INTRODUCTION TO DATA SCIENCE AND ANALYTICS

Project Delivery #3:
Explore your data - Part 2

Project Title:
Social Media Sentiment Analysis

MARMARA UNIVERSITY

1. Chart/Figures of Attribute

Number of Letters

Other than the label feature, there is an attribute named “tweet”. Related charts/figures are given below. By counting the letters of the tweets in the dataset, we created the chart in Figure 1 that shows the frequency and the relative frequency of the letters of the whole tweets.Then, we applied a chi-square test to see whether the distribution of the letters in tweets is the same with what we expect from English texts.

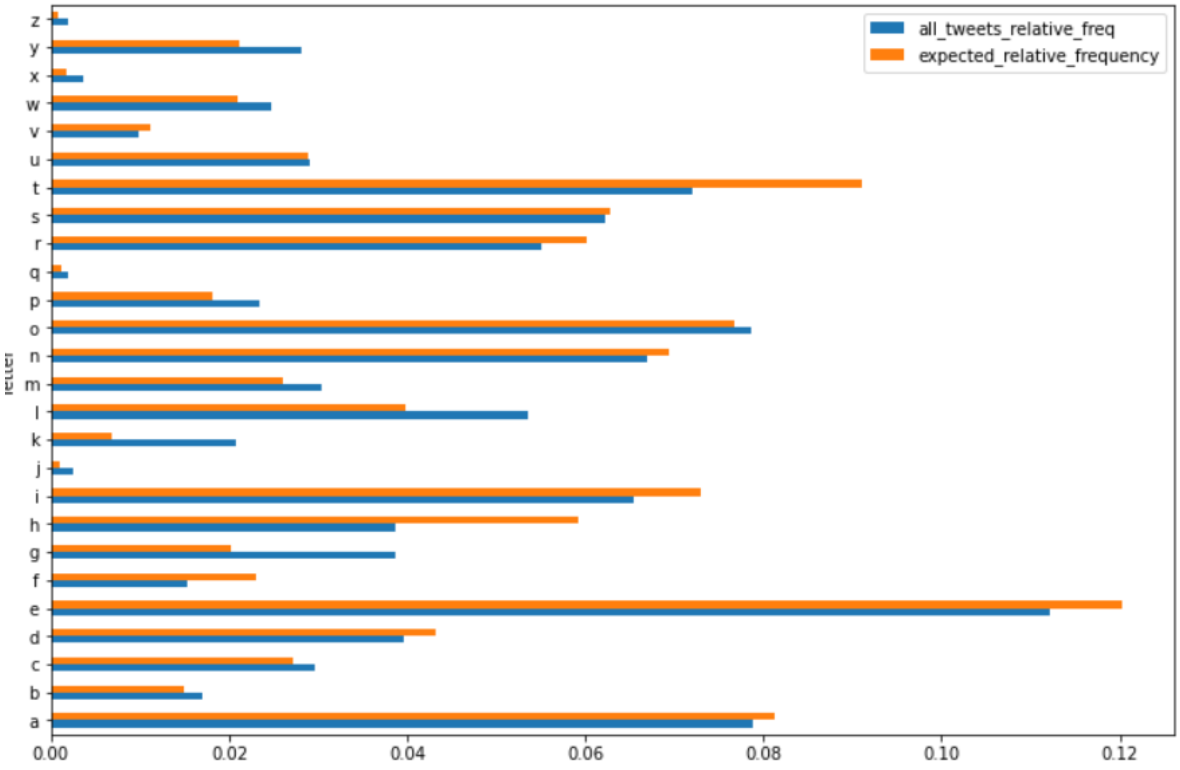
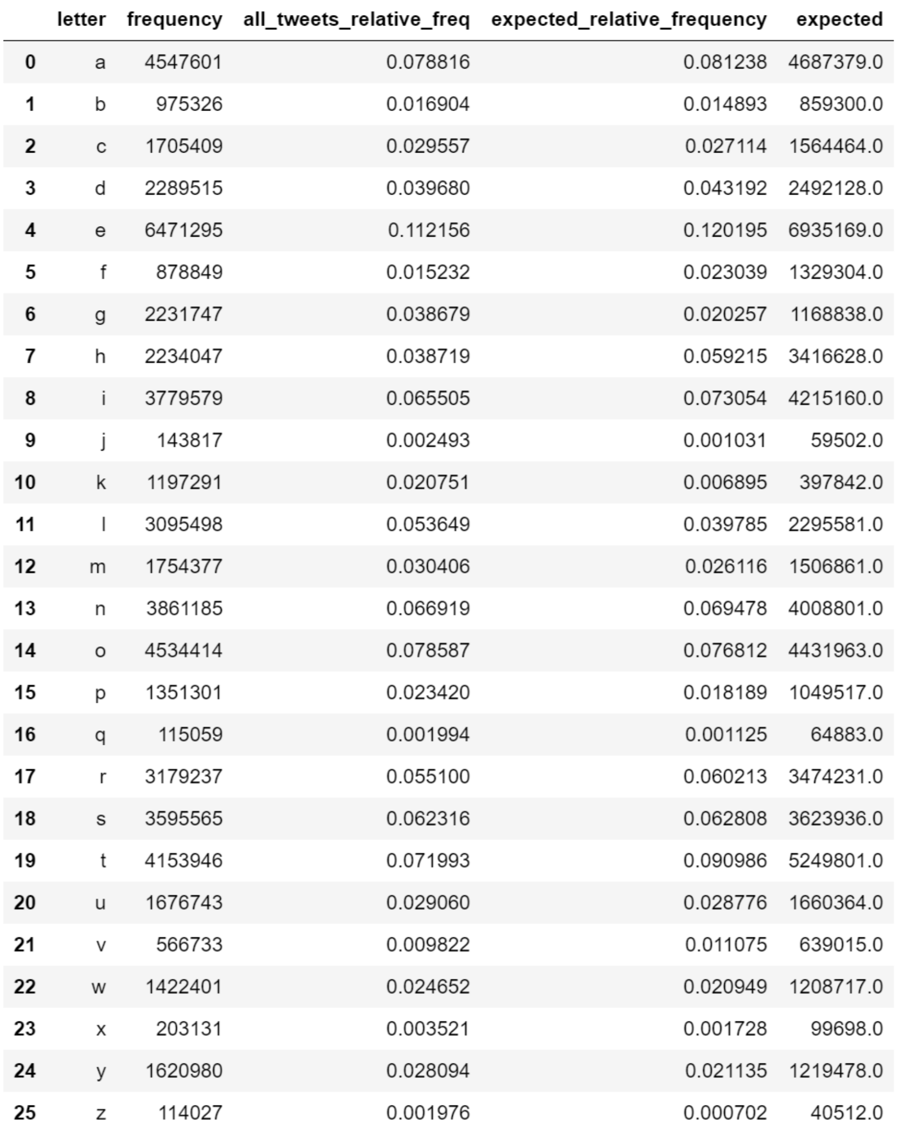


