MASTER THESIS REPORT

Twitter Sentimental Analyis

Introduction

Sentiment analysis of 5000 tweets related to the keyword "bitcoin" is represented in the report.

Analysis Performed:
1) Word Cloud
2) Overall Sentiment Scores and Popularity of Tweets and Hashtags
3) Sentiment Scores by Sentiments Types
4) Sentiments Scores by Hashtags
5) Distribution of Followers, Favorites, Retweet and Friends count
6) Correlation Analysis
7) Sentiments Score Trend Analysis by Time, Retweets and Favorites count
8) Predictive Analysis for Senntiments Scores by Retweet and Favorites counts

Dataset

The dataset was created by compiling data around 5000 tweets which is sent in twitter between 10:30 am and 12:00 pm on 22.Feb.2022. There are 4223 rows after the data cleaning is performed and 14 columns in the dataset. The details about the dataset is described below.

Data fields: The twitter dataset is maintained using JSON which contains several data fields for fetching it for the analysis, in this implementation the below data fields are considered from the main data and appended on this application with custom names.

'tweet': The full text of the tweet from twitter data
'date': The tweet.created_at from twitter data
'author': The user.screen_count from twitter data
'hashtags': The text from twitter JSON data of entities ["entities"] and ["hashtags"]
'followers_count': The user.followers_count from twitter data
'friends_count': The user.friends_count from twitter data
'retweet_count': The retweet_count from twitter data
'favorite_count': The favorite_count from twitter data
'coordinates': Represents the geographic location

Technologies used

Tools: Jupyter Notebook, Apache Spark, Amazon Comprehend
Programming Language: Python
Libraries: Tweepy, ipywidgets, tqdm, numpy, pandas, re, nltk, pyspark, emoji, Boto3 SDK, Matplotlib, Plotly, Seaborn

Results:

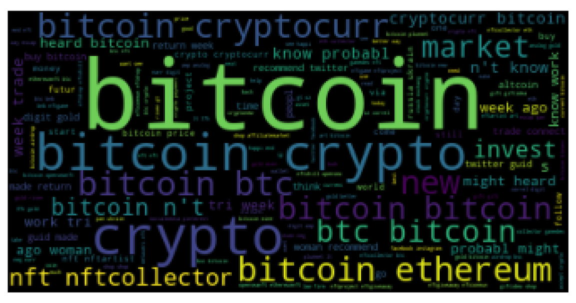
1. Word Cloud

Methods and Objectives

Word Cloud library is used for visual representation the text data. Matplotlib is used to plot the graphs.

Objective: To represent the quick visual of the frequency of words from the tweets and hashtags based on the given input keyword.

1.1. Wordcloud of tweets



1.2. Wordcloud of hashtags

Insights

A quick perception is found from the visual which reflects the frequency of the words used in this data. This helps to depict an idea on which topics the users are more interested.

2. Overall Sentiment Scores and Popularity of Tweets and Hashtags

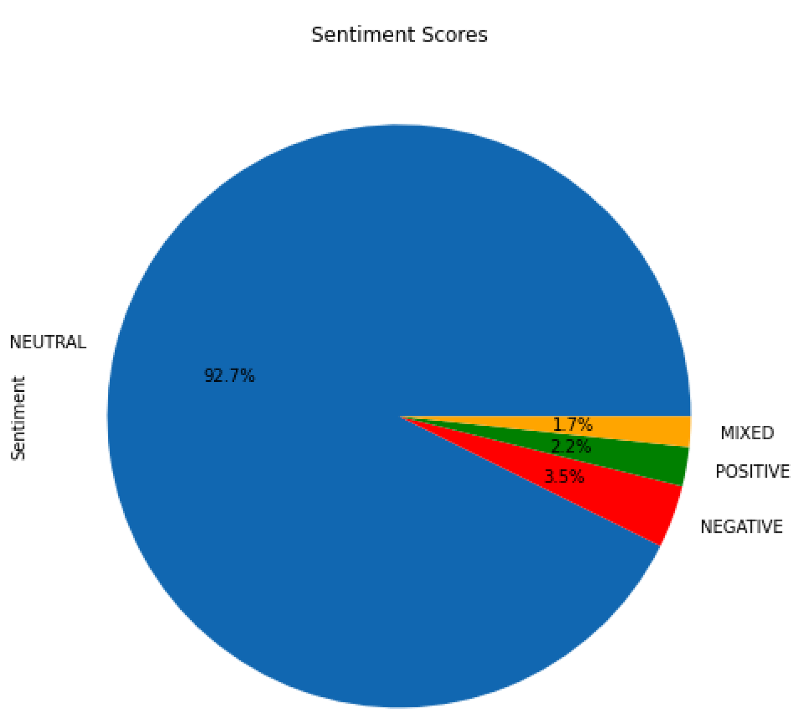
Methods and Objectives

Pandas and matplotlib libraries are used in this section.
Pandas is used to reshape the datasets and .value_counts function is used to count the number of sentiments, hashtags and tweets.
Matplotlib is used to plot the graphs.

Objective: To find the overall sentiment scores and analyze the popular tweets and hashtags.

2.1. Overall Sentiment Score

The below pie chart displays the overall sentiments scores by percentage
2.2% - Positive
3.5% - Negative sentiment
92.7% - Neutral sentiment is the most popular
1.7% - Mixed

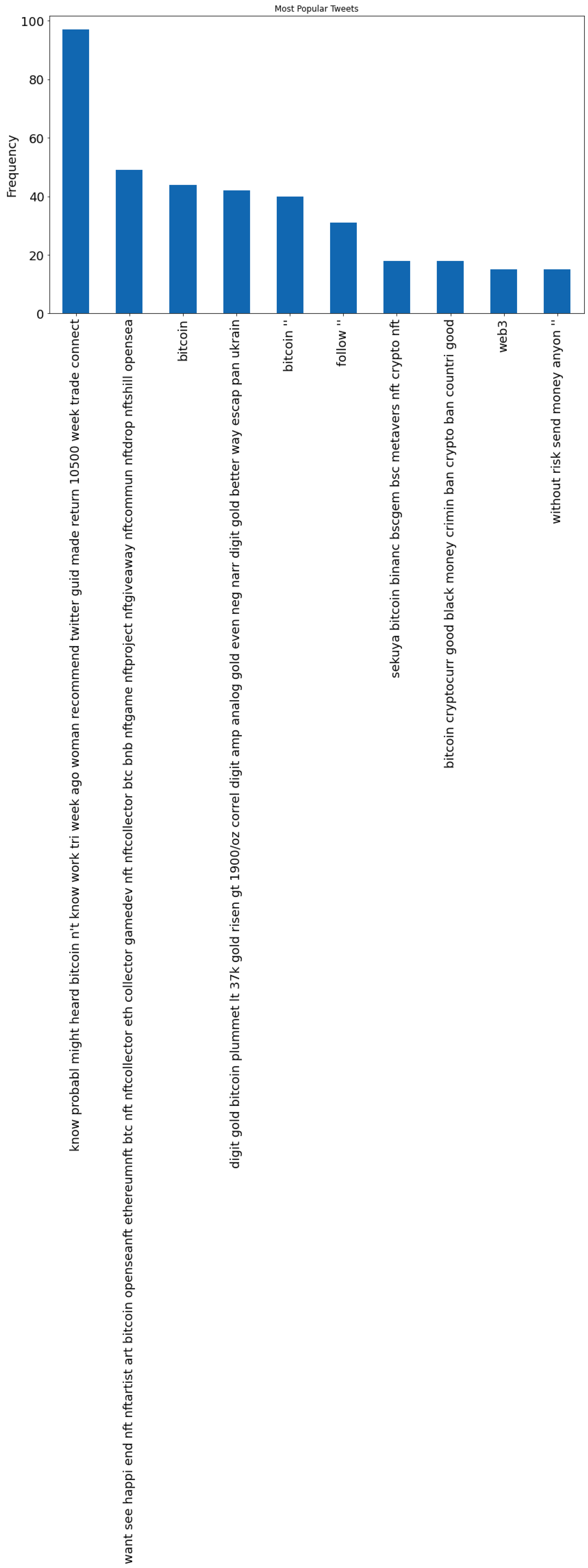


Insights

Based on the above chart with the sentiment types and scores, it is predicted that the given input keyword "bitcoin" has a greater number of neutral sentiments comparing other sentiments. This gives us an understanding that the user's opinion is more unbiased hence the data driven decision related to the bitcoin is not conclusive either in positive or negative.

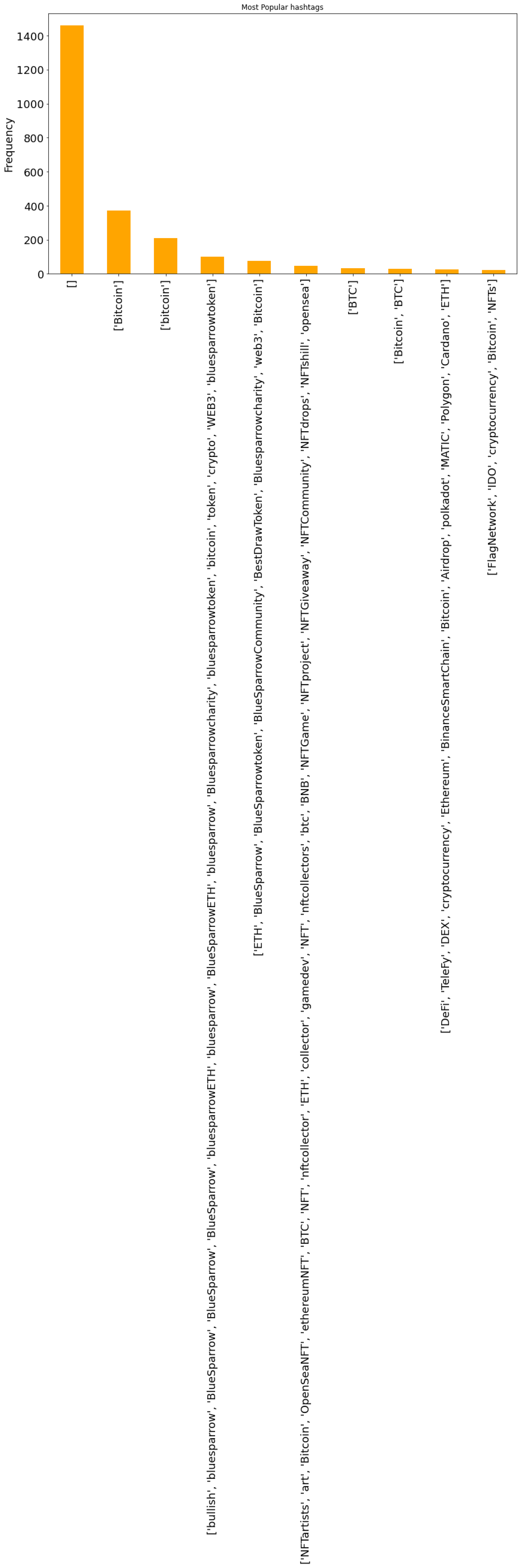
2.2. Popular Tweets

```
know probabl might heard bitcoin n't know work tri week ago woman recommend twitter guid made return 10500 week trade connect                    97
want see happi end nft nftartist art bitcoin openseanft ethereumnft btc nft nftcollector eth collector gamedev nft nftcollector btc bnb nftgame nftproject nftgiveaway nftcommun nftdrop nftshill opensea    49
bitcoin                                                                                                                                          44
digit gold bitcoin plummet lt 37k gold risen gt 1900/oz correl digit amp analog gold even neg narr digit gold better way escap pan ukrain         42
bitcoin ''                                                                                                                                       40
                                                                                                                                                 ..
attent everyon mani peopl n't know space invest bitcoin buy car hous land property.with minimum invest plan 100_ 1'000_ 10,000 car hous land owner inbox    1
linkedin shoppingtim twitter facebook instagram bitcoin socialmedia pinterest medium dogecoin gift gift giftidea shop shop affiliatemarket ad drizli deliv drink beer wine liquor deliv 60 minut    1
crypto bitcoin invest nft ethereum ''                                                                                                             1
best platform concept great futur good team think project success ''                                                                              1
btc famou feed ... bitcoin price plung tension russia ukrain climb cnn                                                                            1
Name: tweet, Length: 3202, dtype: int64
```



