Flip 01 Project Report

Jing Miao

Qingdao University of Technology

January 10th

Overview

Introduction

Project Introduction

With all of the tweets circulating every second it is hard to tell whether the sentiment behind a specific tweet will impact a company, or a person's, brand for being viral (positive), or devastate profit because it strikes a negative tone. Capturing sentiment in language is important in these times where decisions and reactions are created and updated in seconds. The purpose of this task is to detect hate speech in tweets. For the sake of simplicity, if there is racist or sexist sentiment on Twitter, we will say that it contains hate speech. Therefore, the task is to classify racist or sexist tweets from other tweets. If there is a system that can detect this type of text, it will definitely make the Internet and social media a better, non-malicious communication space.

Data Analysis

Statistical Analysis Data

Table 1: The First Five Rows of The Train Data

	textID	text	selected_text	sentiment
0	cb774db0d1	I‘dhaveresponded,ifIweregoing	I‘dhaveresponded,ifIweregoing	neutral
1	549e992a42	SoooSADIwillmissyouhereinSanDiego!!!	SoooSAD	negative
2	088c60f138	mybossisbullyingme...	bullyingme	negative
3	9642c003ef	whatinterview!leavemealone	leavemealone	negative
4	358bd9e861	Sonsof ∗ ∗ ∗ ∗,whycouldn‘ttheyputthemont...	Sonsof ∗ ∗ ∗ ∗,	negative

Table 2: The First Five Rows of The Test Data

	textID	text	sentiment
0	f87dea47db	Lastsessionofthedayhttp://twitpic.com/67ezh	neutral
1	96d74cb729	Shanghaiisalsoreallyexciting(precisely−...	positive
2	eee518ae67	RecessionhitVeroniqueBranquinho,shehasto...	negative
3	01082688c6	happybday!	positive
4	33987a8ee5	http://twitpic.com/4w75p−Ilikeit!!	positive

