

CSE 4062.1 SPRING 2021

INTRODUCTION TO DATA SCIENCE AND ANALYTICS

Project Delivery #5:
Descriptive Analysis

Project Title:
Social Media Sentiment Analysis

Group #3
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)



MARMARA
UNIVERSITY

Features

We use Term Frequency-Inverse Document Frequency (TF-IDF) to transform the text data. You can obtain the tf-idf array from Figure 1.

	00	000	0000	002	00am	00pm	01	02	026	02am	...	½sklov	½ssen	½sunday	½t	½tiei	½tobe	½u	½ve	½y	½i
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Figure 1. tf-idf array.

We used the Elbow method to make sure we choose the optimal number of clusters. We decided to make experiments 2 and 3 number of clusters.

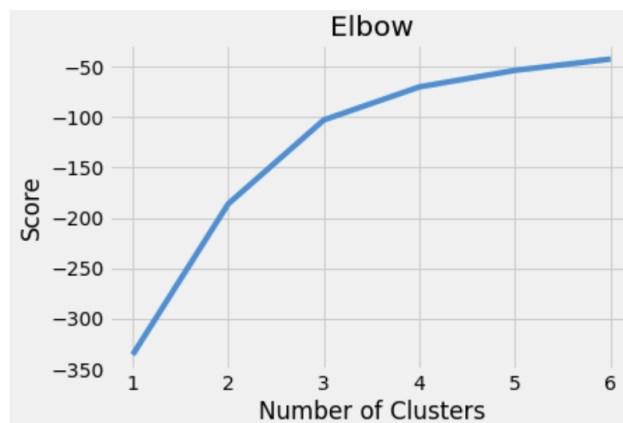
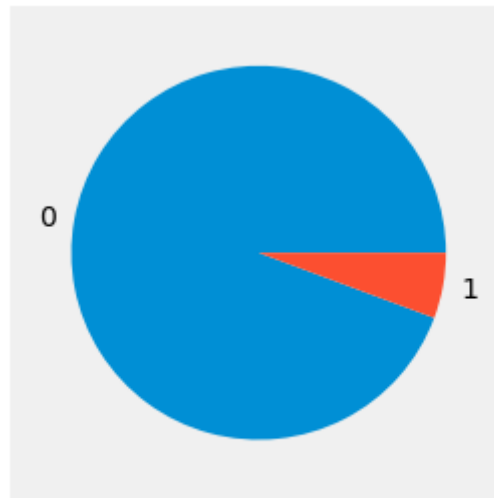


Figure 2. Elbow method to get optimal number of clusters.

Instance Distributions Pie Chart

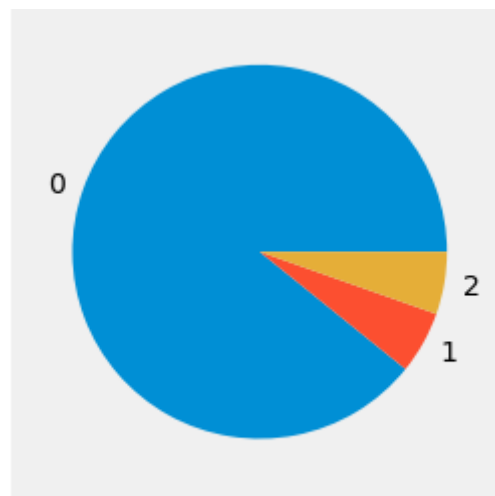


28300 1700

Cluster 0 Percentage = 94.3%

Cluster 1 Percentage = 5.7%

Figure 3. A pie chart showing the instance distributions for 2 clusters.



26584 1609 1609

Cluster 0 Percentage = 88.6%

Cluster 1 Percentage = 5.7%

Cluster 2 Percentage = 5.7%

Figure 4. A pie chart showing the instance distributions for 3 clusters.

