MARMARA UNIVERSITY

CSE 4062.1 SPRING 2021
INTRODUCTION TO DATA SCIENCE AND ANALYTICS
PROJECT DELIVERY #2: Exploring your data

Group #3

Project Title:

Social Media Sentiment Analysis

Group Members:

Arda Bayram 150116029 arda.bayram1998@gmail.com (Computer Engineering)
Göksel Tokur 150116049 gokseltokur04@gmail.com (Computer Engineering)
Gülnihal Erdem 150319554 gulnihalerdem@gmail.com (Industrial Engineering)
İrem Seçmen 170219006 iremsecmenn@gmail.com (Electrical and Electronics Engineering)

Dataset

The dataset [1] contains 1,600,000 tweets extracted using the twitter api. The tweets have been classified from 0 (negative) to 4 (positive). The dataset contains 6 fields which are target as integer, ids as integer, date as date, flag as string, user as string and text as string.These 6 fields are shown below.

- target: The polarity of the tweet (0 - negative, 2 - neutral, 4 - positive)
- ids: The id of the tweet.
- date: The date of the tweet.
- flag: The query. If there is no query, then this value is NO_QUERY.
- user: The user that tweeted.
- text: The text of the tweet

Target	Ids	Date	Flag	User	Text
0	1467810369	Mon Apr 06 22:19:45 PDT 2009	NO_QUERY	_TheSpecialOn	@switchfoot http://twitpic.com/2y1zl - Awww, tl
0	1467810672	Mon Apr 06 22:19:49 PDT 2009	NO_QUERY	scotthamilton	is upset that he can't update his Facebook by text
0	1467810917	Mon Apr 06 22:19:53 PDT 2009	NO_QUERY	mattycus	@Kenichan I dived many times for the ball. Mana
0	1467811184	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	ElleCTF	my whole body feels itchy and like its on fire
0	1467811193	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	Karoli	@nationwideclass no, it's not behaving at all. i'm
0	1467811372	Mon Apr 06 22:20:00 PDT 2009	NO_QUERY	joy_wolf	@Kwesidei not the whole crew
0	1467811592	Mon Apr 06 22:20:03 PDT 2009	NO_QUERY	mybirch	Need a hug
0	1467811594	Mon Apr 06 22:20:03 PDT 2009	NO_QUERY	coZZ	@LOLTrish hey long time no see! Yes.. Rains a bit
0	1467811795	Mon Apr 06 22:20:05 PDT 2009	NO_QUERY	2Hood4Hollyw	@Tatiana_K nope they didn't have it
0	1467812025	Mon Apr 06 22:20:09 PDT 2009	NO_QUERY	mimismo	@twittera que me muera ?

Figure 1. A sample from the dataset

The dataset has a dimension of 1600000×2 after necessary data reduction has been applied(It can be seen in Figure 2).

	label	tweet
0	Negative	@switchfoot http://twitpic.com/2y1zl - Awww, t...
1	Negative	is upset that he can't update his Facebook by ...
2	Negative	@Kenichan I dived many times for the ball. Man...
3	Negative	my whole body feels itchy and like its on fire
4	Negative	@nationwideclass no, it's not behaving at all....
...	...	...
1599995	Positive	Just woke up. Having no school is the best fee...
1599996	Positive	TheWDB.com - Very cool to hear old Walt interv...
1599997	Positive	Are you ready for your MoJo Makeover? Ask me f...
1599998	Positive	Happy 38th Birthday to my boo of alll time!!! ...
1599999	Positive	happy #charitytuesday @theNSPCC @SparksCharity...

Figure 2. Dataset

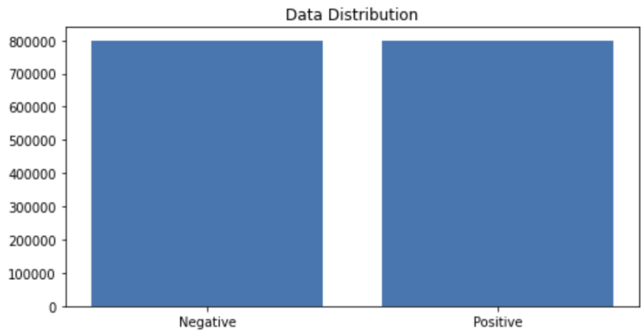
The features/attributes of the dataset is as follows after data reduction is applied:

