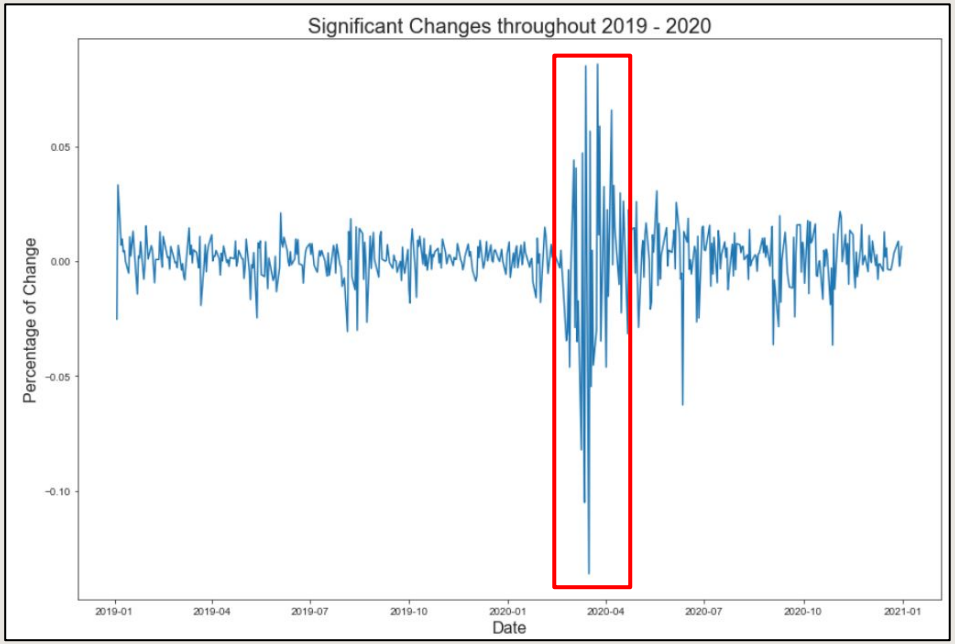
```
False    194
True      57
Name: title_cleaned, dtype: int64

True     200
False     51
Name: title_cleaned, dtype: int64
```