2.3. Popular Hashtags

The below plot is displaying the top 10 hashtags. "Bitcoin" is the most popular hashtag.



Insights

The topmost tweets and hashtags are relevant with the given keyword "bitcoin". The top 10 tweets show the familiarity of input keyword and its influences. The top 10 hashtags show us the relevant topics related to bitcoin which the users are tweeting.

3. Sentiment Scores by Sentiment Types

Methods and Objectives

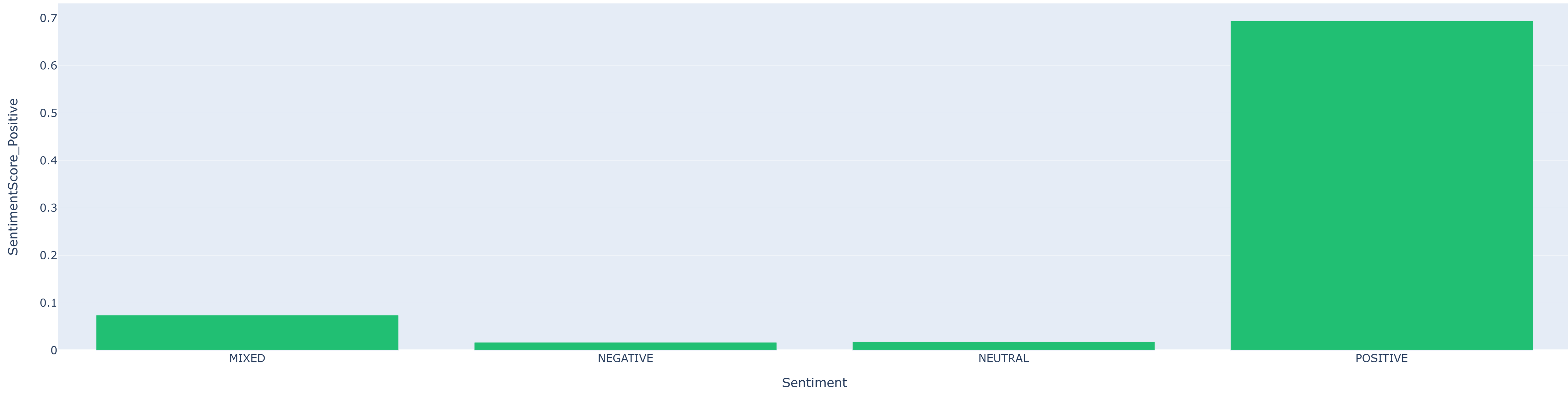
The Pandas and Plotly libraries were used to build this section.
The data is moulded with Pandas, and the average sentiment scores for each sentiment are calculated using the.mean() method.
Plotly was used to make the bar graph. Sentiment mean value is represented on the X-axis, and sentiment score type is plotted on the y-axis.

Objective: To get the impact of each sentiment with its own sentiment types positive, negative, neutral, and mixed. It is visualized by calculating the mean value of each sentiment score with its own types.

3.1. Positive

	Sentiment	SentimentScore_Positive
0	MIXED	0.074293
1	NEGATIVE	0.016968

	Sentiment	SentimentScore_Positive
2	NEUTRAL	0.018010
3	POSITIVE	0.694082

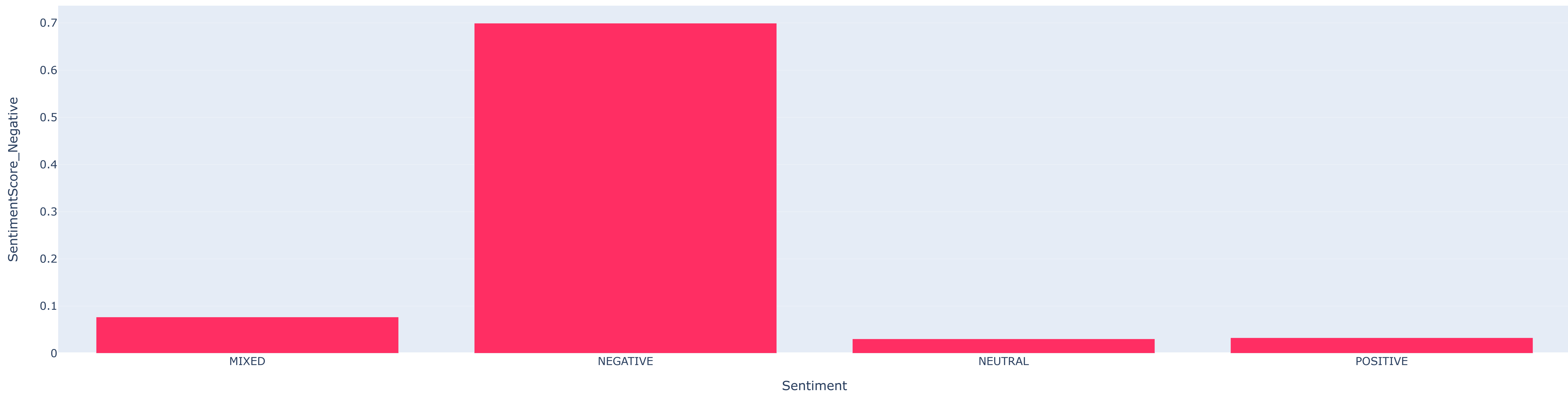


Insights - Positive Sentiment

Positive sentiment score is high for positive sentiment type and low for negative sentiment type.

3.2. Negative

	Sentiment	SentimentScore_Negative
0	MIXED	0.076926
1	NEGATIVE	0.699517
2	NEUTRAL	0.030662
3	POSITIVE	0.033010

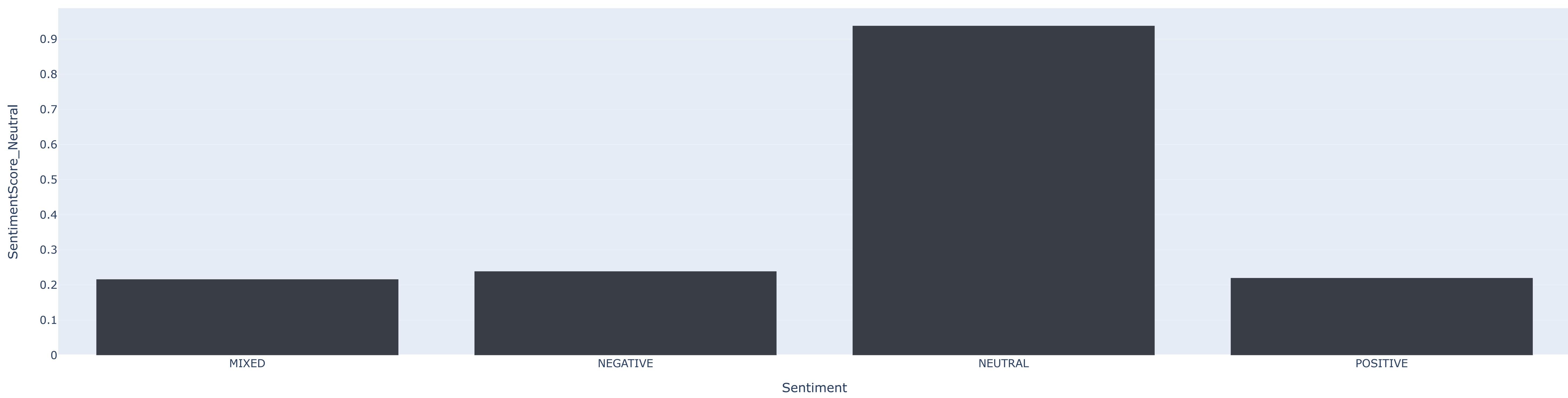


Insights - Negative Sentiment

The Negative sentiment score is high for the negative sentiment type and low for neutral sentiment type.