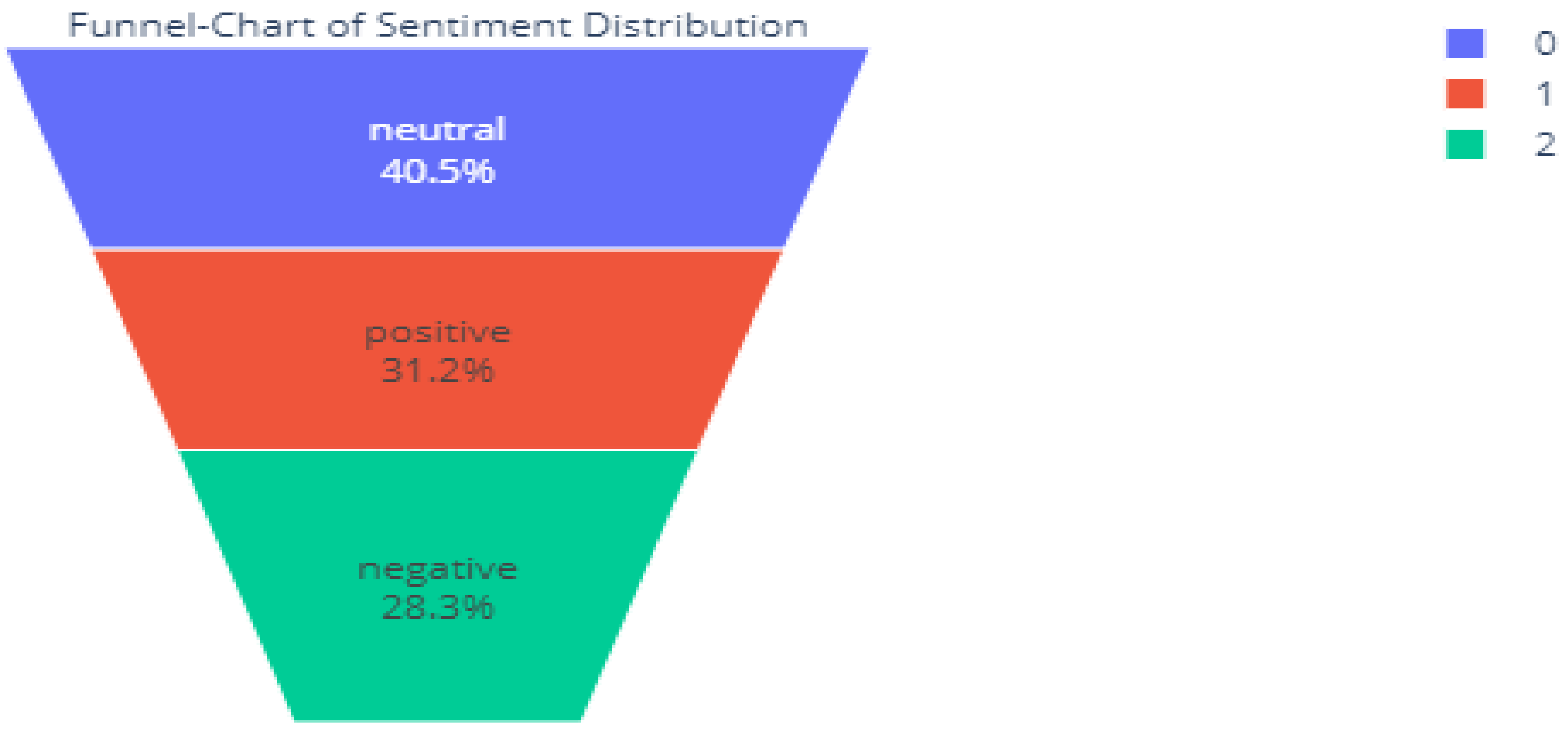
Statistical Analysis Data

Table 3: Train Data Describe

	textID	text	selected_text	sentiment
count	27481	27480	27480	27481
unique	27481	27480	22463	3
top	703*d*8*ea*662	*hiiiimonmyipod...icantfallasleep*	*good*	*neutral*
freq	1	1	199	11118

Data Visualization

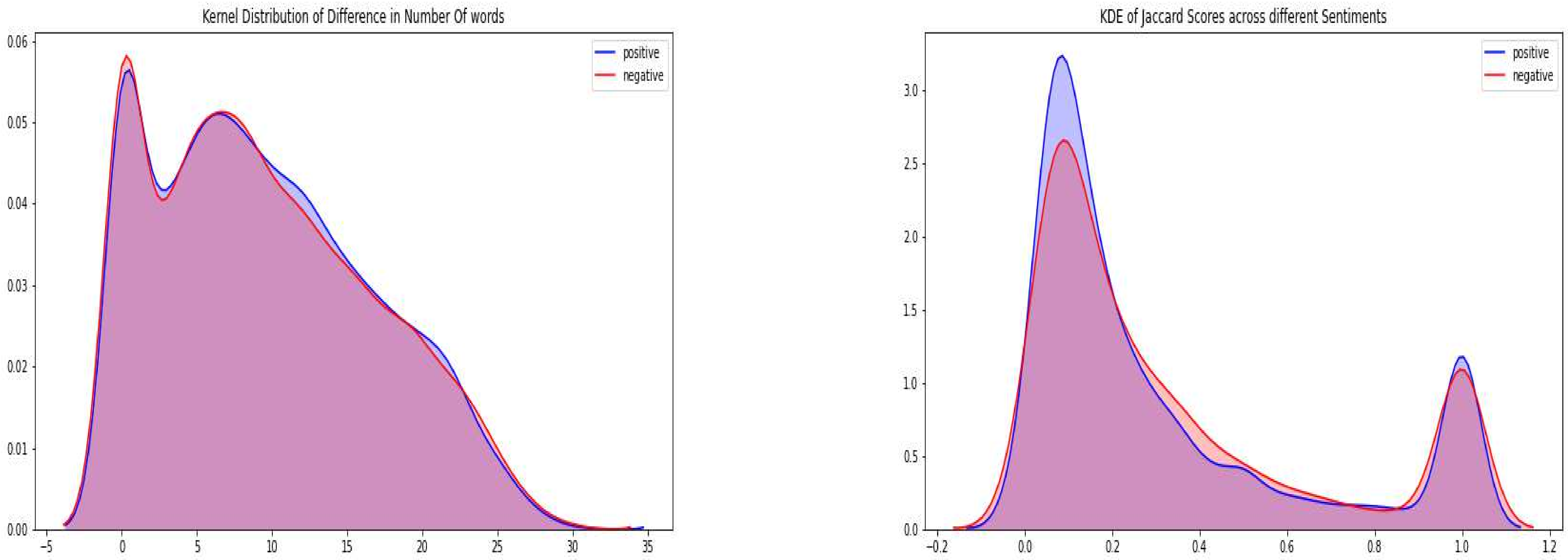
- To draw a Funnel-Chart for better visualization



- From the above Funnel-Chart,the neutral sentiment accounted for the majority, followed by the positive and the least negative.