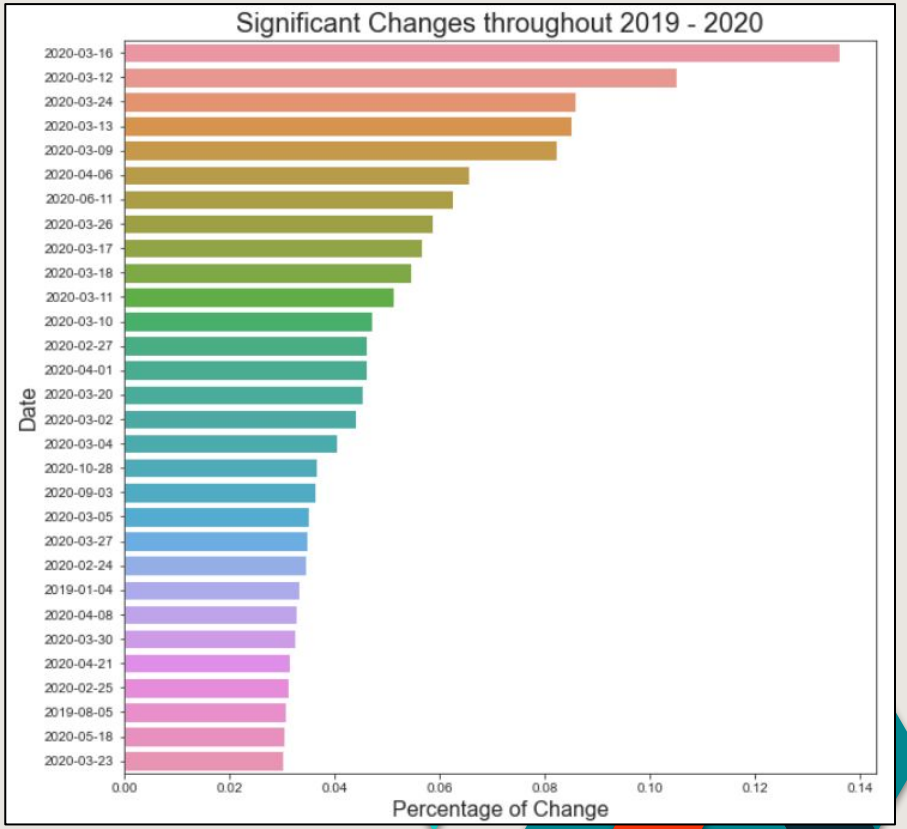
Huge Correction in US Stock Market



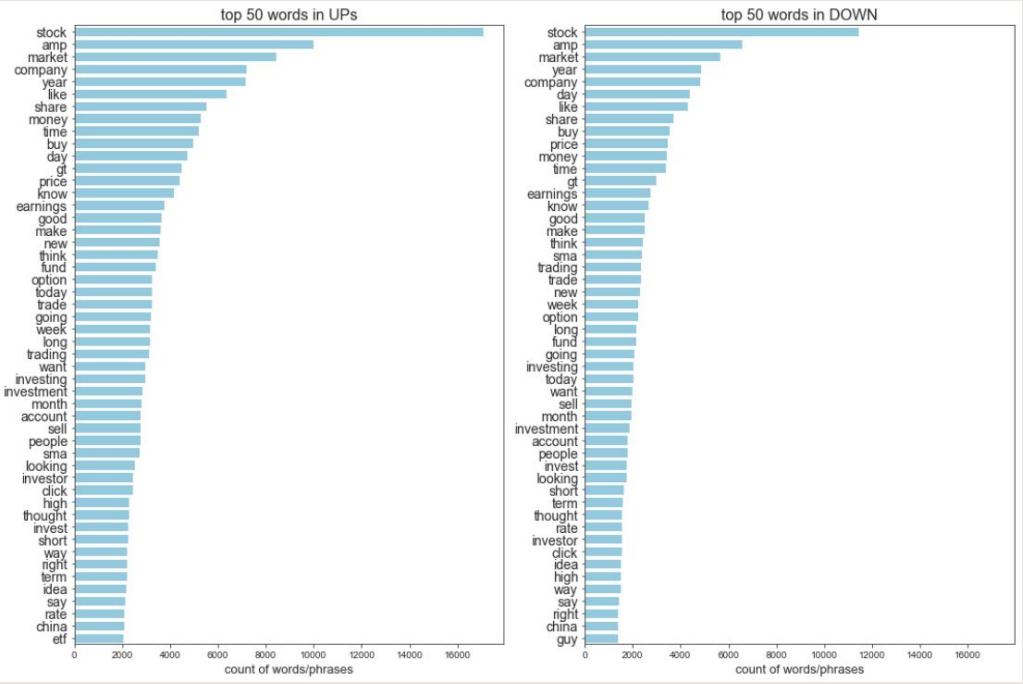


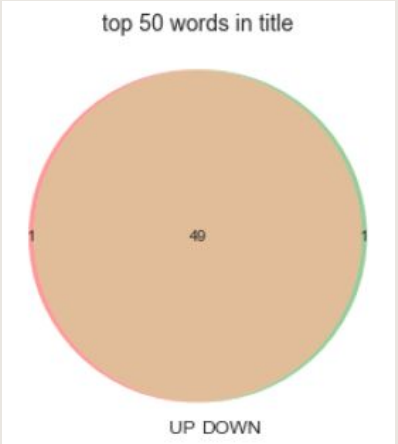
Significant Changes





Preprocessing (Initial Cleaning - Unigram)