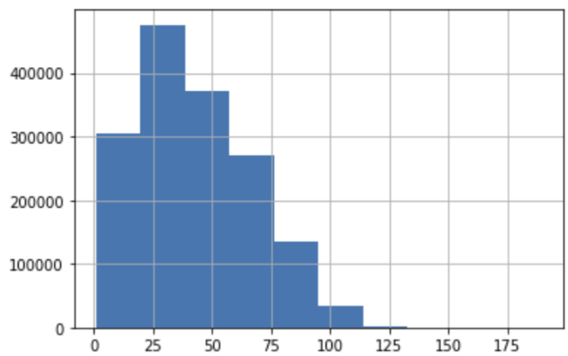
#	Feature Name	Description	Type	# of values	Missing Values %
1	label	Negative or Positive	nominal	1600000	%0.4795
2	tweet	Tweets	text	1600000	%0.4795

Figure 3. Dataset features/attributes.

We remove tweets that have a length of 0. After this process, the dataset has a dimension of 1592328×2

Positive and negative samples are equal. The dataset distribution has not any skewness as shown in Figure 4.






```
count    1.592328e+06
mean     4.279740e+01
std      2.415896e+01
min      1.000000e+00
25%      2.300000e+01
50%      3.900000e+01
75%      6.000000e+01
max      1.890000e+02
dtype: float64
```

Figure 4. Dataset distribution

Number of Letters

We provide the frequency and the relative frequency of the letters of the whole tweets. Finally, we will apply a chi-square test to test if the distribution of the letters in tweets is the same with what we see in English texts.