Data Visualization

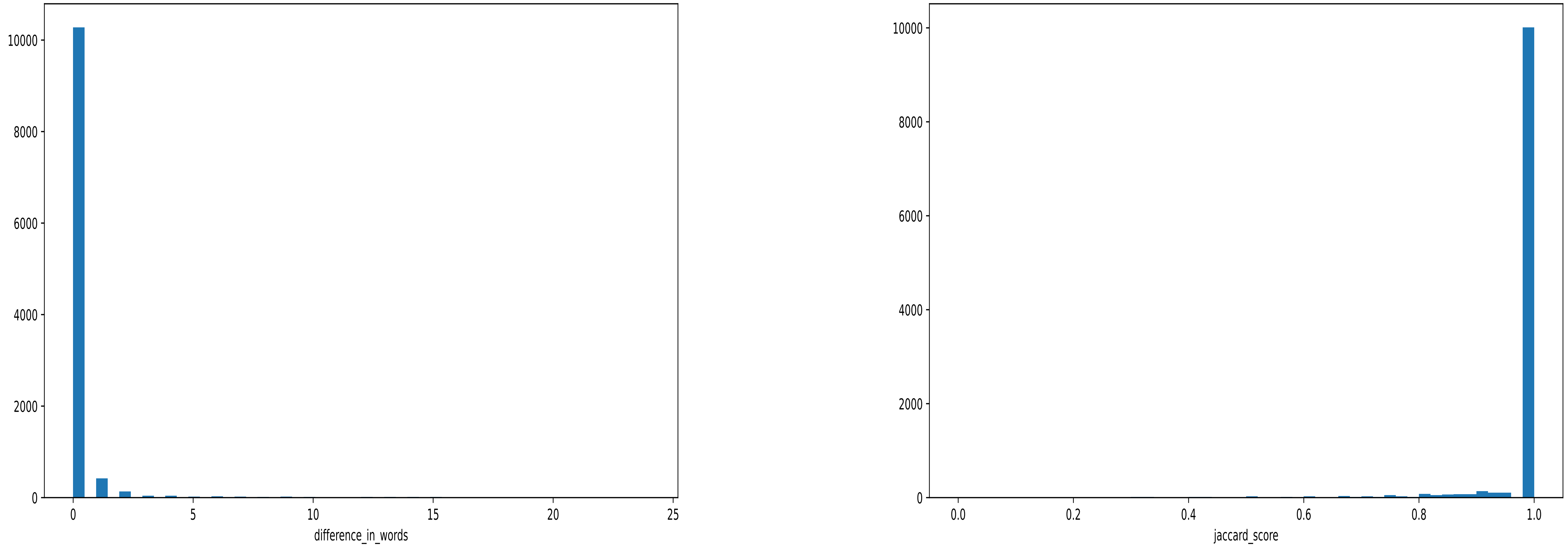
- So further analysis is needed to check the difference in Number of Words and the Jaccard Scores similarity across Different Sentiments between Text and selected_text.