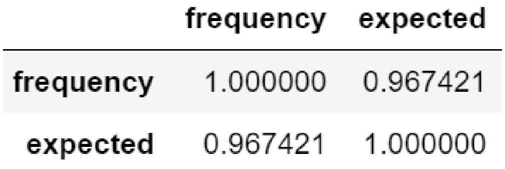
Figure 1. Letter frequencies of each 26 characters in English Alphabet.



	letter	frequency	all_tweets_relative_freq	expected_relative_frequency	expected
0	a	4547601	0.078816	0.081238	4687379.0
1	b	975326	0.016904	0.014893	859300.0
2	c	1705409	0.029557	0.027114	1564464.0
3	d	2289515	0.039680	0.043192	2492128.0
4	e	6471295	0.112156	0.120195	6935169.0
5	f	878849	0.015232	0.023039	1329304.0
6	g	2231747	0.038679	0.020257	1168838.0
7	h	2234047	0.038719	0.059215	3416628.0
8	i	3779579	0.065505	0.073054	4215160.0
9	j	143817	0.002493	0.001031	59502.0
10	k	1197291	0.020751	0.006895	397842.0
11	l	3095498	0.053649	0.039785	2295581.0
12	m	1754377	0.030406	0.026116	1506861.0
13	n	3861185	0.066919	0.069478	4008801.0
14	o	4534414	0.078587	0.076812	4431963.0
15	p	1351301	0.023420	0.018189	1049517.0
16	q	115059	0.001994	0.001125	64883.0
17	r	3179237	0.055100	0.060213	3474231.0
18	s	3595565	0.062316	0.062808	3623936.0
19	t	4153946	0.071993	0.090986	5249801.0
20	u	1676743	0.029060	0.028776	1660364.0
21	v	566733	0.009822	0.011075	639015.0
22	w	1422401	0.024652	0.020949	1208717.0
23	x	203131	0.003521	0.001728	99698.0
24	y	1620980	0.028094	0.021135	1219478.0
25	z	114027	0.001976	0.000702	40512.0

Figure 2. Letter frequency of the dataset, relative frequencies of the dataset, expected relative frequency according to the English language and expected character length according to the English language.