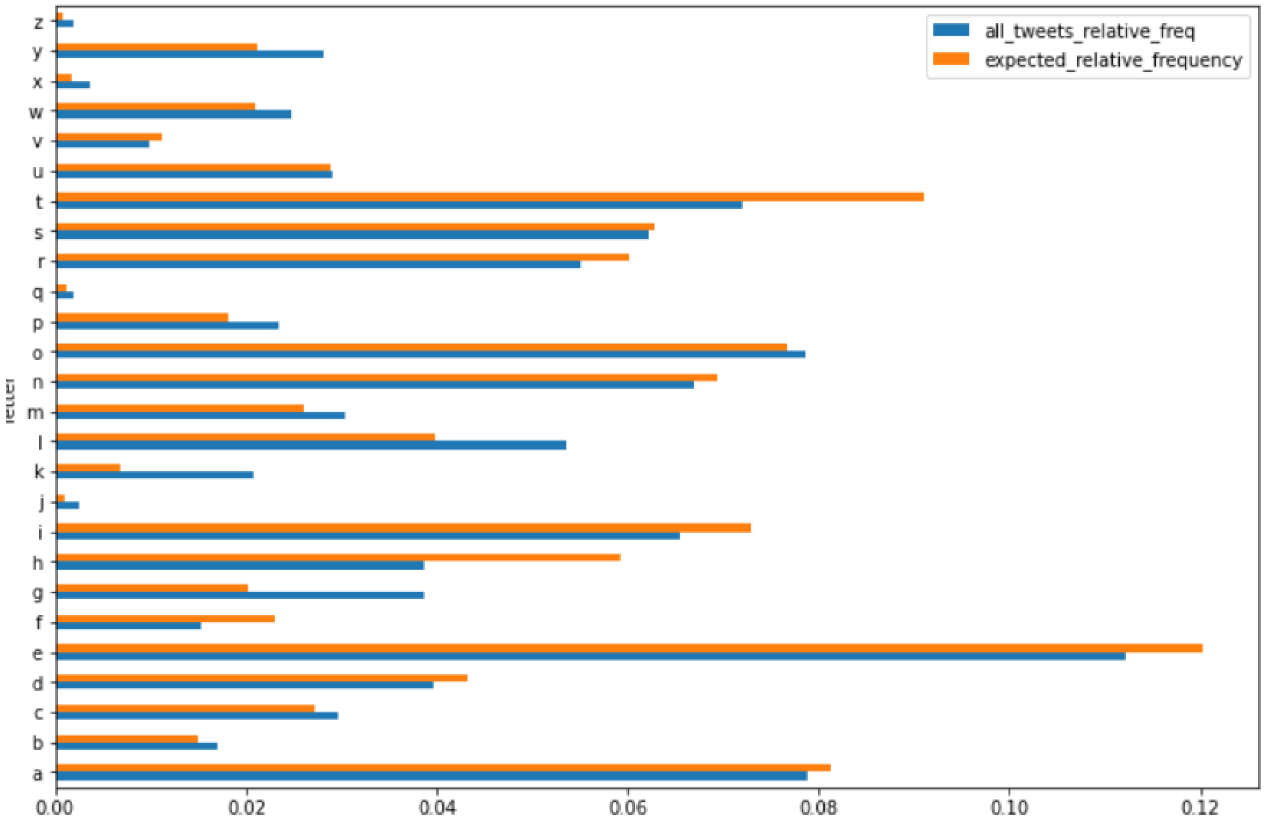


Figure 5. Letter frequencies of each 26 characters in English Alphabet.

	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 6. 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

	frequency	expected
frequency	1.000000	0.967421
expected	0.967421	1.000000

Figure 7. Correlation.

We counted the number of characters for each tweet 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 9.

	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 8. 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 9. 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 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 11.

	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 10. 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 11. Max, min, mean and standard deviation of each tweet in terms of number of words.

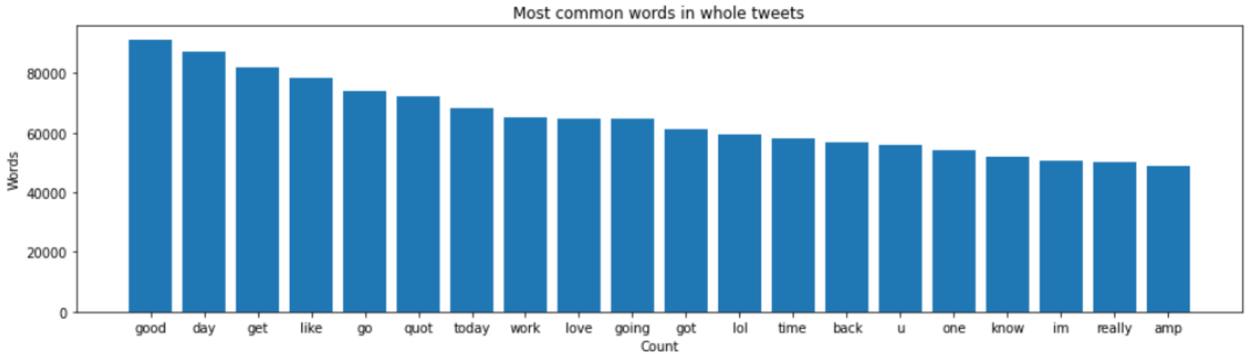


Figure 11. Most common words in our dataset.

Positive Tweets

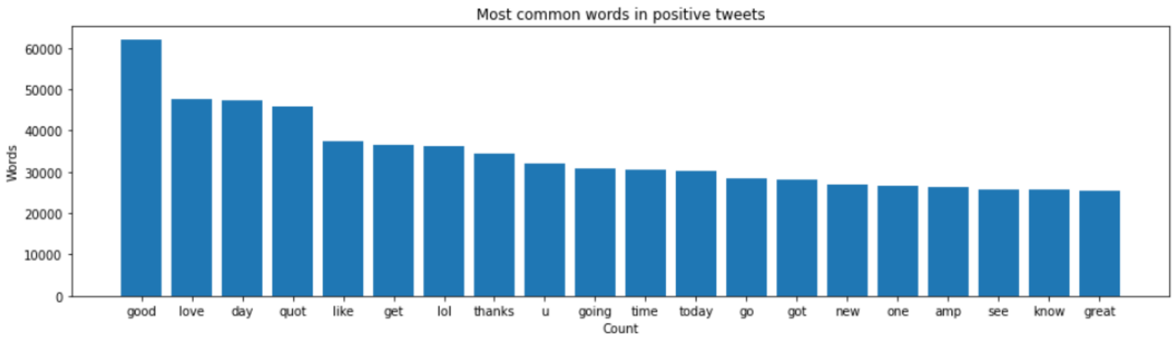


Figure 12. Most common words in positive tweets in our dataset.