3.3. Neutral

	Sentiment	SentimentScore_Neutral
0	MIXED	0.216999
1	NEGATIVE	0.239703
2	NEUTRAL	0.938223
3	POSITIVE	0.220765

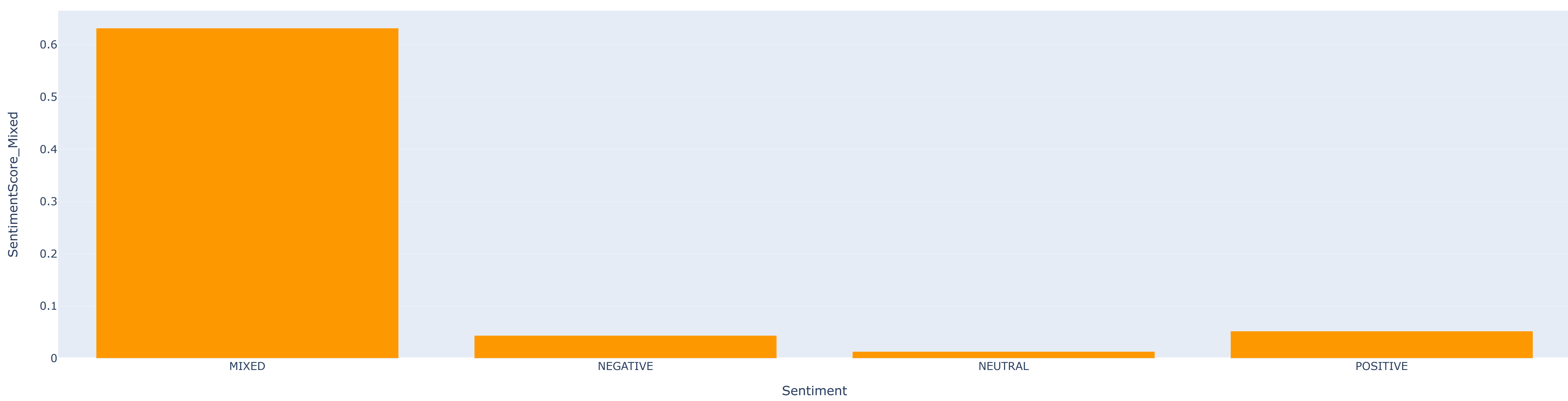


Insights - Neutral Sentiment:

Neutral sentiment score is high for the neutral sentiment type and low for the mixed sentiment type.

3.4. Mixed

	Sentiment	SentimentScore_Mixed
0	MIXED	0.631783
1	NEGATIVE	0.043812
2	NEUTRAL	0.013105
3	POSITIVE	0.052143



Insights - Mixed Sentiment:

The mixed sentiment score is high for the mixed sentiment type and low for neutral sentiment type.

4. Sentiments score by Hashtags

Objective: To find how the hashtags impacts the sentiment scores and to depict about the topic from the data.

4.1. Positive Sentiment with Top 100 Hashtags

	hashtags	SentimentScore_Positive
0	['39K', 'Bitcoin']	0.376727
1	['ADA', 'BUSD', 'Stablecoin', 'BearMarket']	0.005403
2	['ADA', 'Bitcoin']	0.015105
3	['ADA', 'bitcoin']	0.007699
4	['AI', 'NFTs', 'Metaverse', 'metaverse', 'web3...	0.000389
...	...	...
1276	['yieldfields', 'yieldfarming', 'crypto', 'bit...	0.000258
1277	['youtube']	0.001780
1278	['youtubeshort', 'ETH', 'Ethereum', 'Bitcoin',...	0.009844
1279	['zenotrading', 'forex', 'bitcoin', 'zenomarke...	0.000280
1280	[]	0.043467

1281 rows × 2 columns

	hashtags	SentimentScore_Positive
1146	['love', 'tech']	0.964151
726	['TippingTuesday', 'doge', 'dogecoin', 'shib',...	0.927233
1229	['seamlessswapdefi', 'bitcoin', 'Crypto']	0.924746
510	['FlagNetwork', 'IDO', 'cryptocurrency', 'Bitc...	0.918179
1150	['manic']	0.807664
...	...	...
949	['crypto', 'Bitcoin', 'BTC']	0.050383
5	['ALTCOINS', 'BITCOIN']	0.049984
32	['AnjiEco', 'AnjiToken']	0.047844
720	['Thailand', 'Burma', 'Myanmar', 'Laos', 'Mala...	0.047652
732	['Twosday', 'Bitcoin', 'Alts']	0.047173

100 rows × 2 columns

4.2. Negative Sentiments with Top 100 Hashtags

	hashtags	SentimentScore_Negative
0	['39K', 'Bitcoin']	0.006800
1	['ADA', 'BUSD', 'Stablecoin', 'BearMarket']	0.037786
2	['ADA', 'Bitcoin']	0.008253
3	['ADA', 'bitcoin']	0.021263
4	['AI', 'NFTs', 'Metaverse', 'metaverse', 'web3...	0.000688
...	...	...
1276	['yieldfields', 'yieldfarming', 'crypto', 'bit...	0.000190
1277	['youtube']	0.016264
1278	['youtubeshort', 'ETH', 'Ethereum', 'Bitcoin',...	0.007162
1279	['zenotrading', 'forex', 'bitcoin', 'zenomarke...	0.000118
1280	[]	0.080772

1281 rows × 2 columns

	hashtags	SentimentScore_Negative
349	['CRYPTO', 'BTC', 'BITCOIN', 'FUCK_PUTIN']	0.980393
217	['Bitcoin', 'Monero']	0.917629
99	['Binance', 'bitcoin']	0.896226
1173	['newbie', 'Crypto', 'Trader', 'bitcoin']	0.864263
542	['HYPRR']	0.802966
...	...	...
800	['bitcoin', 'MetaverseNFT']	0.074230
815	['bitcoin', 'bitcoin']	0.073893
928	['buythedip', 'Bitcoin', 'BitcoinCrash', 'Cryp...	0.073548
770	['altcoins', 'bitcoin', 'bitdao']	0.073182
682	['Russia', 'Ukraine', 'crypto', 'Bitcoin']	0.071999

100 rows × 2 columns

4.3. Neutral Sentiment with Top 100 Hashtags

	hashtags	SentimentScore_Neutral
0	['39K', 'Bitcoin']	0.614766
1	['ADA', 'BUSD', 'Stablecoin', 'BearMarket']	0.952853
2	['ADA', 'Bitcoin']	0.974812
3	['ADA', 'bitcoin']	0.961021
4	['AI', 'NFTs', 'Metaverse', 'metaverse', 'web3...	0.998916
...	...	...
1276	['yieldfields', 'yieldfarming', 'crypto', 'bit...	0.999534
1277	['youtube']	0.981515
1278	['youtubeshort', 'ETH', 'Ethereum', 'Bitcoin',...	0.981003
1279	['zenotrading', 'forex', 'bitcoin', 'zenomarke...	0.999585
1280	[]	0.834241

1281 rows × 2 columns

	hashtags	SentimentScore_Neutral
1260	['tradewithcasper', 'crypto', 'bitcoin', 'ethe...	0.999610
763	['affiliatemarketing', 'affiliate', 'socialmed...	0.999603
833	['bitcoin', 'coin', 'nft', 'etherium']	0.999600
619	['NFT', 'NFTs', 'NFTCommmunity', 'NFTCollectio...	0.999599
513	['Forex', 'Crypto', 'USDJPY', 'USDX', 'Gold', ...	0.999597
...	...	...
391	['Crypto', 'cryptocurrency', 'CryptoMining', '...	0.999342
959	['crypto', 'bitcoin', 'cryptocurrency', 'block...	0.999341
207	['Bitcoin', 'HJPatel', 'coding', '100DaysOfCod...	0.999340
87	['BTCUSDT', 'BTC', 'USDT', 'Bitcoin', 'Binance...	0.999340
368	['Crypto', 'Bitcoin', 'Ethereum', 'Blockchain']	0.999335

100 rows × 2 columns

4.4. Mixed Sentiment with Top 100 Hashtags

	hashtags	SentimentScore_Mixed
0	['39K', 'Bitcoin']	0.001707
1	['ADA', 'BUSD', 'Stablecoin', 'BearMarket']	0.003958
2	['ADA', 'Bitcoin']	0.001830
3	['ADA', 'bitcoin']	0.010017
4	['AI', 'NFTs', 'Metaverse', 'metaverse', 'web3...	0.000006
...	...	...
1276	['yieldfields', 'yieldfarming', 'crypto', 'bit...	0.000017
1277	['youtube']	0.000442
1278	['youtubeshort', 'ETH', 'Ethereum', 'Bitcoin',...	0.001991

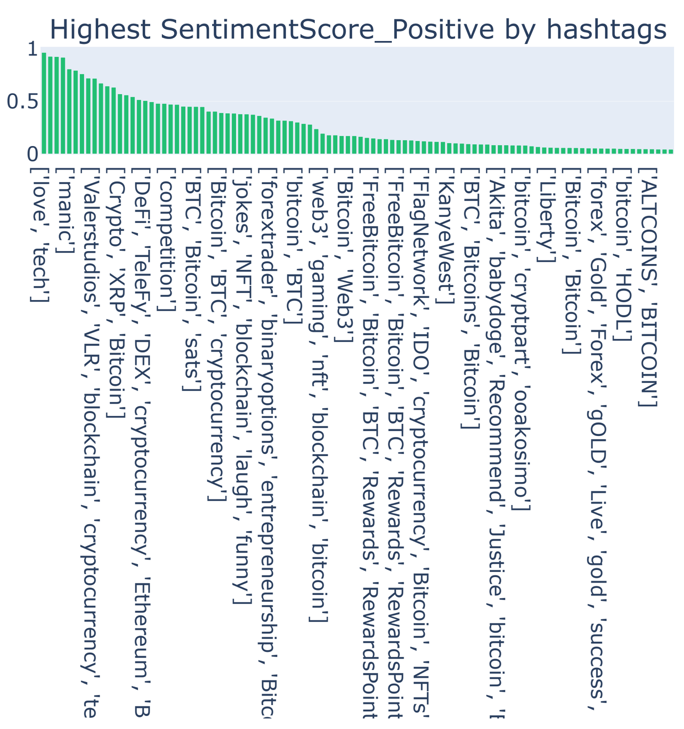
	hashtags	SentimentScore_Mixed
1279	['zenotrading', 'forex', 'bitcoin', 'zenomarke...	0.000017
1280	[]	0.041519