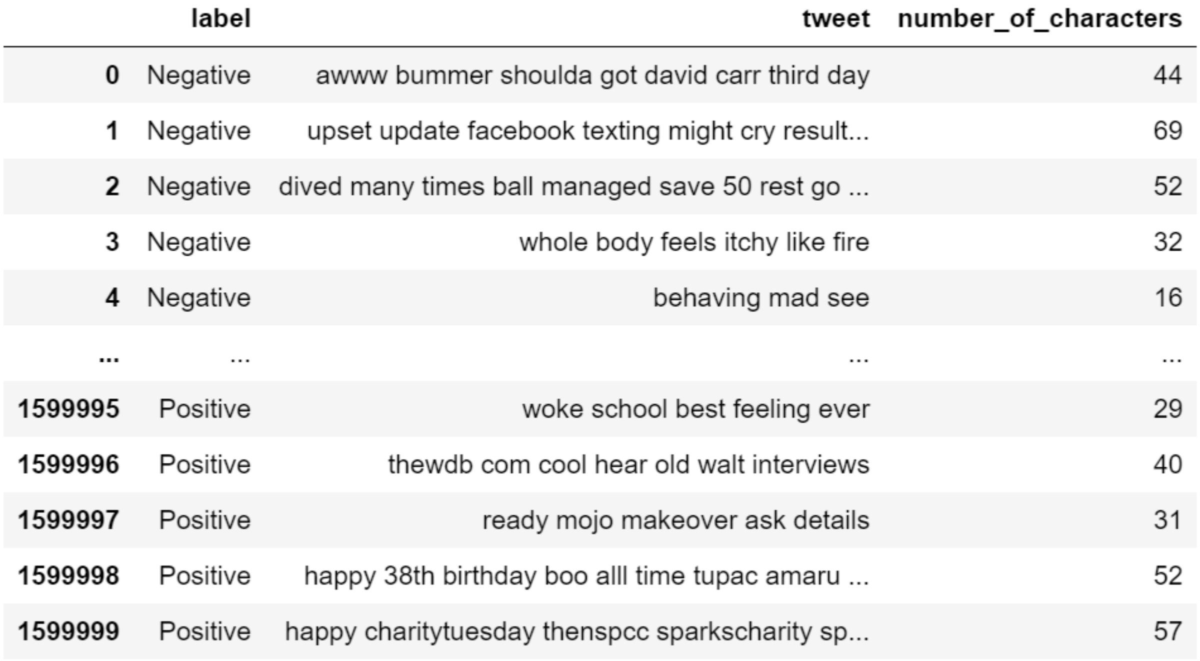
We got the p-value (p) as 0 which implies that the letter frequency does not follow the same distribution with what we see in English tests, although the Pearson correlation is too high (~96.7%) as shown in Figure 3.



	frequency	expected
frequency	1.000000	0.967421
expected	0.967421	1.000000

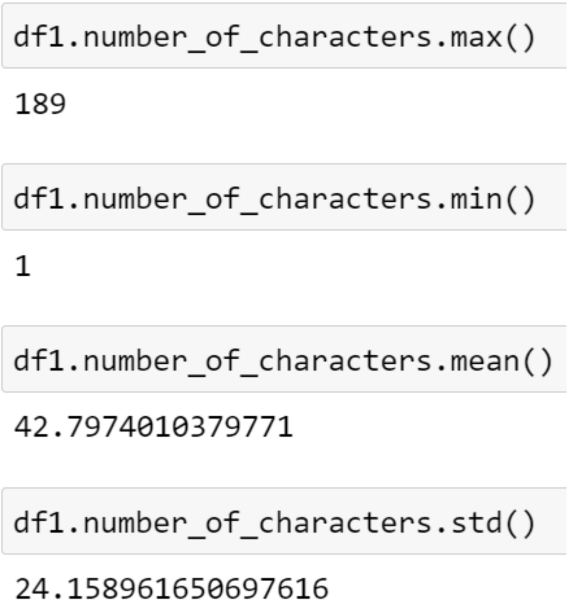
Figure 3. Correlation.