Evaluation of Clustering Experiments

- Experiment 1 - Number of clusters = 2

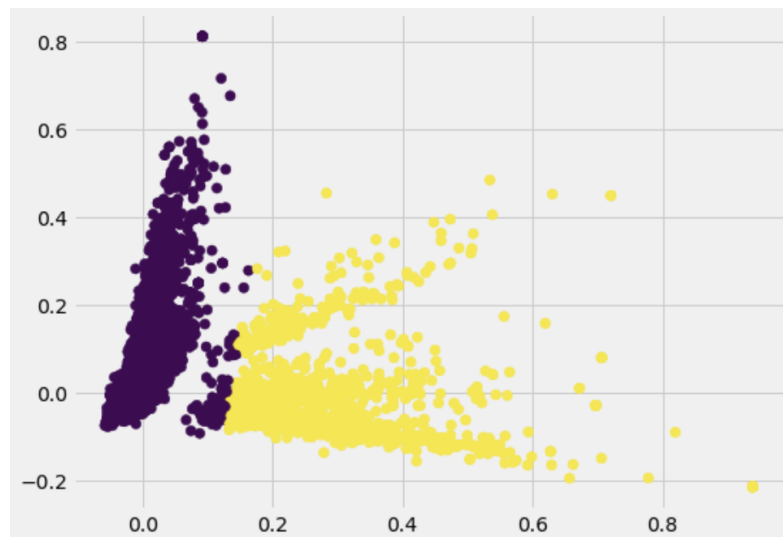


Figure 5. 2 Clusters

init	time	inertia	homo	compl	v-meas	ARI	AMI	NMI	silhouette
k-means++	0.093s	38122	0.973	0.970	0.971	0.991	0.971	0.971	0.814
random	0.108s	38122	0.975	0.970	0.972	0.991	0.972	0.972	0.749
PCA-based	0.050s	38985	0.011	0.010	0.010	0.068	0.010	0.010	0.723

Figure 6. Evaluation metrics for 2 clusters.

Most important words in Cluster 0:

	word	score
0	just	0.015132
1	day	0.012346
2	today	0.011476
3	like	0.010374
4	want	0.010060
5	going	0.010016
6	don	0.009887
7	really	0.009350
8	got	0.009332
9	sad	0.008994
10	good	0.008851
11	miss	0.008415
12	time	0.008402
13	know	0.008327
14	im	0.008257
15	wish	0.008104
16	home	0.008088
17	sorry	0.007745
18	sleep	0.007660
19	night	0.007330

Most important words in Cluster 1:

	word	score
0	work	0.302107
1	tomorrow	0.028365
2	day	0.027841
3	today	0.027249
4	going	0.023979
5	ready	0.017308
6	time	0.015752
7	home	0.015085
8	got	0.014318
9	want	0.014249
10	morning	0.014045
11	bed	0.013746
12	getting	0.013460
13	don	0.012802
14	just	0.012368
15	tired	0.011848
16	night	0.011170
17	sleep	0.010848
18	gotta	0.010505
19	hours	0.010329