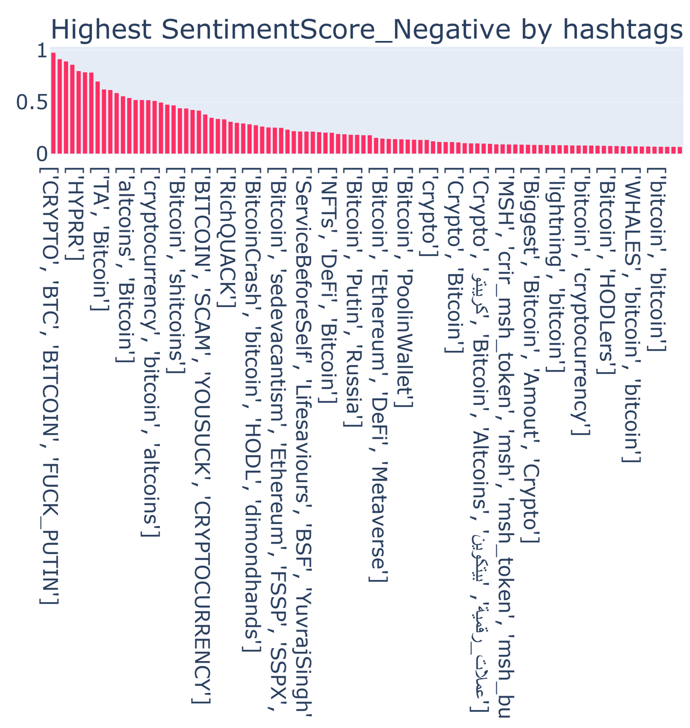
1281 rows × 2 columns

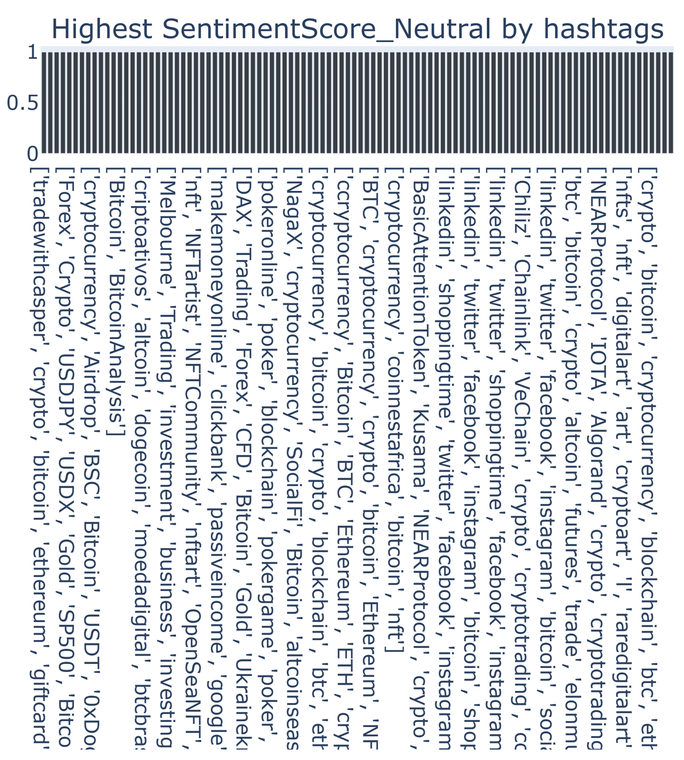
	hashtags	SentimentScore_Mixed
455	['Defi', 'Bitcoin', 'Luna', 'UST']	0.908490
868	['bitcoin', 'thebitcoinstandard', 'crypto']	0.853191
1014	['cryptocurrency', 'FOMO', 'bitcoin']	0.835652
874	['bitcoiners', 'bitcoin', 'forthekids', 'alpha...	0.783056
856	['bitcoin', 'fomo']	0.693800
...	...	...
429	['CryptocurrencyNews', 'cryptocurrency', 'Cryp...	0.026387
1069	['fall', 'Bitcoin', 'Ethereum']	0.026307
810	['bitcoin', 'bike']	0.025901
61	['BTC', 'Bitcoin', 'Crypto']	0.025721
730	['Trudeau', 'Bitcoin']	0.025716

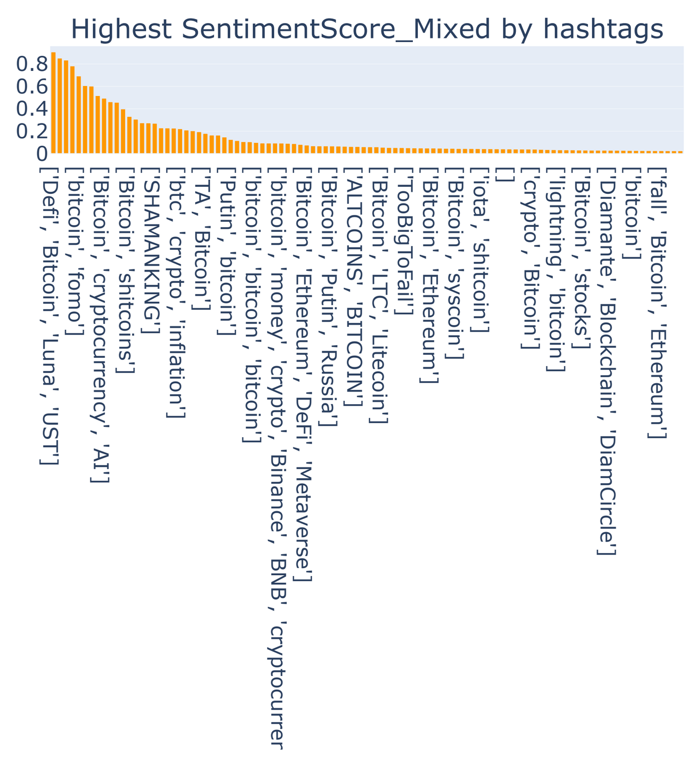
100 rows × 2 columns

The bar plots displays the top 100 hashtags by average sentiment scores.
The bar plots are plotted by taking hashtags at x-axis and Sentiment scores at y-axis.









Insights

The above charts helped to identify the top 100 frequently used hashtags with all types of sentiment based on the keyword "bitcoin", also it shows the entities or topics posted related to the data.

5. Distribution of Followers, Favorites, Retweet and Friends count

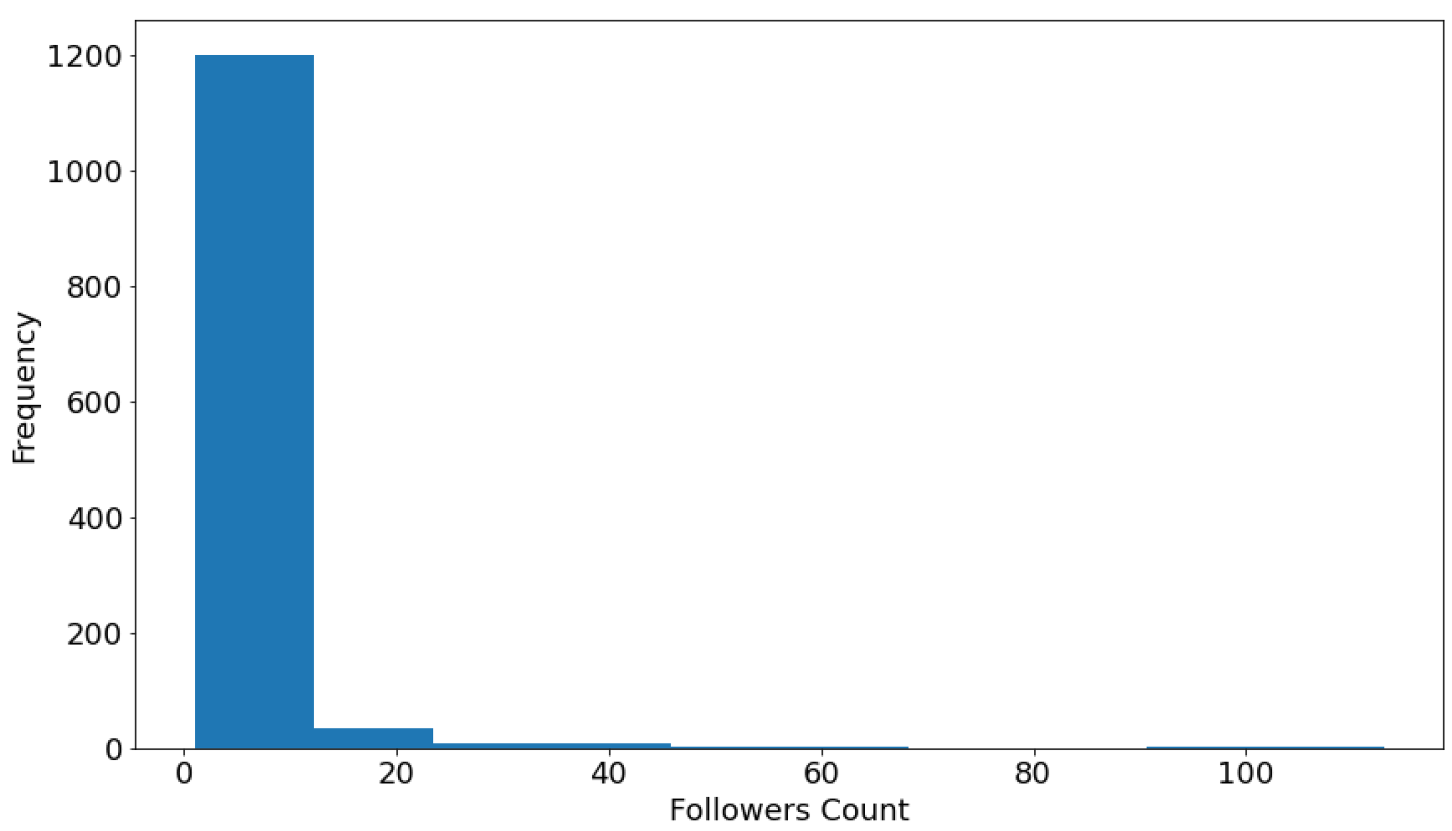
Methods and Objectives

The pandas value counts function is used to count the number of followers, and then the data is plotted using the Matplotlib library.

Objective: To find the frequency distribution of the tweets with respect to followers, favourites, retweet, and friends count.