We counted the number of characters for each tweet (Figure 4) and analyzed the data frame according to maximum number of characters, minimum number of characters, mean of the number of characters column and its standard deviation. Our longest tweet is 189 characters long, the shortest tweet is 1 character long and mean of all tweets' character length 42.78. The standard deviation of all tweet character length is 24.16 as shown in Figure 5.



	label	tweet	number_of_characters
0	Negative	awww bummer shoulda got david carr third day	44
1	Negative	upset update facebook texting might cry result...	69
2	Negative	dived many times ball managed save 50 rest go ...	52
3	Negative	whole body feels itchy like fire	32
4	Negative	behaving mad see	16
...	...	...	...
1599995	Positive	woke school best feeling ever	29
1599996	Positive	thewdb com cool hear old walt interviews	40
1599997	Positive	ready mojo makeover ask details	31
1599998	Positive	happy 38th birthday boo alll time tupac amaru ...	52
1599999	Positive	happy charitytuesday thenspcc sparkscharity sp...	57

Figure 4. Number of characters.



```
df1.number_of_characters.max()
```

```
189
```

```
df1.number_of_characters.min()
```

```
1
```

```
df1.number_of_characters.mean()
```

```
42.7974010379771
```

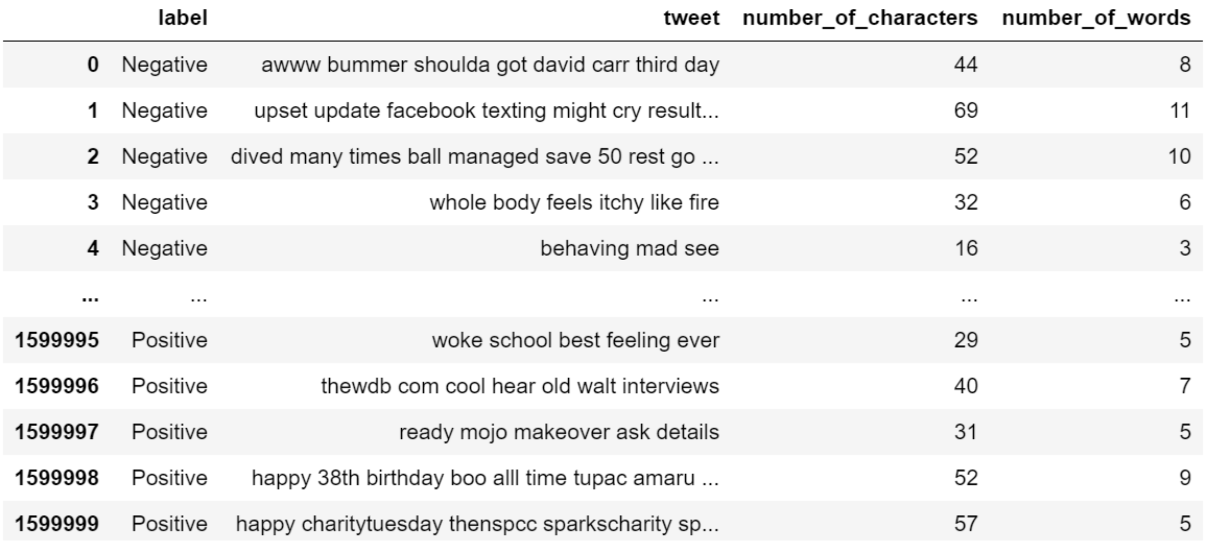
```
df1.number_of_characters.std()
```

```
24.158961650697616
```

Figure 5. Max, min, mean and standard deviation of each tweet in terms of character length.