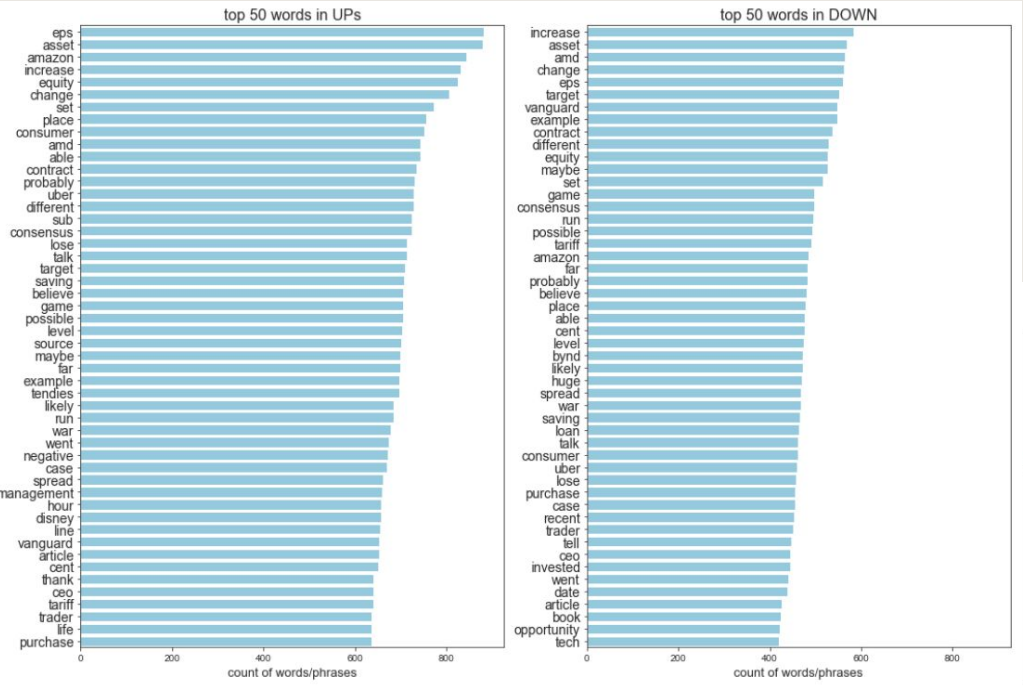
UP Trend

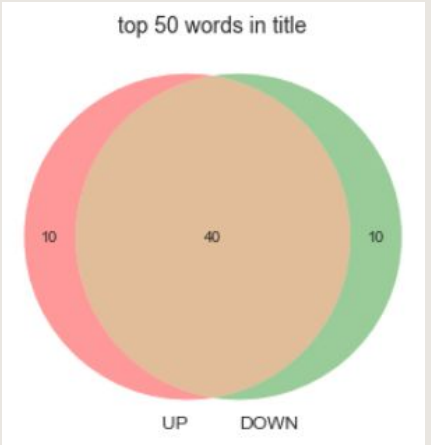
DOWN Trend

Preprocessing (More Cleaning - Stopwords Update)

```
stops.update(['notext', 'sma', 'year', 'month', 'day', 'gt', 'click', 'long', 'term', 'stock', 'market', 'way', \
              'global', 'idea', 'ty', 'amp', 'ta', 'cap', 'st', 'live', 'pre', 'company', 'like', 'time', \
              'price', 'earnings', 'removed', 'option', 'trading', 'trade', 'today', 'want', 'sell', 'thought', \
              'investing', 'invest', 'money', 'good', 'new', 'week', 'think', 'going', 'account', 'option', \
              'stocks', 'know', 'make', 'shares', 'spy', 'million', 'buying', 'fund', 'revenue', ' short', \
              'investement', 'looking', 'buy', 'investment', 'people', 'share', 'china', 'guy', 'right', 'curr', \
              'key', 'avg', 'fn', 'tn', 'tp', 'bb', 'rsi', 'hang', 'seng', 'chart', 'sch', 'stofu', 'yr', 'feel', \
              'morning', 'coo', 'elon', 'mask', 'wall', 'street', 'stoxx', 'usd', 'calendar', 'kong', 'united', \
              'macd', 'quote', 'thanks', 'advance', 'information', 'technology', 'tl', 'dr', 'dae', 'growth', 'news', \
              'investor', 'dividend', 'billion', 'roth', 'ira', 'etf', 'question', 'index', 'hong', 'state', 'fy', 'fp',
              'rate', 'cut', 'goldman', 'sachs', 'economic', 'appreciated', 'reached', 'sector', 'dt', 'tgt', 'warren',
              'real', 'estate', 'basis', 'say', 'discretionary', 'yahoo', 'finance', 'morgan', 'stanley', 'south', \
              'korea', 'oversold', 'according', 'com', 'briefing', 'president', 'captial', 'student', 'buffet', \
              'td', 'ameritrade', 'yield', 'charles', 'schwab', 'quarter', 'reported', 'high', 'short', 'portfolio', \
              'robinhood', 'donald', 'trump', 'bln', 'federal', 'reserve', 'bank', 'currently', 'robin', 'hood',  \
              'cash', 'changed', 'webp', 'dow', 'jones', 'dir', 'wti', 'capital', 'large', 'hit', 'beat', 'map', \