- Experiment 2 - Number of clusters = 3

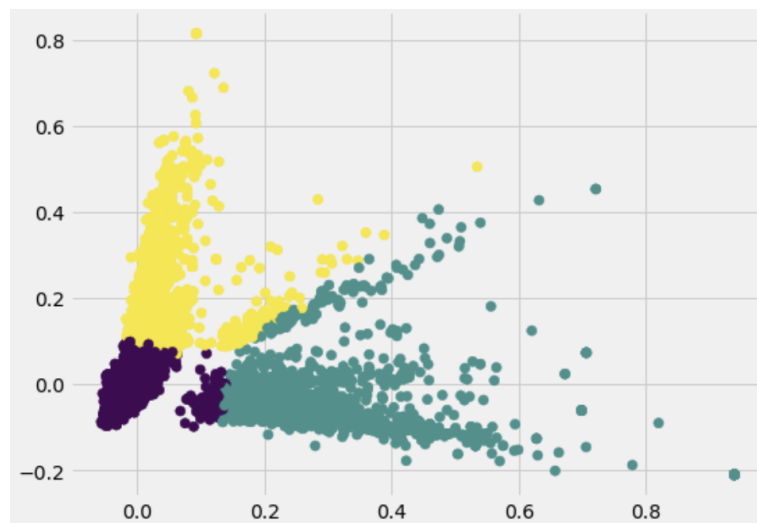


Figure 7. 3 Clusters

init	time	inertia	homo	compl	v-meas	ARI	AMI	NMI	silhouette
k-means++	0.172s	49102	0.455	0.426	0.440	0.471	0.440	0.440	0.681
random	0.210s	49102	0.454	0.424	0.438	0.470	0.438	0.438	0.717
PCA-based	0.062s	49102	0.455	0.425	0.439	0.471	0.439	0.439	0.662

Figure 8. Evaluation metrics for 2 clusters.

Most important words in Cluster 0: Most important words in Cluster 1: Most important words in Cluster 2:

word	score	word	score	word	score
0 just 0.015528		0 work 0.308336		0 day 0.199634	
1 like 0.010524		1 tomorrow 0.027660		1 today 0.065946	
2 want 0.009992		2 today 0.025722		2 school 0.059420	
3 don 0.009947		3 going 0.024375		3 tomorrow 0.057504	
4 got 0.009460		4 day 0.018545		4 going 0.028303	
5 really 0.009307		5 ready 0.017846		5 good 0.017077	
6 sad 0.008864		6 time 0.016309		6 long 0.015431	
7 going 0.008797		7 home 0.015314		7 beautiful 0.013727	
8 miss 0.008738		8 got 0.014329		8 break 0.013477	
9 know 0.008549		9 want 0.014295		9 bad 0.012683	
10 im 0.008349		10 morning 0.014260		10 home 0.012522	
11 good 0.008327		11 getting 0.014222		11 bed 0.012221	
12 time 0.008284		12 bed 0.014176		12 want 0.011225	
13 wish 0.008086		13 don 0.013260		13 morning 0.010797	
14 sorry 0.008076		14 just 0.012587		14 sad 0.010686	
15 today 0.007920		15 tired 0.012140		15 feeling 0.010684	
16 home 0.007797		16 sleep 0.011213		16 work 0.010277	
17 sleep 0.007639		17 night 0.010812		17 spring 0.010034	
18 need 0.007373		18 hours 0.010564		18 time 0.010008	
19 night 0.007292		19 need 0.010554		19 really 0.009986	

Result

K-means is a very simple and powerful algorithm to cluster a dataset. However, one of the problems is that clusters are spherical. Therefore, it can not be reliable for all situations.

We are using text data for our project. So, we need to represent the data as the model understands. For this reason, firstly, we vectorize our data with tf-idf vectorizer. Then, we use the elbow method to make sure we choose the optimal number of clusters. We decided to make experiments with 2 and 3 numbers of clusters.

Therefore, we have two different experiments with 2 and 3 clusters, we have 2 different instance distributions pie charts. For two clusters, we can see that Cluster-0 has a really huge ratio, with 94.3%, in 30,000 instances. When we applied the experiment with 3 clusters, we can see that Cluster-1 did not lose any instances, but Cluster-0 is split into two. Cluster-2 has an equal ratio with Cluster-1.

We compare three approaches kmeans++, random initialization and initialization based on PCA projection for 2 and 3 numbers of clusters. Evaluation metrics for each 2 experiments as shown in Figure 6 and Figure 8.

We score the words in each cluster in order of importance. In Experiment 1 (number of cluster is 2), we obtain most important words of Cluster-0 are like, good, wish, sorry. They can be said mostly positive words. On the other hand, Cluster-1's most important words are not distinguishable.

In Experiment 2 (number of cluster is 3), Cluster-0 has mostly positive words such as good, like, miss, sorry, wish. Cluster-1 seems like neutral and Cluster-2 has some negative words such as bad and sad.

The K-means is clustering words according to some semblance of meaning in our experiments, but experiments can be developed with even more accurate parameters.