(a) Negative and Positive Sentiment

(b) Negative and Positive Sentiment

Data Visualization



(c) Difference in Number of Words in Neutral Sentiment

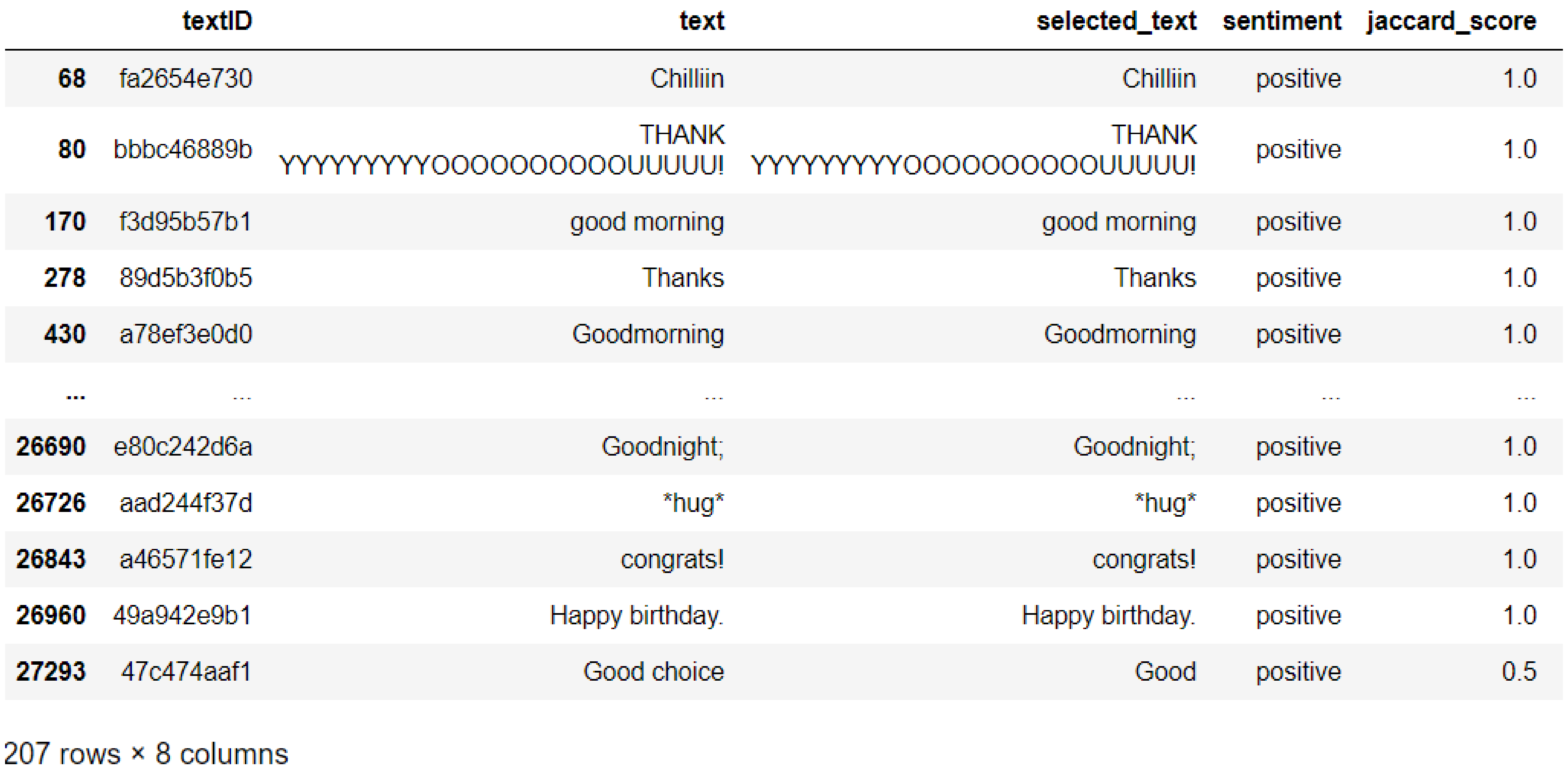
(d) Jaccard Scores in Neutral Sentiment

Conclusion Of EDA

- It can be seen that the number of tweets with a jaccard similarity of 1 between text extraction and text is more neutral sentiment.In conclusion maybe we can use neutral "text" as it is for "selected_text" in test data submission.
- We can see from the Jaccard Score Plot that there is peak for negative and positive plot around score of 1 .That means there is a cluster of tweets where there is a high similarity between text and selected texts ,if we can find those clusters then we can predict text for selected texts for those tweets irrespective of segment.