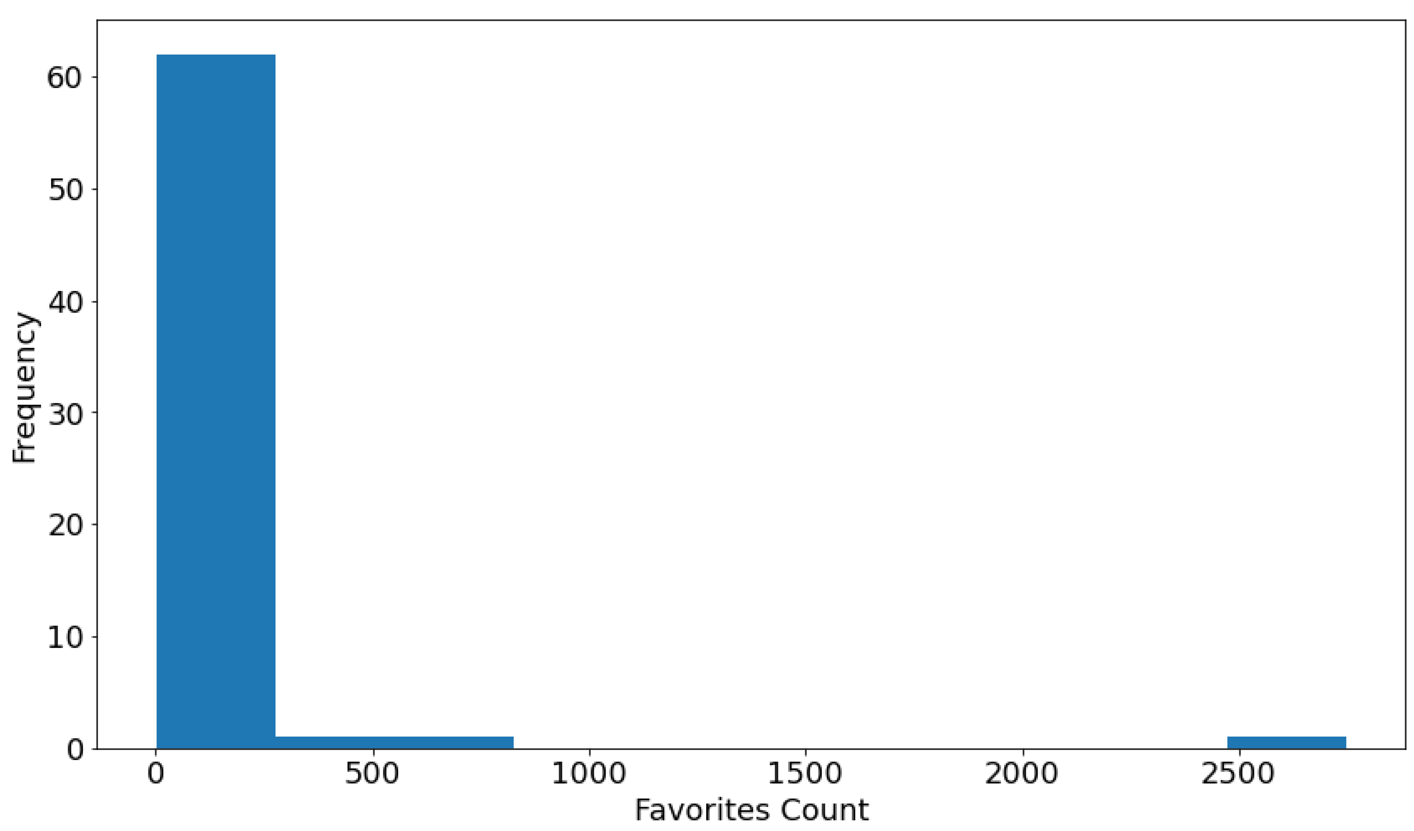
5.1. followers_count Insights:

The number of followers is normally distributed between 0 to 20 with high frequency of the keyword, and it is distributed between 20 to 120 with a relatively very less or no frequency.



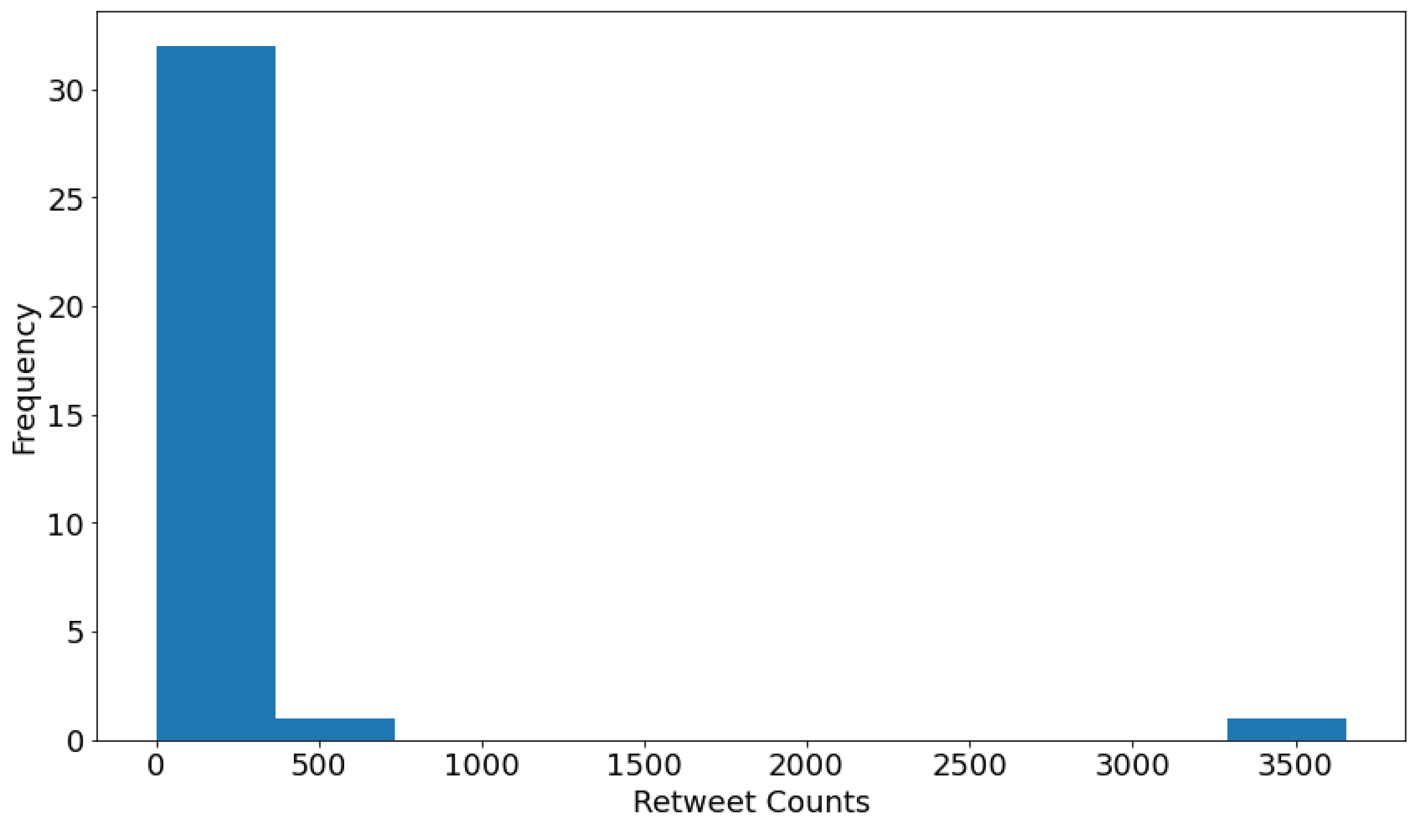
5.2. favorite_count Insights:

The number of favourites count with the highest frequency are distributed between 0 to 250. Between 250 to 500 there is very low frequency. Then from 750 to 2500 there is no frequency and after 2500 there is very low frequency.



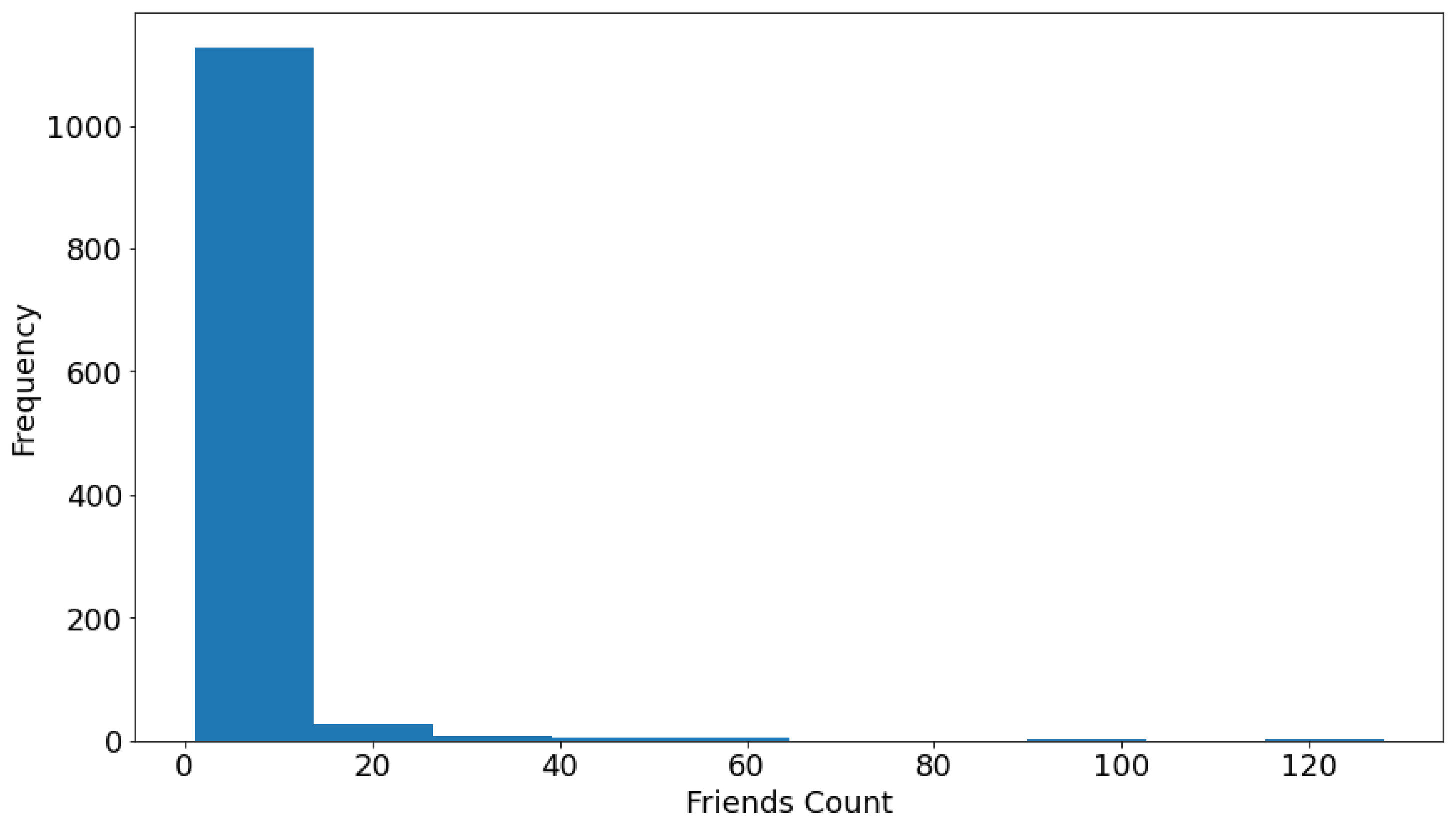
5.3. retweet_count Insights:

The number of retweets count with the highest frequency are distributed between 0 to 500, almost no frequency between 700 to 3250, and a very low frequency between 3250 to 3700.