              'credit', 'card', 'expected', 'neutral', 'rated', 'exp', 'dax', 'glf', 'finished', 'cac', 'euro', \
              'format', 'png', 'auto', 'nikkei', 'yesterday', 'technical', 'shanghai','need', 'said', 'thing', 'work', \
              'balance', 'sheet', 'estimate', 'interactive', 'broker', 'overbought', 'low', 'bond', 'future', 'risk', \
              'position', 'advice', 'little', 'strategy', 'read', 'higher', 'apple', 'symbol', 'ago', 'tesla', 'pay', \
              'holding', 'average', 'cnbc', 'margin', 'th', 'product', 'small', 'data', 'open', 'total', 'app', 'world',
              'profit', 'look', 'start', 'gain', ' future', 'lot', 'loss', 'let', 'free', 'post', \
              'big', 'business', 'return', 'sale', 'tax', 'shit', 'financial', 'report', 'best', 'bought', \
              'ratio', 'closing', 'closed', 'value', 'really', 'help', 'hold', 'come', 'end', 'got', 'play', \
              'service', 'dollar', 'better', 'worth', 'trying', 'recession', 'thinking', 'gold', 'getting', 'debt', \
              'mini', 'nasdaq', 'oz', 'link', 'afternoon', 'ipo', 'bbl', 'insider', 'lyft', 'ba', 'yes', 'health', 'care
              'profit', 'use', 'point', 'deal', 'cost', 'le', ' selling', 'recently', 'understand', \
              'japanese', 'release', 'nd', 'mean', 'number', 'tomorrow', 'wondering', 'actually', 'drop', 'platform',\
              'current', 'selling', 'plan', 'sure', 'great', 'world', 'industry', 'fed', 'using', 'income', 'potential',
              'signal', 'result', 'txn', 'io', 'past', 'close', 'making', 'order', 'bad', 'fuck', 'analyst', 'opinion',\
              'fee', 'based', 'pretty', 'started', 'dd', 'bit', 'monday', 'tuesday', 'wednesday', ' thursday', 'friday',
              'sold', 'perf', 'economy', 'job', 'lower', 'coming', 'research', 'hi', 'fucking', 'chinese', 'google', 're
              'old'])
```

Preprocessing (More Cleaning - Unigram)