Data Visualization

- View the Jaccard value of positive tweets with words less than or equal to 2



	textID	text	selected_text	sentiment	jaccard_score
68	fa2654e730	Chilliin	Chilliin	positive	1.0
80	bbbc46889b	THANK YYYYYYYYYOOOOOOOOOOUUUUU!	THANK YYYYYYYYYOOOOOOOOOOUUUUU!	positive	1.0
170	f3d95b57b1	good morning	good morning	positive	1.0
278	89d5b3f0b5	Thanks	Thanks	positive	1.0
430	a78ef3e0d0	Goodmorning	Goodmorning	positive	1.0
...	...	...	...	...	...
26690	e80c242d6a	Goodnight;	Goodnight;	positive	1.0
26726	aad244f37d	*hug*	*hug*	positive	1.0
26843	a46571fe12	congrats!	congrats!	positive	1.0
26960	49a942e9b1	Happy birthday.	Happy birthday.	positive	1.0
27293	47c474aaf1	Good choice	Good	positive	0.5

207 rows × 8 columns

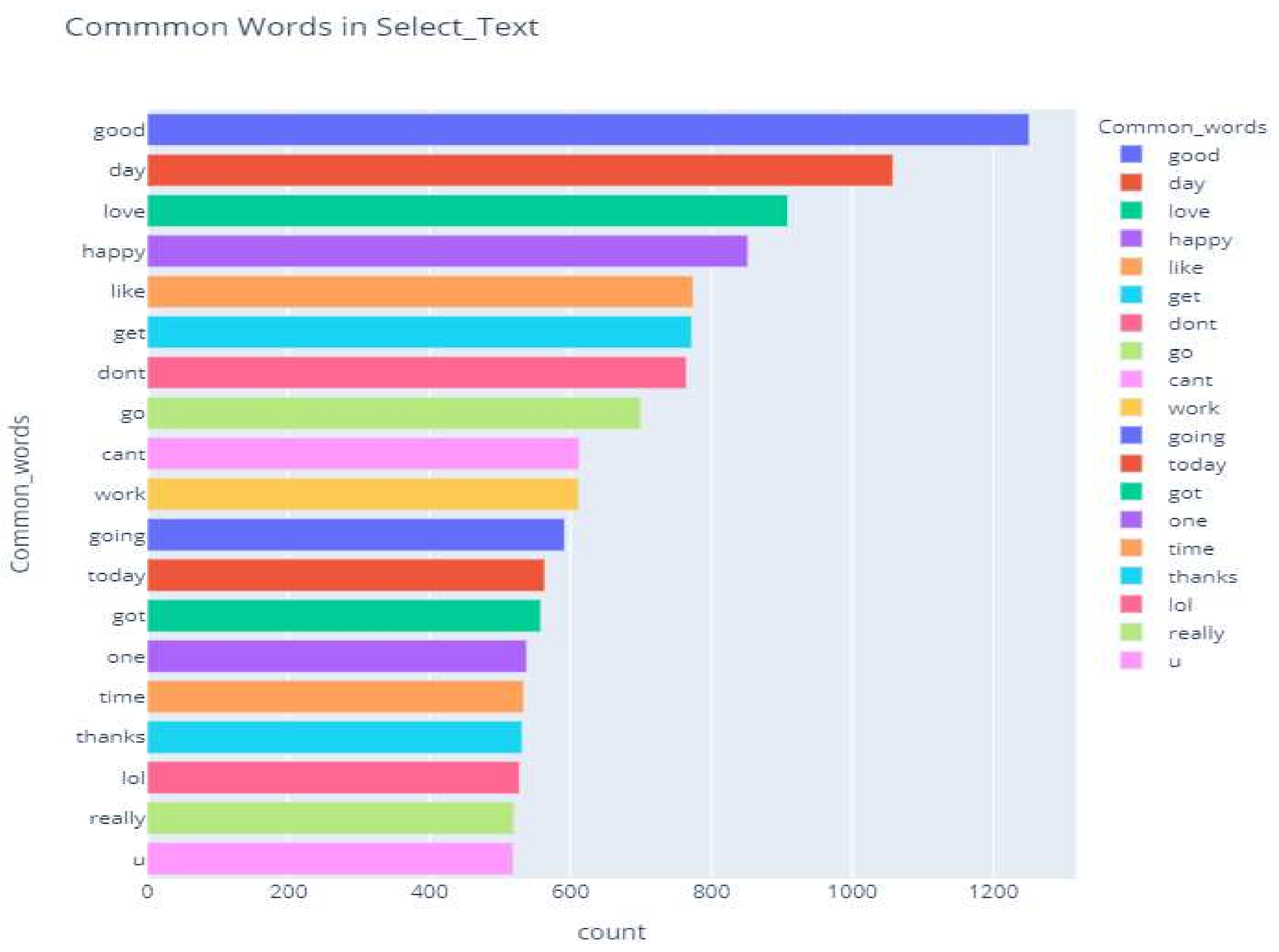
Data cleaning

Data cleaning

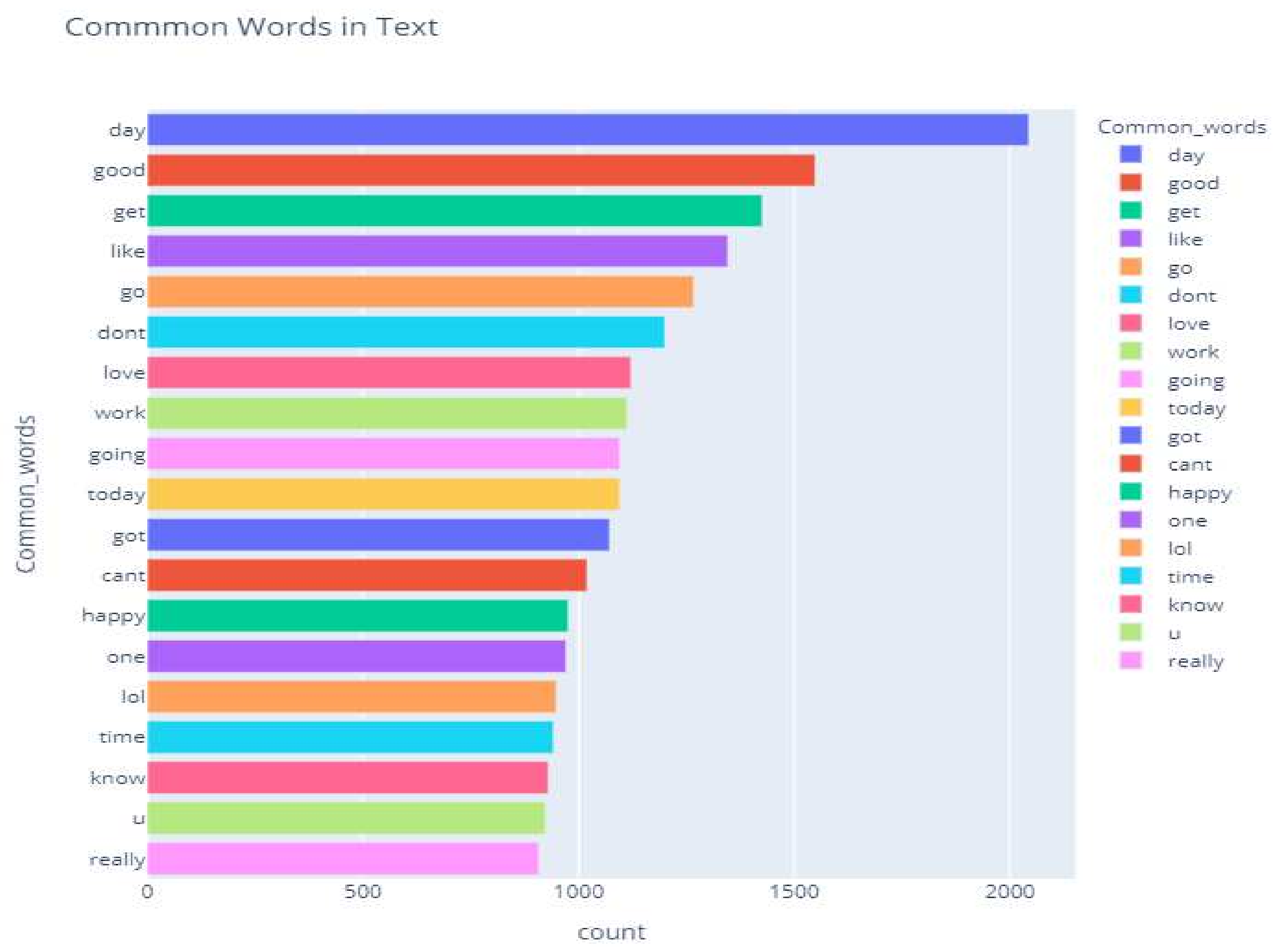
- First,make text lowercase,remove text in square brackets,remove links,remove punctuation and remove words containing numbers.
- Then to remove the stopwords.