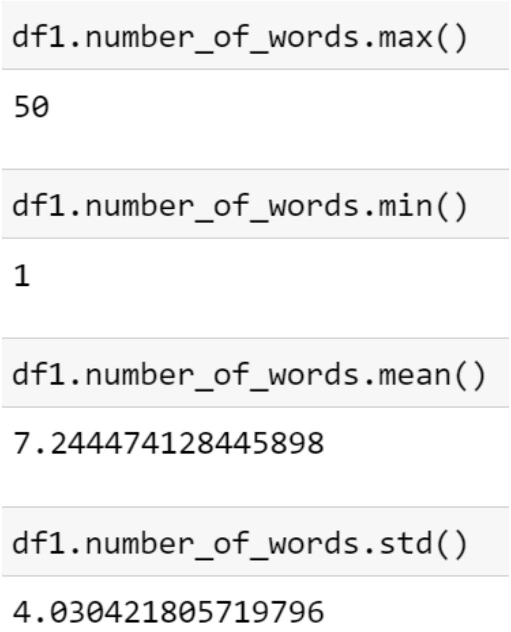
Number of Words

We counted the number of words for each tweet (Figure 6) and analyzed the data frame according to maximum number of words, minimum number of words, mean of the number of words column and its standard deviation. Our longest tweet is 50 words long, the shortest tweet is 1 word long and the mean of all tweets' word length is 7.24. The standard deviation of all tweet character length is 4.03 as shown in Figure 7.



	label	tweet	number_of_characters	number_of_words
0	Negative	awww bummer shoulda got david carr third day	44	8
1	Negative	upset update facebook texting might cry result...	69	11
2	Negative	dived many times ball managed save 50 rest go ...	52	10
3	Negative	whole body feels itchy like fire	32	6
4	Negative	behaving mad see	16	3
...	...	...	...	...
1599995	Positive	woke school best feeling ever	29	5
1599996	Positive	thewdb com cool hear old walt interviews	40	7
1599997	Positive	ready mojo makeover ask details	31	5
1599998	Positive	happy 38th birthday boo alll time tupac amaru ...	52	9
1599999	Positive	happy charitytuesday thenspcc sparkscharity sp...	57	5

Figure 6. Number of words of each tweet.



```
df1.number_of_words.max()

50

df1.number_of_words.min()

1

df1.number_of_words.mean()

7.244474128445898

df1.number_of_words.std()

4.030421805719796
```

Figure 7. Max, min, mean and standard deviation of each tweet in terms of number of words.

Most common words in the whole dataset and positive/negative tweets of the dataset are given in the following figures.

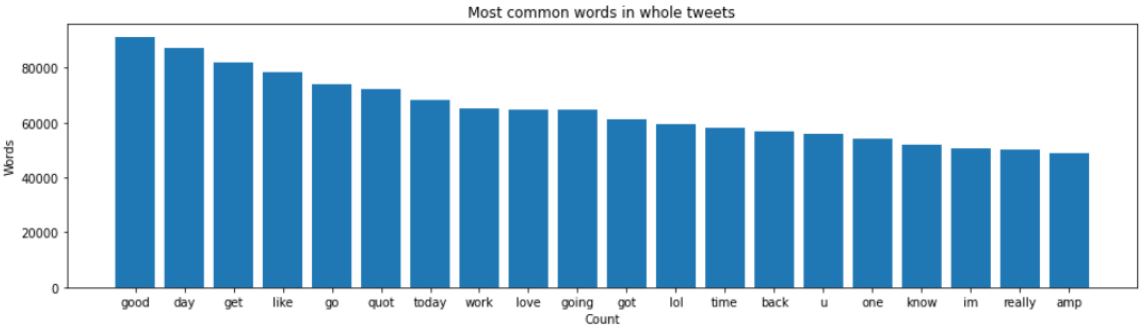


Figure 8. Most common words in our dataset.

Positive Tweets:

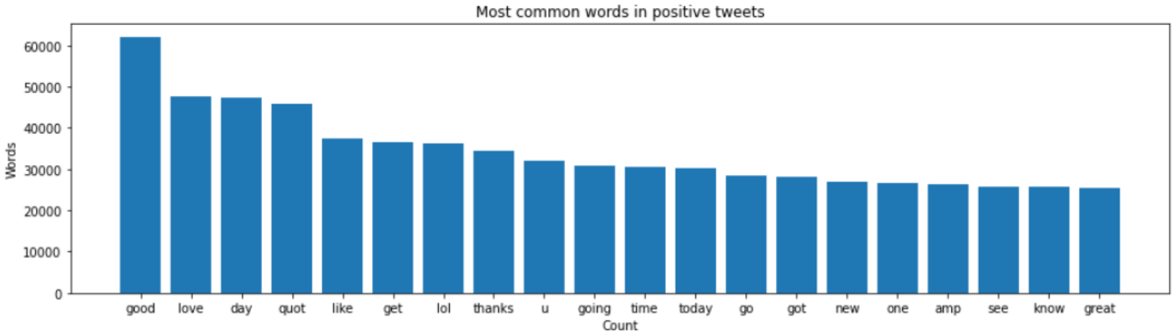


Figure 9. Most common words in positive tweets in our dataset.