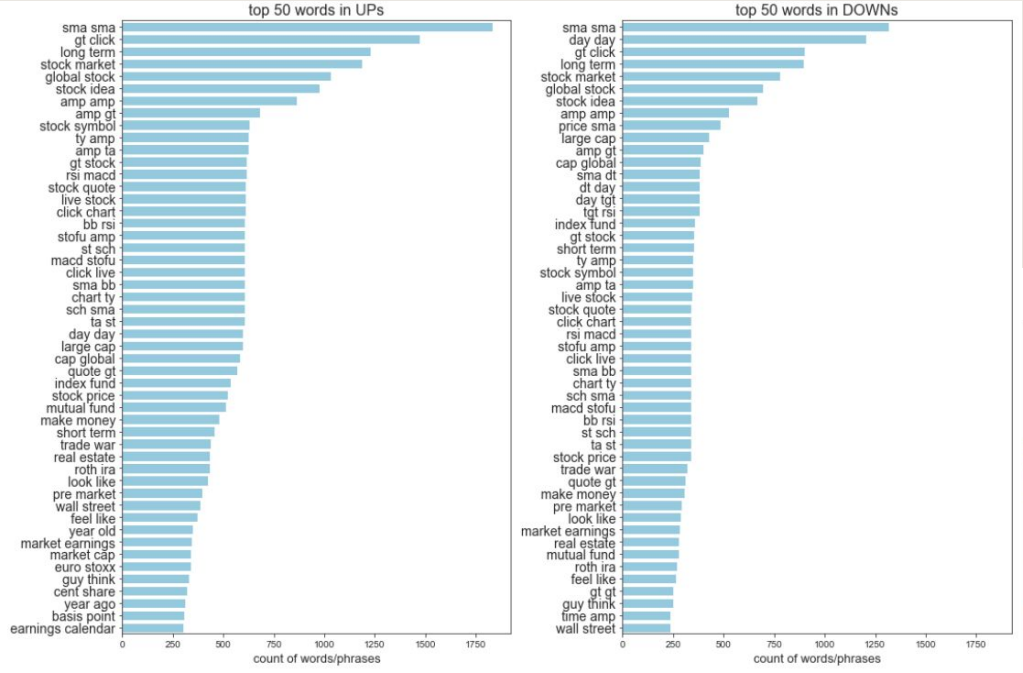


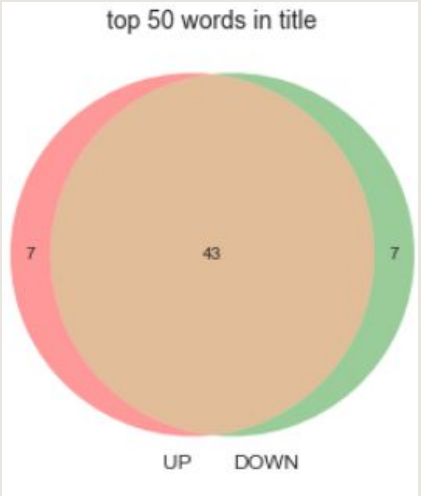


UP Trend

DOWN Trend

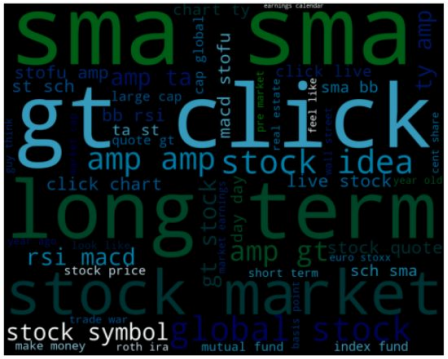
Preprocessing (Initial Cleaning - Bigram)