5.4. friends_count Insights:

The number of friends count is normally distributed between 0 to 20 with high frequency of the keyword, and between 20 to 120 with a relatively low frequency or no frequency.



Insights

The major frequency distribution based on the followers and friends count lies between 0 to 20. The major frequency distribution based on the retweets and favourites count lies between 0 to 500. The frequency of the keyword is high for less the less followers and friends count users. It seems, the frequency is resulting high when counts are less.

6. Correlation Analysis

Correlation is a method for determining the relationships between two variables.

A correlation coefficient that is near to 0, but either positive or negative, indicates that the two variables have less or no relationship.

A correlation coefficient near to 1 indicates that there is a positive relationship between the two variables, with increases in one variable causing increases in the other.

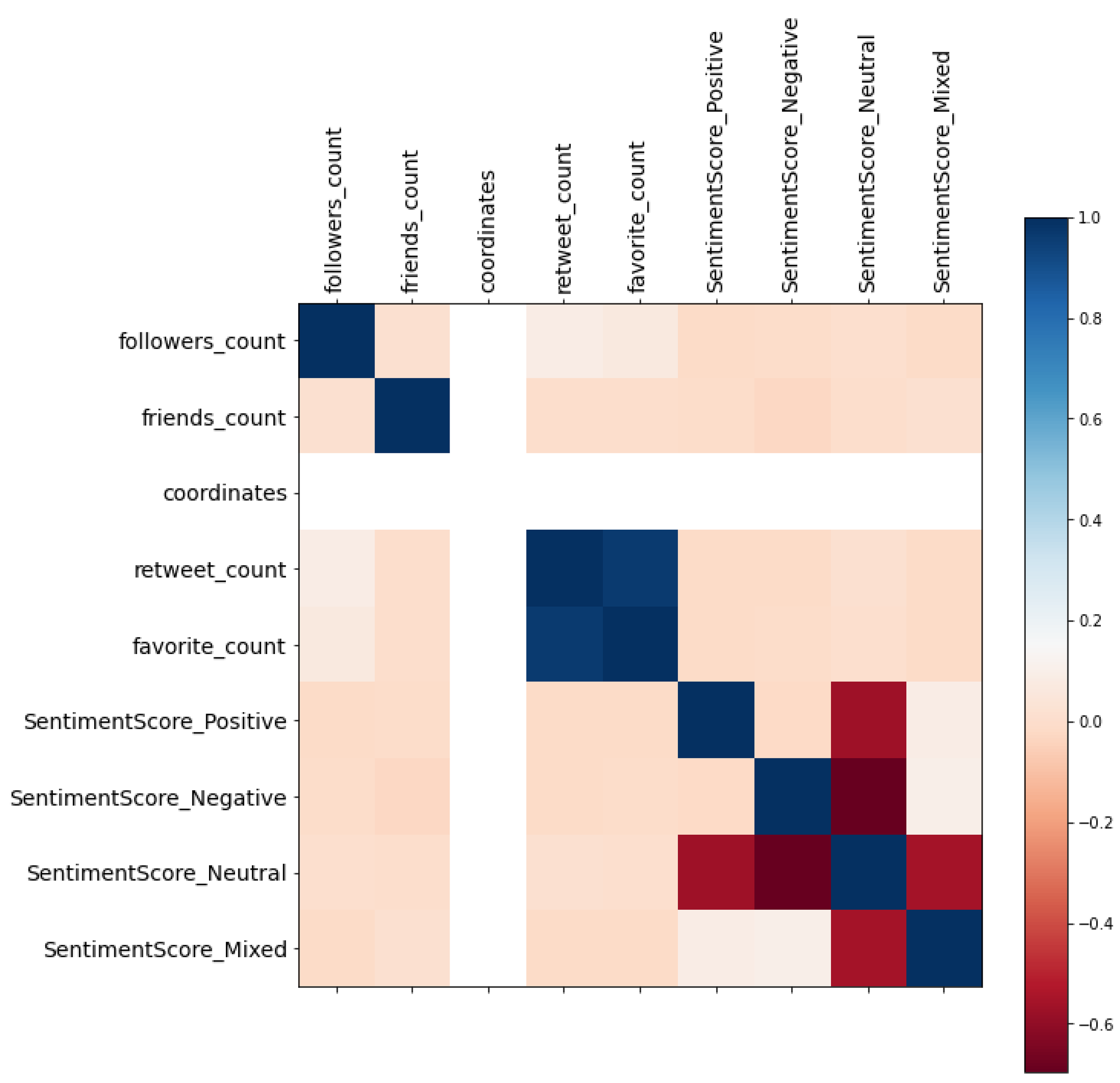
A correlation coefficient near to -1 suggests that there is a negative relationship between two variables.

Method and objectives

Matplotlib library is used to create the dataset's correlation model. The plt.figure() function sets the figure size, and the plt.matshow() function sets the colors.
Finally, the plt.xticks() and plt.yticks() functions plot the columns.

Objective: To investigate the relationship between the retweets, favourites, followers, and friends counts with the sentiment scores.

	followers_count	friends_count	coordinates	retweet_count	favorite_count	SentimentScore_Positive	SentimentScore_Negative	SentimentScore_Neutral	SentimentScore_Mixed
followers_count	1.000000	0.019383	NaN	0.087980	0.072916	-0.008134	-0.005967	0.011801	-0.008122
friends_count	0.019383	1.000000	NaN	0.004601	0.005212	-0.001035	-0.020095	0.005726	0.018401
coordinates	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
retweet_count	0.087980	0.004601	NaN	1.000000	0.966057	-0.011805	-0.008332	0.016904	-0.011769
favorite_count	0.072916	0.005212	NaN	0.966057	1.000000	-0.009696	-0.002194	0.010715	-0.009401
SentimentScore_Positive	-0.008134	-0.001035	NaN	-0.011805	-0.009696	1.000000	-0.013313	-0.565590	0.092674
SentimentScore_Negative	-0.005967	-0.020095	NaN	-0.008332	-0.002194	-0.013313	1.000000	-0.695176	0.104868
SentimentScore_Neutral	0.011801	0.005726	NaN	0.016904	0.010715	-0.565590	-0.695176	1.000000	-0.554221
SentimentScore_Mixed	-0.008122	0.018401	NaN	-0.011769	-0.009401	0.092674	0.104868	-0.554221	1.000000



Insights

We can identify when there is a huge increase in the retweets and favorites count the neutral score will be more positive. Also, when the friends count increases, the mixed sentiment scores will be more positively correlated.

7. Sentiments Score Trend Analysis by Time, Retweets and Favorites Count

Methods and objectives

This analysis uses pandas and plotly libraries. The dataset is reshaped with Pandas, the average scores are calculated with the.mean() function. The scatter plots are plotted with the plotly library.

Objective: To identify the trends or occupancy of the sentiment scores by Time, Retweets, and Favorites count.

	date	SentimentScore_Positive
719	2022-02-22 10:53:37+00:00	0.996263
2452	2022-02-22 11:51:05+00:00	0.995342

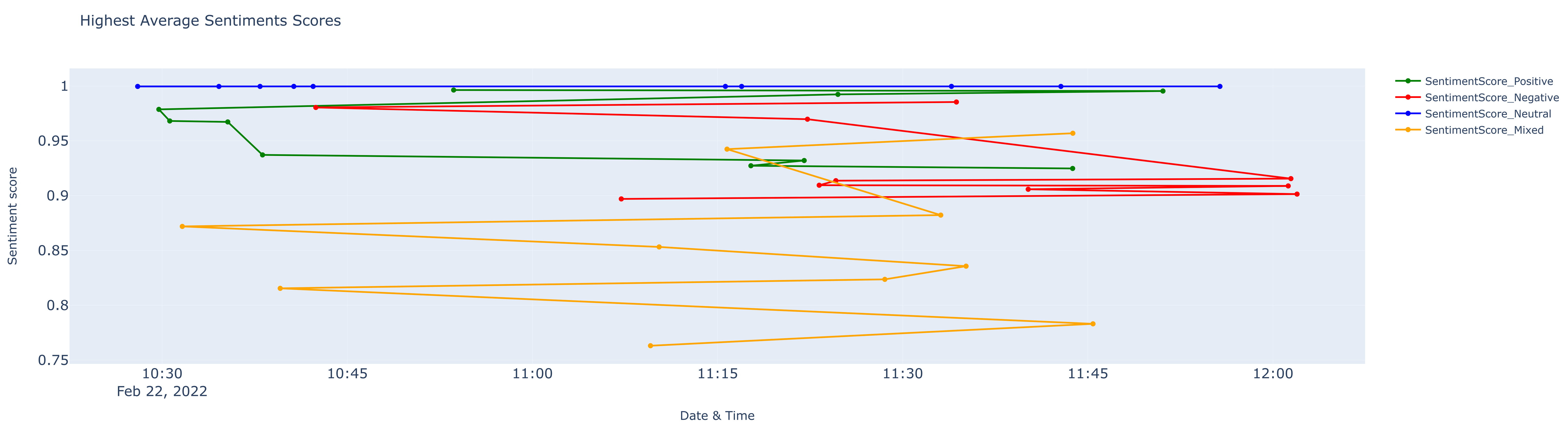
	date	SentimentScore_Positive
1667	2022-02-22 11:24:45+00:00	0.992300
48	2022-02-22 10:29:44+00:00	0.978629
79	2022-02-22 10:30:37+00:00	0.968057
217	2022-02-22 10:35:19+00:00	0.967188
284	2022-02-22 10:38:08+00:00	0.937121
1586	2022-02-22 11:22:01+00:00	0.931993
1473	2022-02-22 11:17:42+00:00	0.927233
2236	2022-02-22 11:43:46+00:00	0.924746