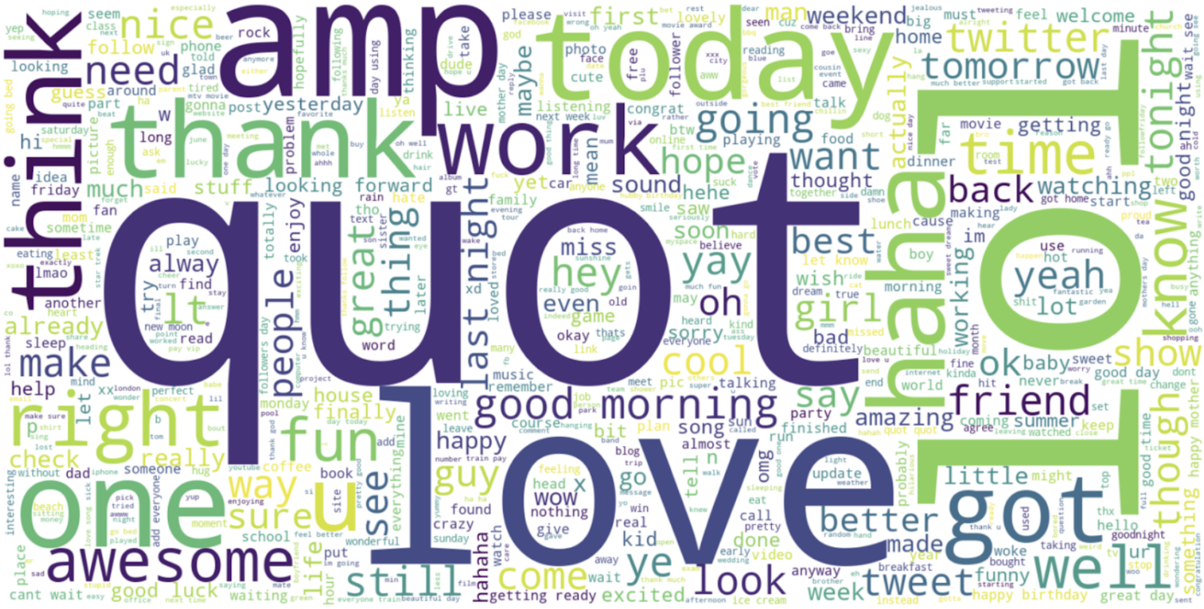


Figure 10. Word cloud of positive tweets.

Negative Tweets:

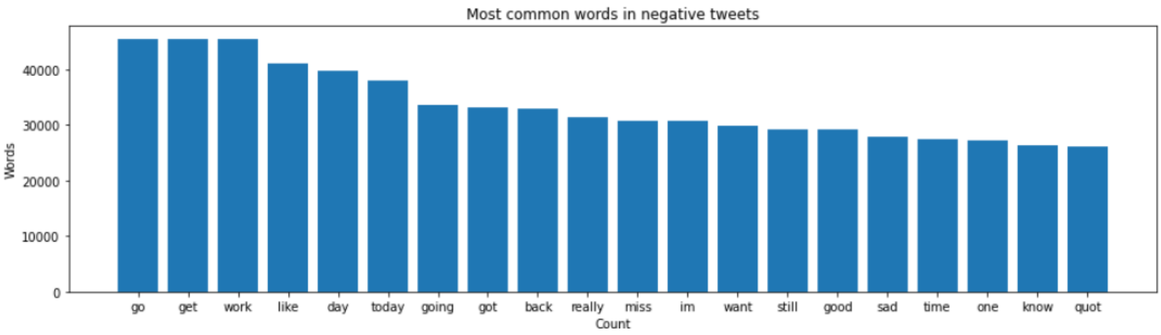


Figure 11. Most common words in negative tweets in our dataset.