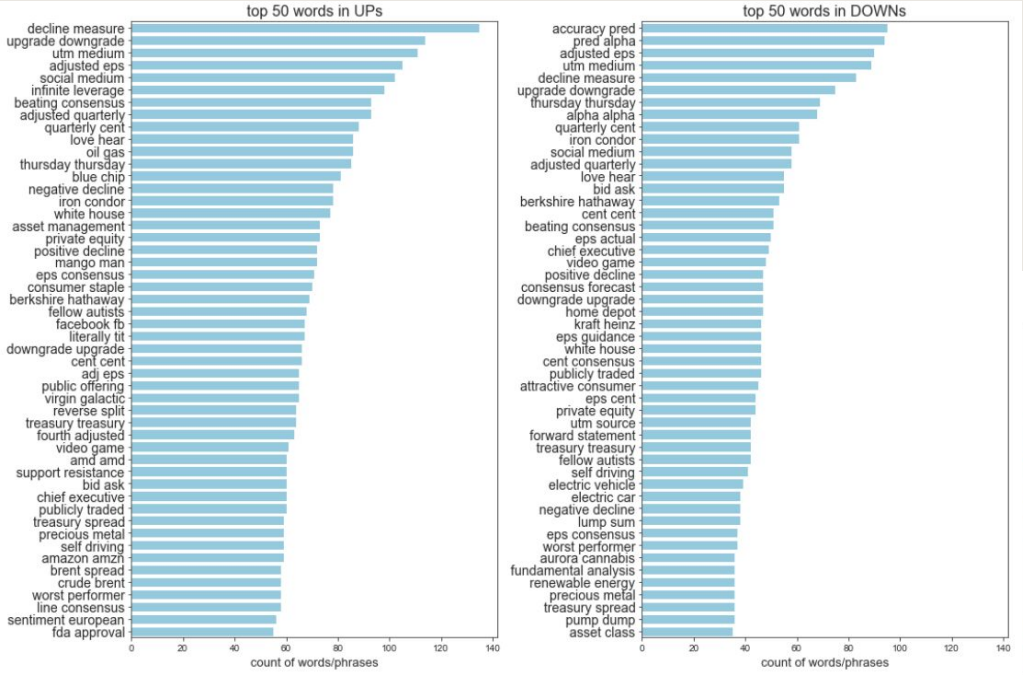


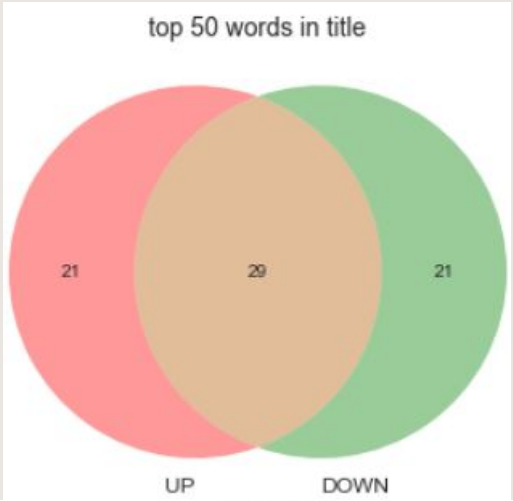
UP Trend

DOWN Trend

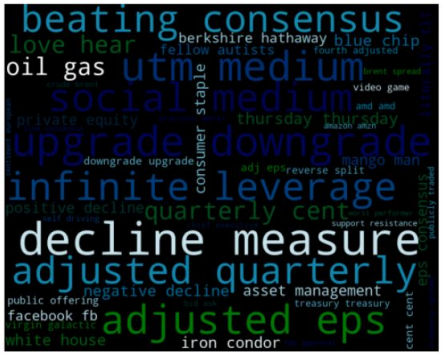


Preprocessing (More Cleaning - Bigram)