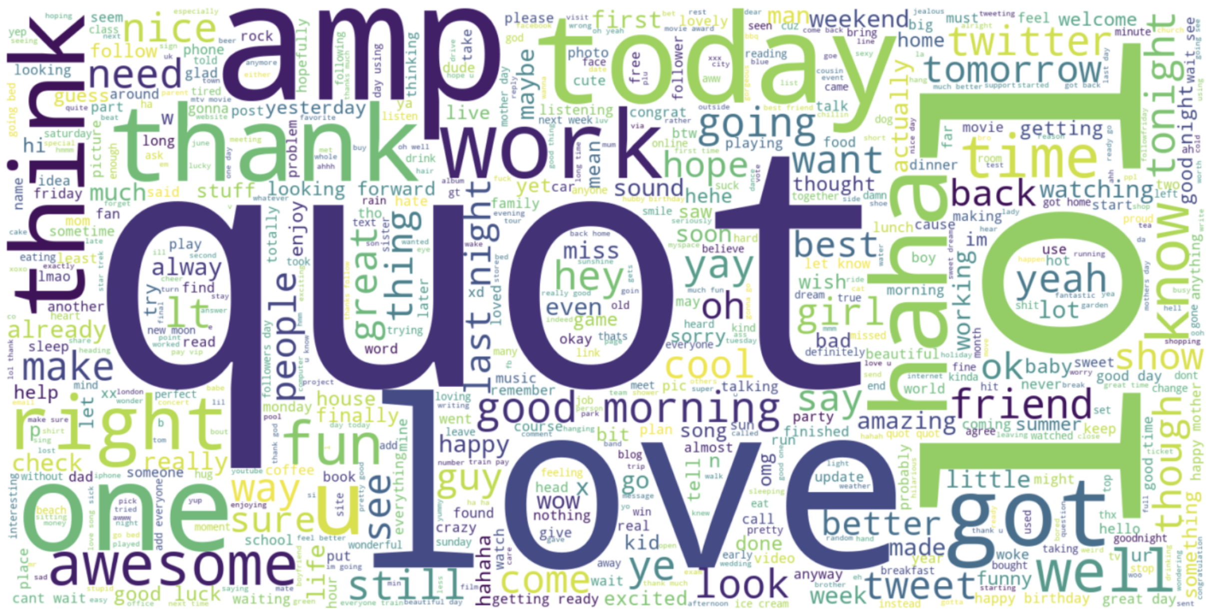


Figure 13. Word cloud of positive tweets.

Negative Tweets

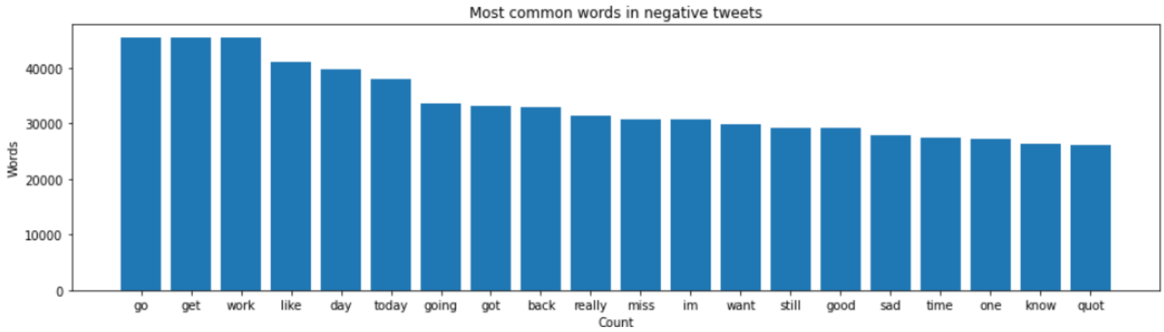


Figure 14. Most common words in negative tweets in our dataset.

Figure 15. Word cloud of positive tweets.

GloVe: Global Vectors for Word Representation [2]

We can train the embedding ourselves. However, that approach can take a long time to train. So, we use transfer learning technique, and we use GloVe: Global Vectors for Word Representation.

The Global Vectors for Word Representation, or GloVe, algorithm is an extension to the word2vec method for efficiently learning word vectors, developed by Pennington, et al. at Stanford. It is an unsupervised learning algorithm for obtaining vector representations for words. Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space.

We download the GloVe. Then we initialize an embedding index that has 400000 word vectors, and embedding matrix.

References

[1] Sentiment140, http://help.sentiment140.com/home

[2] GloVe: Global Vectors for Word Representation, Jeffrey Pennington, Richard Socher, Christopher D. Manning