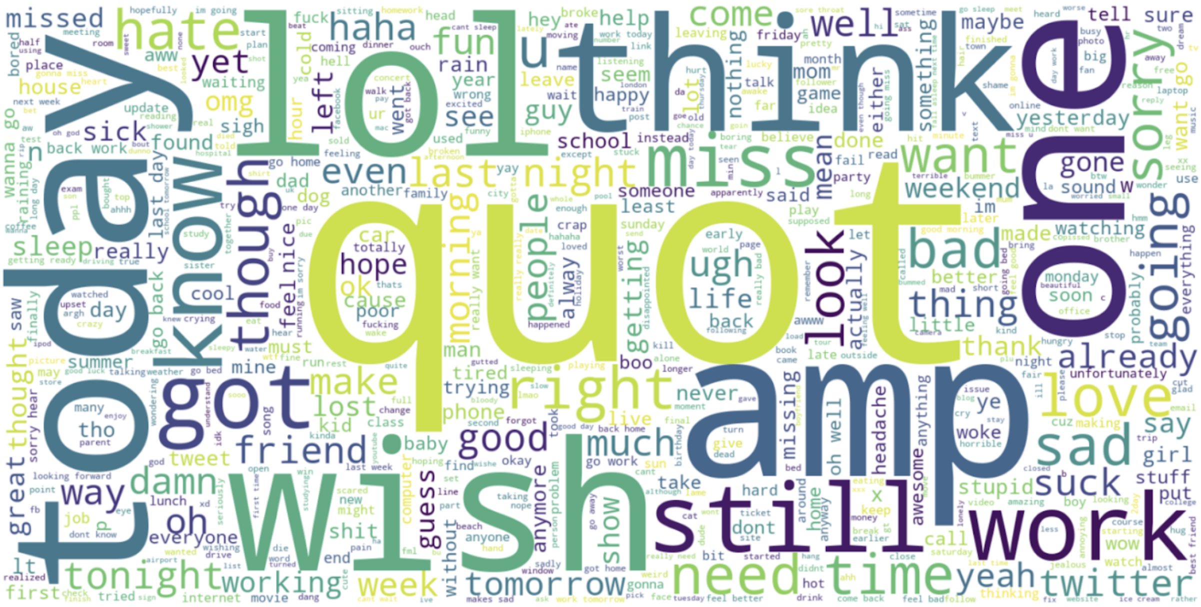


Figure 12. Word cloud of positive tweets.

Positive and negative samples are equal. The dataset distribution has not any skewness as shown in Figure 13.

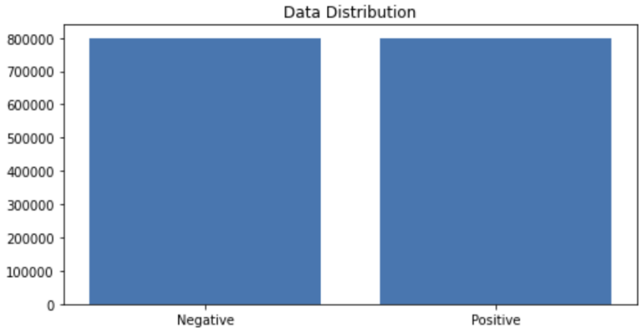


Figure 13. Positive and negative tweets distribution.

2. Scatter Plot

We used feature extraction methods, bag-of-words, and word embedding. Bag of words with TF-IDF is a common and simple way of feature extraction. Bag-of-Words is a representation model of text data and TF-IDF is a calculation method to score the importance of words in a document.

After applying bag-of-words with TF-IDF, we create the scatter plot according to these results.