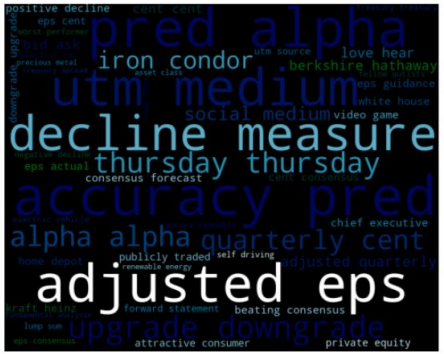




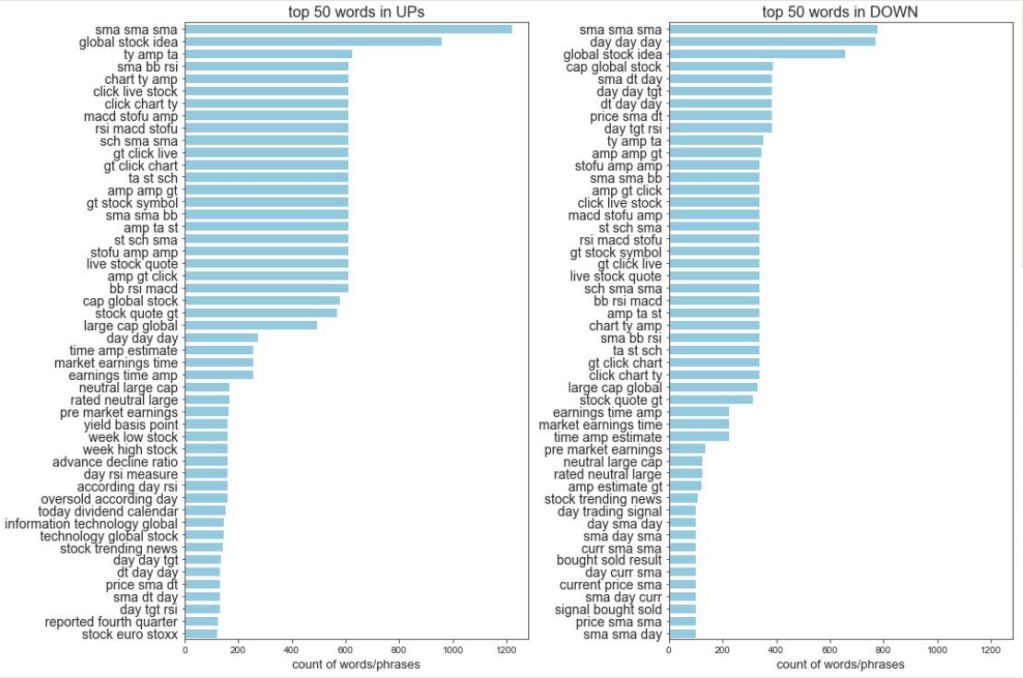
UP Trend

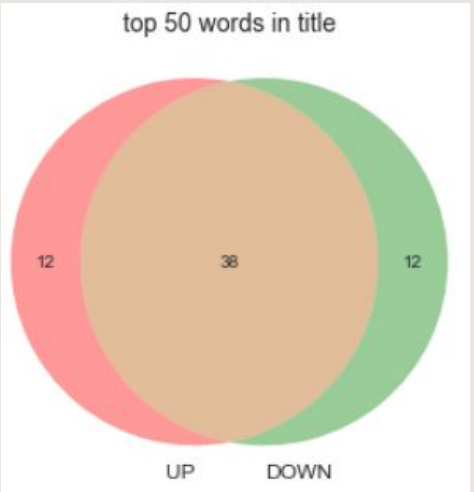


DOWN Trend



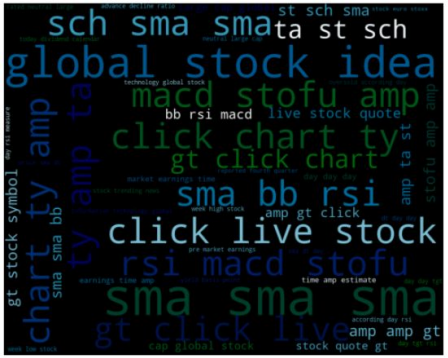
Preprocessing (Initial Cleaning - Trigram)