	date	SentimentScore_Negative
1961	2022-02-22 11:34:21+00:00	0.985341
400	2022-02-22 10:42:27+00:00	0.980393
1595	2022-02-22 11:22:17+00:00	0.969666
2763	2022-02-22 12:01:27+00:00	0.915479
1662	2022-02-22 11:24:35+00:00	0.913612
1626	2022-02-22 11:23:14+00:00	0.909454
2754	2022-02-22 12:01:14+00:00	0.908822
2140	2022-02-22 11:40:10+00:00	0.905784
2780	2022-02-22 12:01:57+00:00	0.901380
1153	2022-02-22 11:07:12+00:00	0.896975

	date	SentimentScore_Neutral
1946	2022-02-22 11:33:57+00:00	0.999599
351	2022-02-22 10:40:40+00:00	0.999594
2578	2022-02-22 11:55:42+00:00	0.999585
1413	2022-02-22 11:15:38+00:00	0.999580
277	2022-02-22 10:37:56+00:00	0.999571
393	2022-02-22 10:42:14+00:00	0.999569
2212	2022-02-22 11:42:49+00:00	0.999561
10	2022-02-22 10:28:01+00:00	0.999559
1449	2022-02-22 11:16:57+00:00	0.999550
201	2022-02-22 10:34:36+00:00	0.999546

	date	SentimentScore_Mixed
2237	2022-02-22 11:43:47+00:00	0.956861
1416	2022-02-22 11:15:46+00:00	0.942364
1920	2022-02-22 11:33:05+00:00	0.882220
104	2022-02-22 10:31:38+00:00	0.871897
1249	2022-02-22 11:10:16+00:00	0.853191
1987	2022-02-22 11:35:08+00:00	0.835652
1769	2022-02-22 11:28:33+00:00	0.823624
319	2022-02-22 10:39:34+00:00	0.815435
2292	2022-02-22 11:45:25+00:00	0.783056
1224	2022-02-22 11:09:34+00:00	0.763101

The plots are represented using go.scatter() function. The graph displays the highest average scores across the time period.
The plot dispalys the highest average sentiment scores for each sentiment type.
X-axis shows the date and Y-Axis shows the sentiments score.

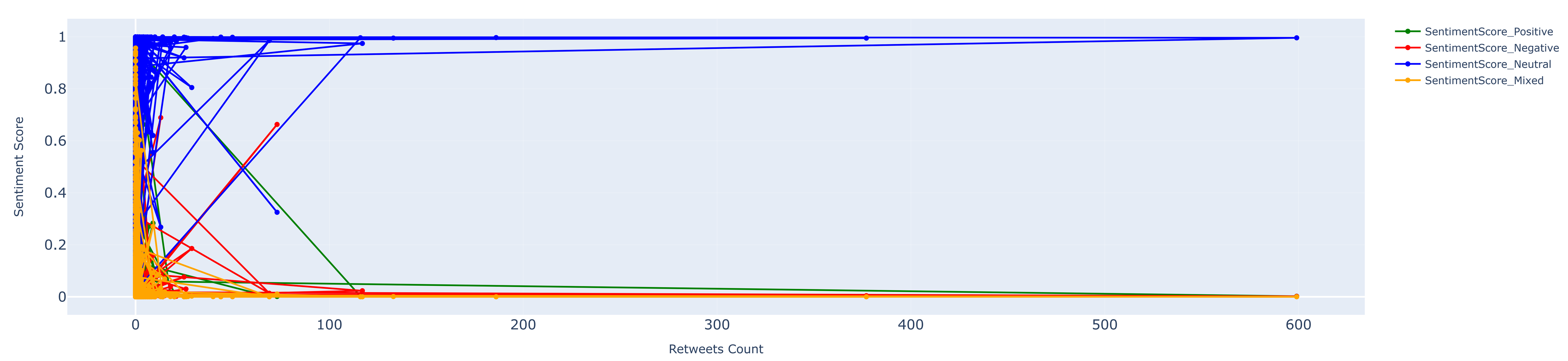


Insights

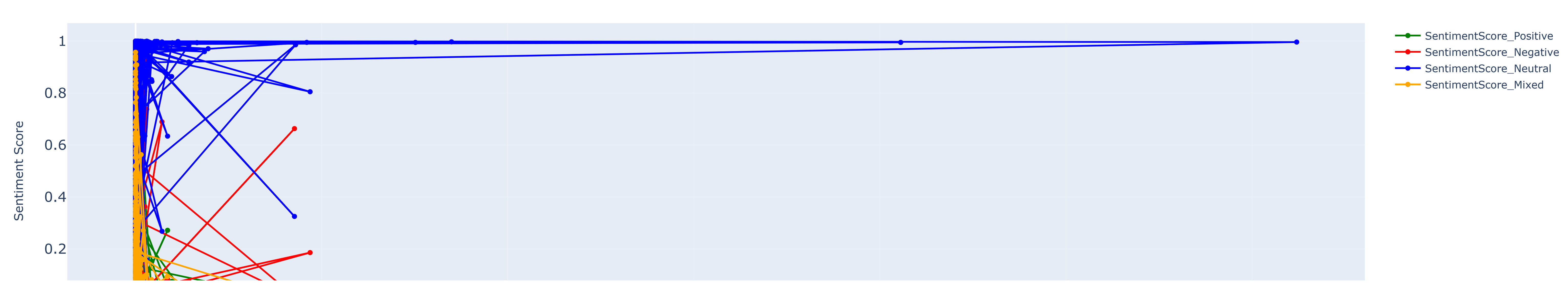
The top 10 highest average sentiment score within the time duration taken to scrap the data, shows that the neutral sentiment is scattered more with high sentiment score when compared to other types.

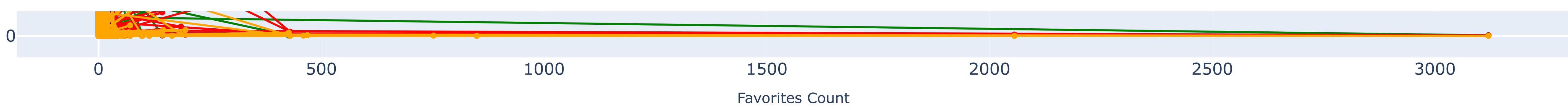
Overall Sentiment Score by Retweet Counts

X-Axis shows the retweet count
Y-Axis shows the sentiment scores



Overall Sentiment Score by Favorites Count





Insights

The neutral sentiment score is high or increasing with respect to the retweets count, while the positive, negative, and mixed sentiment types are less or decreasing.
The neutral sentiment score is high or increasing with respect to the favourites count, while the positive, negative, and mixed sentiment types are less or decreasing.

8. Predictive Analysis for Sentiment Scores by Retweet and Favorites counts

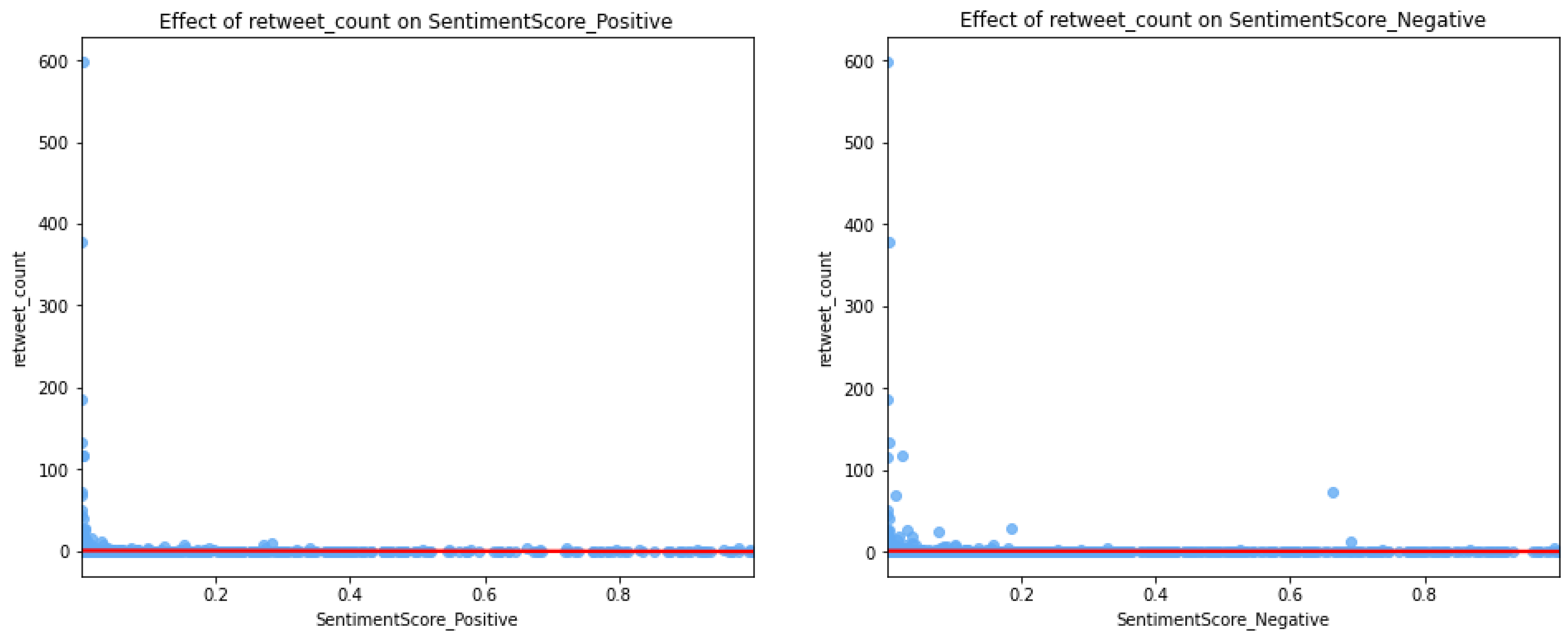
Methods and objectives

Seaborn library is used to plot the regression plots.