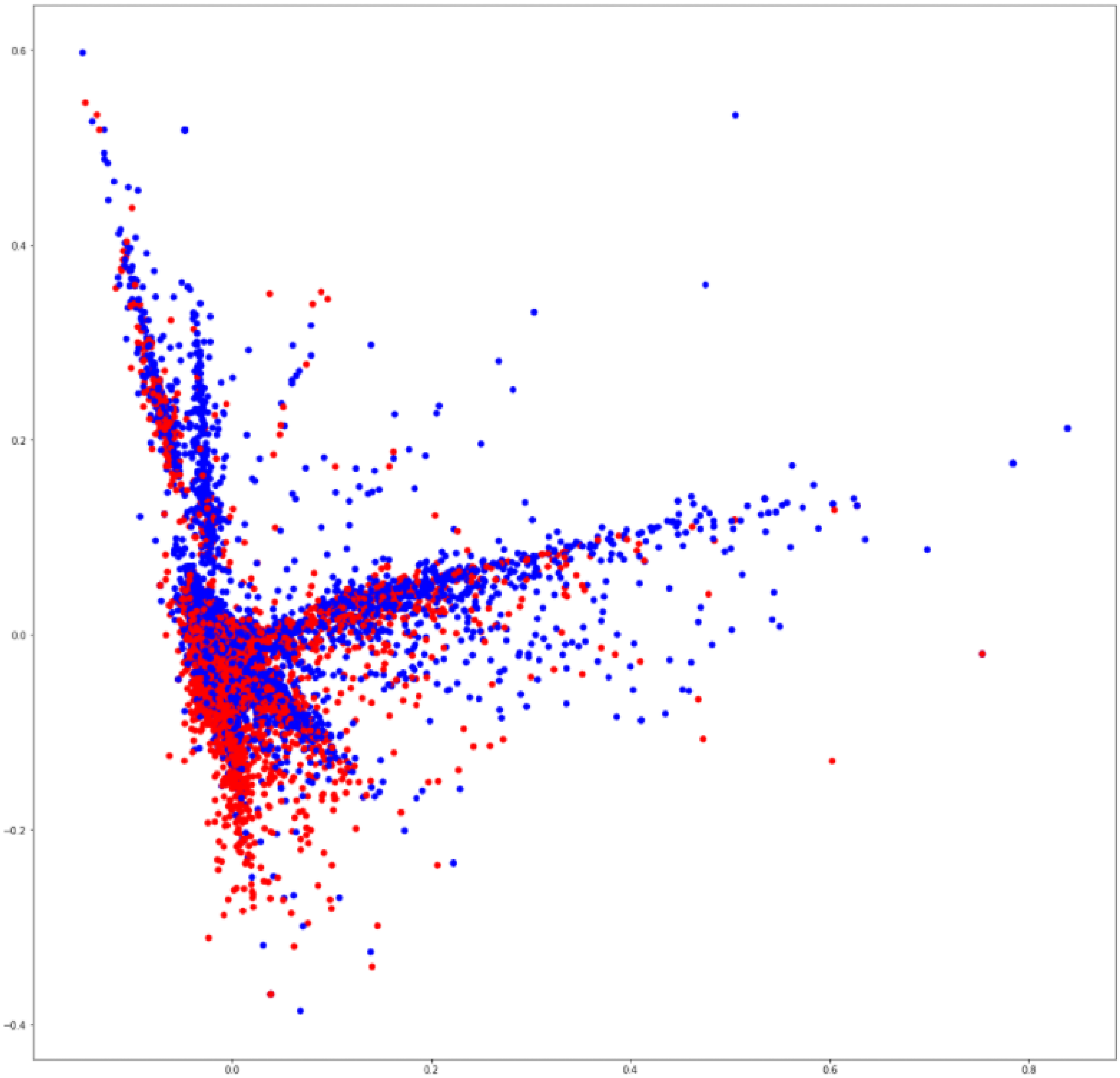


Figure 14. Scatter plot that shows correlation of words in the corpus: red indicates negatives, blue indicates positives.

3. Results

In this delivery, we explored our dataset by applying some analyses to the attributes and created related charts. There are 2 attributes in our dataset including label attribute. We applied these analyses on them.

We explored the tweets by looking at the letters and words in them. First of all, we counted the letters of all tweets and calculated the letter frequencies. Then we compared the letter frequency of our data with the expected frequency of the letters of the alphabet of English. In Figure 1, this comparison is shown. As seen in the graph, even though there are some exceptions, for most of the letter, the frequencies of our data is really close to the expected ones.

The number of characters and words are also counted and analysed. Minimum number of characters of all tweets is 1 whereas the maximum number is 189. Since the mean is around 42 and standard deviation is around 24, it can be said that a small number of tweets has a high number of characters. The similar result can be seen in word analysis . When the number of words counted, it is seen that the maximum number of words in tweets is 50 whereas the minimum number is 1. Mean is around 7 and standard deviation is around 4 which gives a similar result with the number of characters. Very small number of tweets has a high number of words. According to these results, it can be interpreted that both the number of characters and number of words graphs are skewed graphs.

After counting the number of words used in tweets, word usages are analysed. Since the stop words are usually the most used words in texts and they may prevent us from getting the right results, they are calculated by filtering the stopwords. The results are shown in Figure 8. Also, most common words for positive and negative labels are separated and shown in Figures 9, 10, 11 and 12.

Then, as mentioned in Part 2, by using some feature extraction methods, a scatter plot is obtained. The plot (Figure 14) shows the correlation between the words.