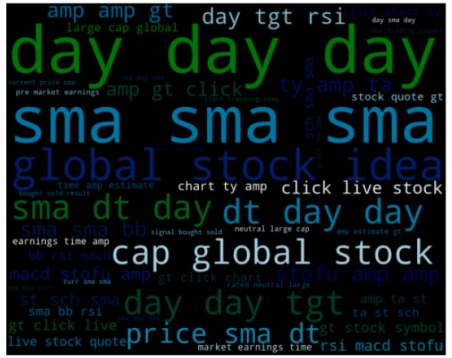




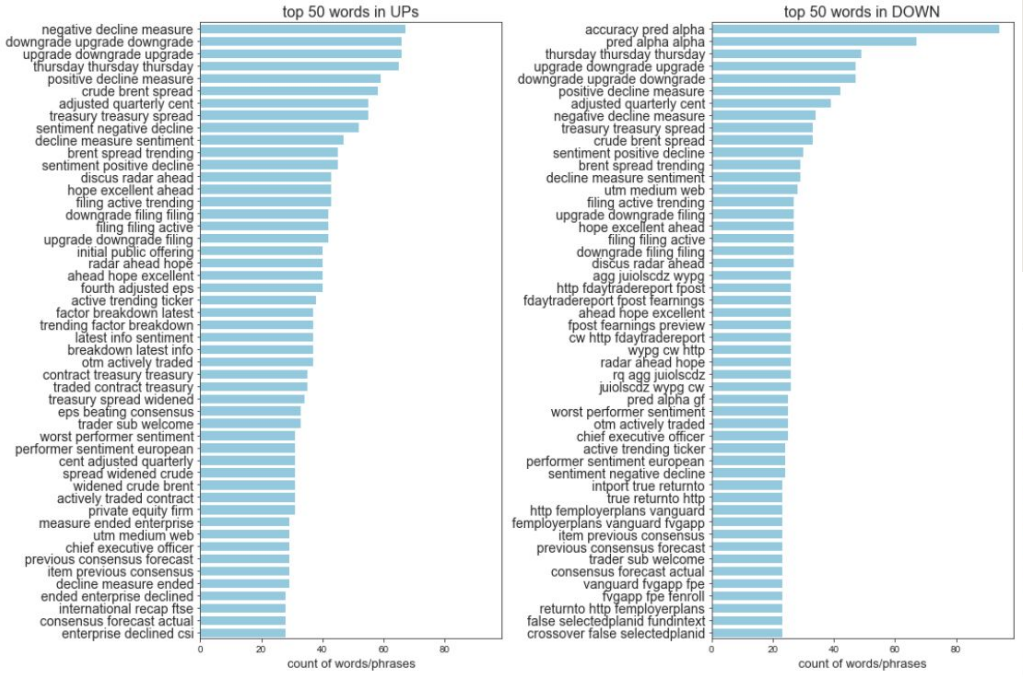
UP Trend

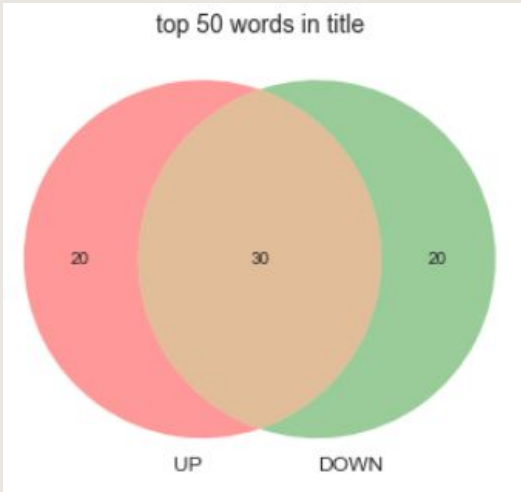
DOWN Trend



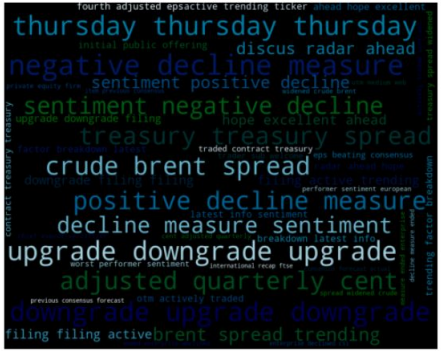


Preprocessing (More Cleaning - Trigram)

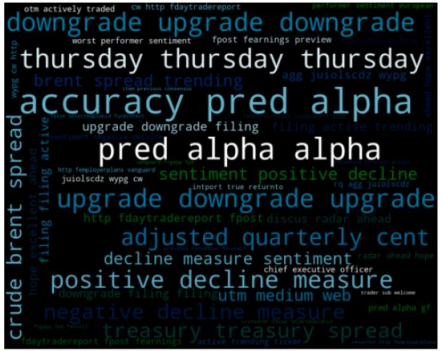




UP Trend



DOWN Trend



NLTK - Vader

Most Positive

	title_cleaned	percent_change_class
24	first researched pick pcty well first well res...	up
33	newb question tell stock overbought thought go...	up
45	long term let play game thought hain china gdp...	up
36	general electric stock first robbin hood margi...	down
65	ar augmented reality daily profit get weed fol...	down

Most Negative

	title_cleaned	percent_change_class
255	hidden fee charles schwab thinking getting tod...	up
292	robinhood much robinhood make lose morning tes...	up
315	exactly msft microsof affected purchase profit...	up
367	sma line moving average line someone explain h...	down
309	short long term tax question newer investor ta...	down

Conclusion

Conclusion

- Reddit community cannot determine whether the stock market will be bullish or bearish the next day

- Sentiments are still mixed regardless of the performance of the stock market
  - (invest.Buy_Low_Sell_High) | (invest.FOMO)

- Recommendation:
  - Data from official news outlets
  - Data from people of influence

- Importance of EDA!!!

THANKS!