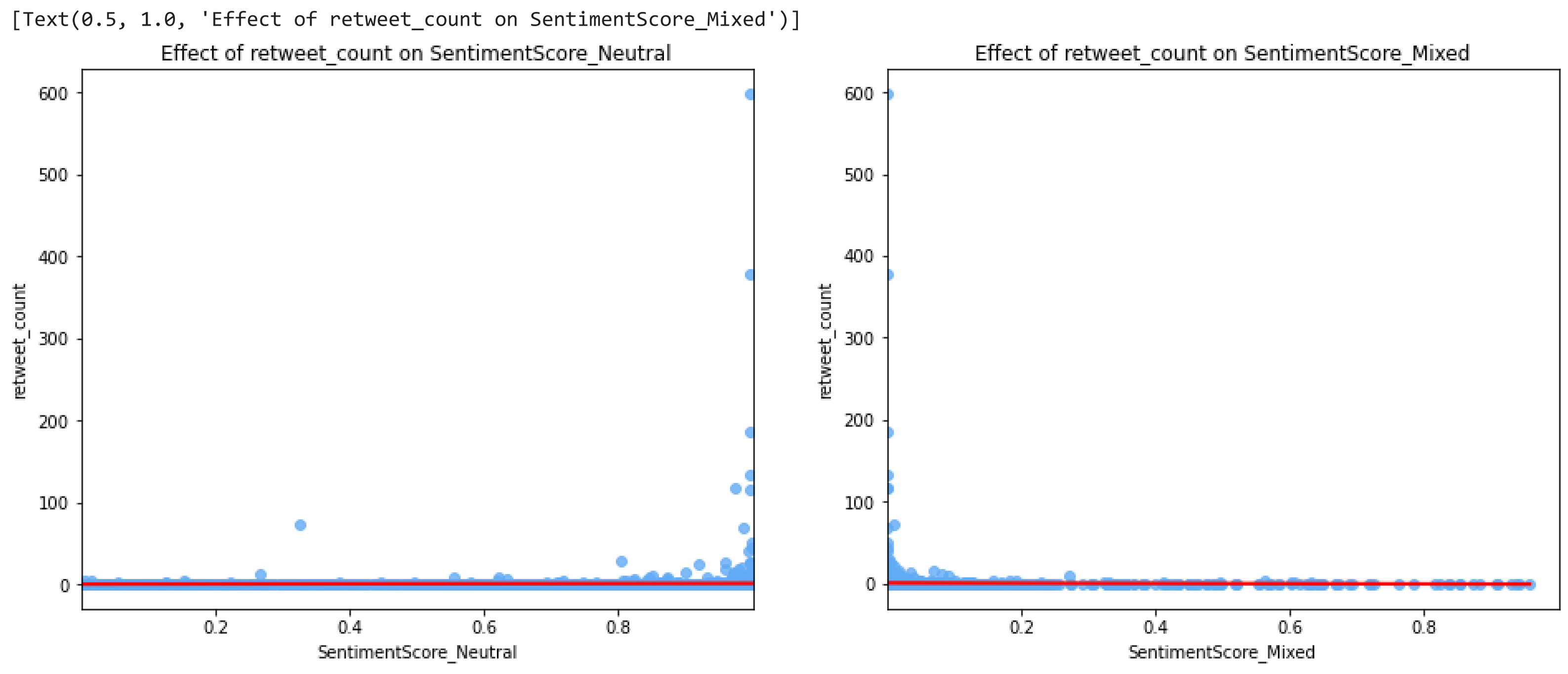
Objective: To find out how the sentiment scores will be affected in the future with the increase in retweets and favorite counts

8.1. Retweet - Positive, Negative, Neutral & Mixed

```
[Text(0.5, 1.0, 'Effect of retweet_count on SentimentScore_Negative')]
```

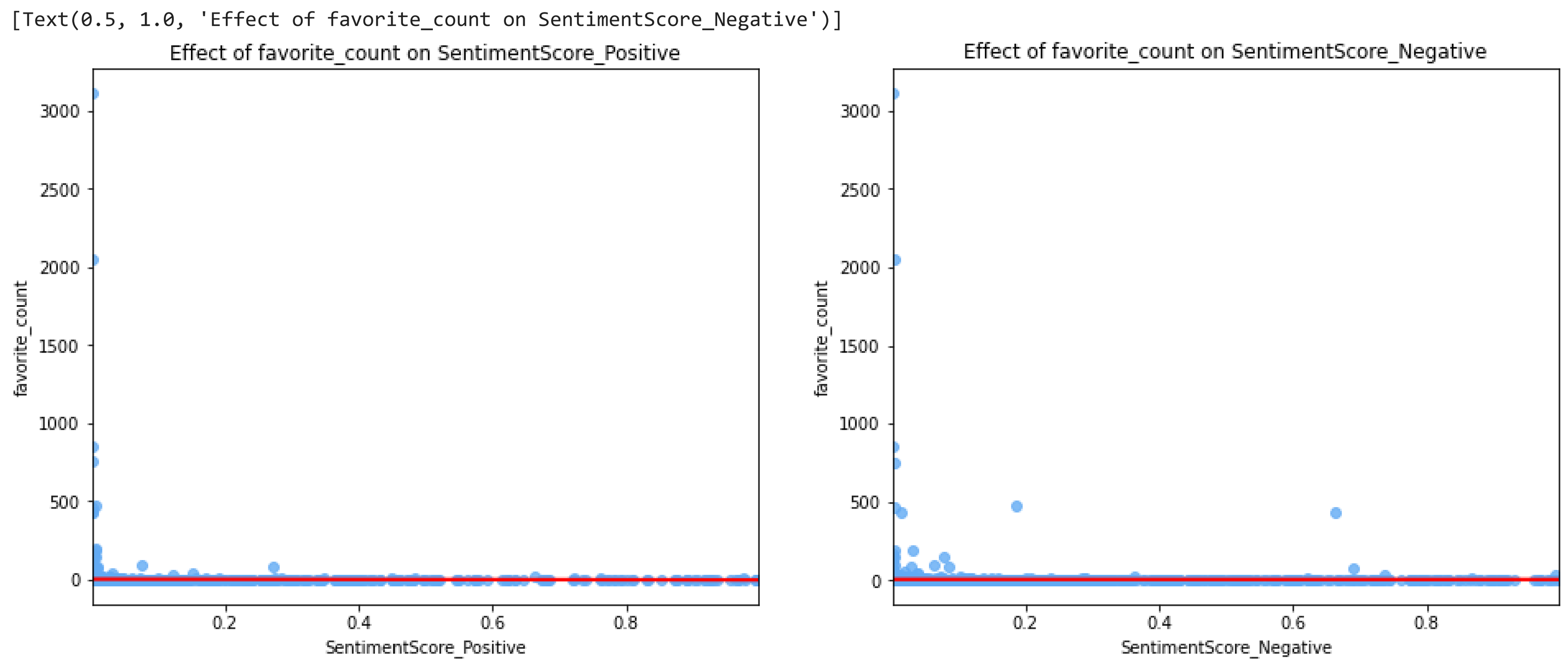



```
[Text(0.5, 1.0, 'Effect of retweet_count on SentimentScore_Mixed')]
```

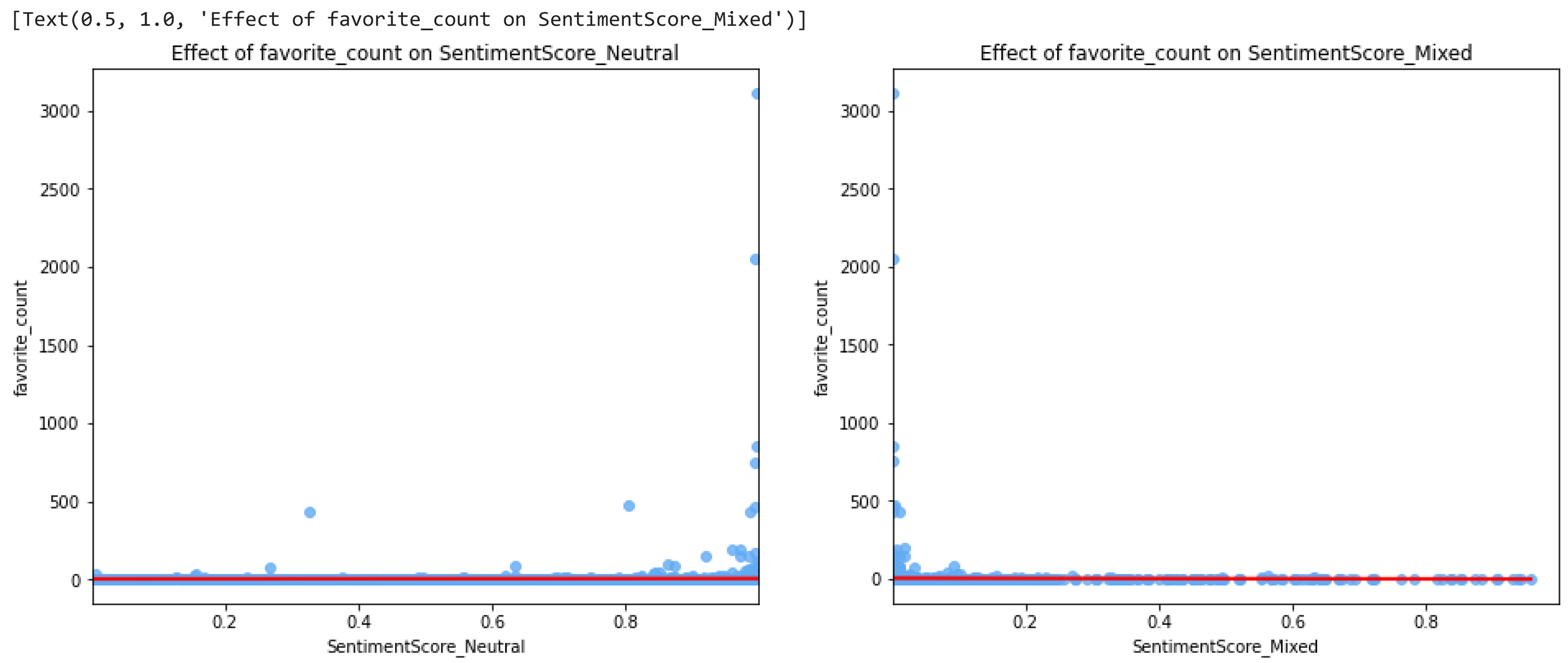


8.2. Favourite - Positive, Negative, Neutral & Mixed

```
[Text(0.5, 1.0, 'Effect of favorite_count on SentimentScore_Negative')]
```




```
[Text(0.5, 1.0, 'Effect of favorite_count on SentimentScore_Mixed')]
```



Insights

There is no increase in the sentiment scores with the increase of retweet and favourite counts except neutral sentiments. Neutral sentiment scores increase with respect to retweets and favourite counts.