Data cleaning

- We can see that the top 20 most common words and the text in the selected text are almost the same.



(e)

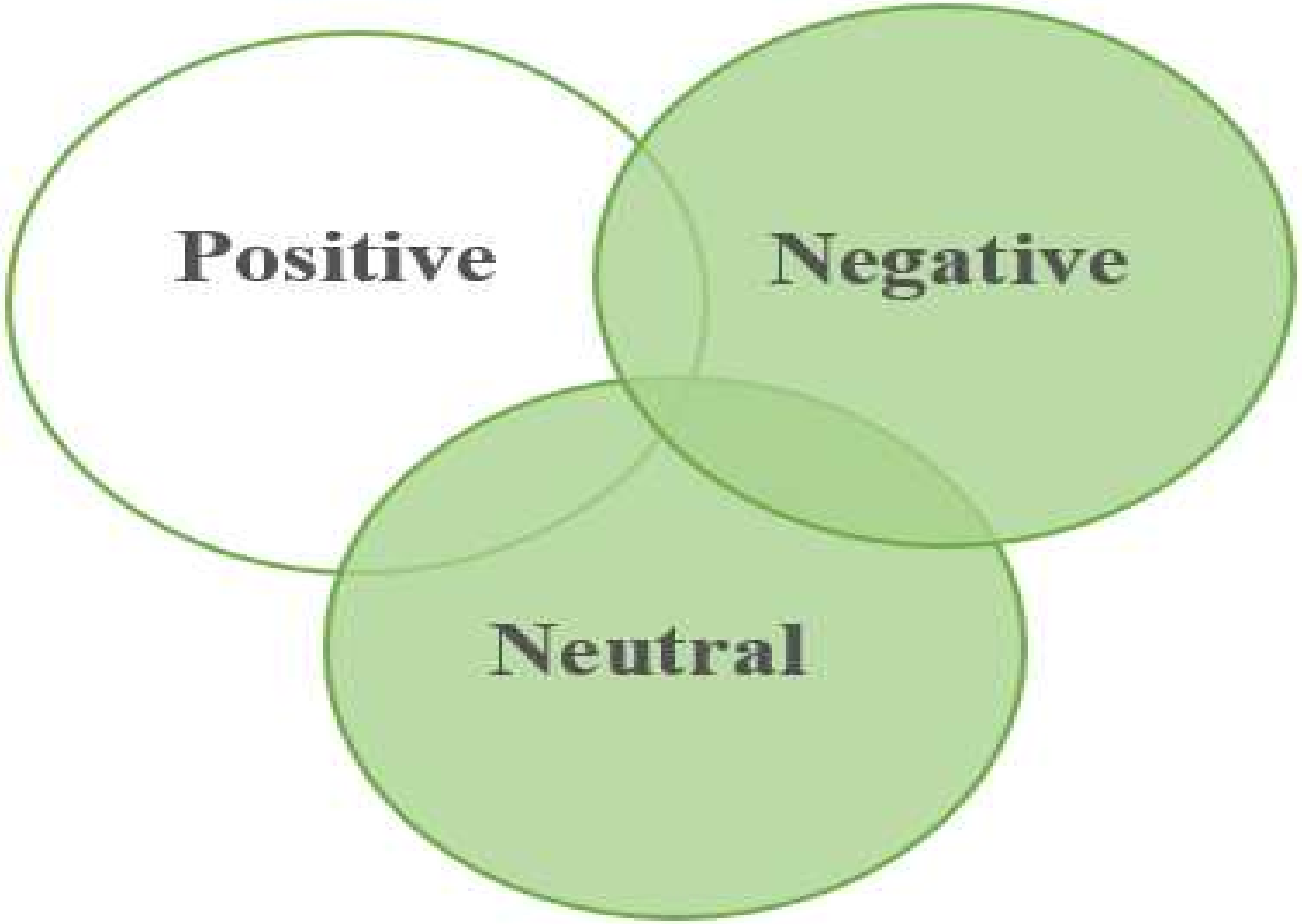


(f)

Data Feature Analysis

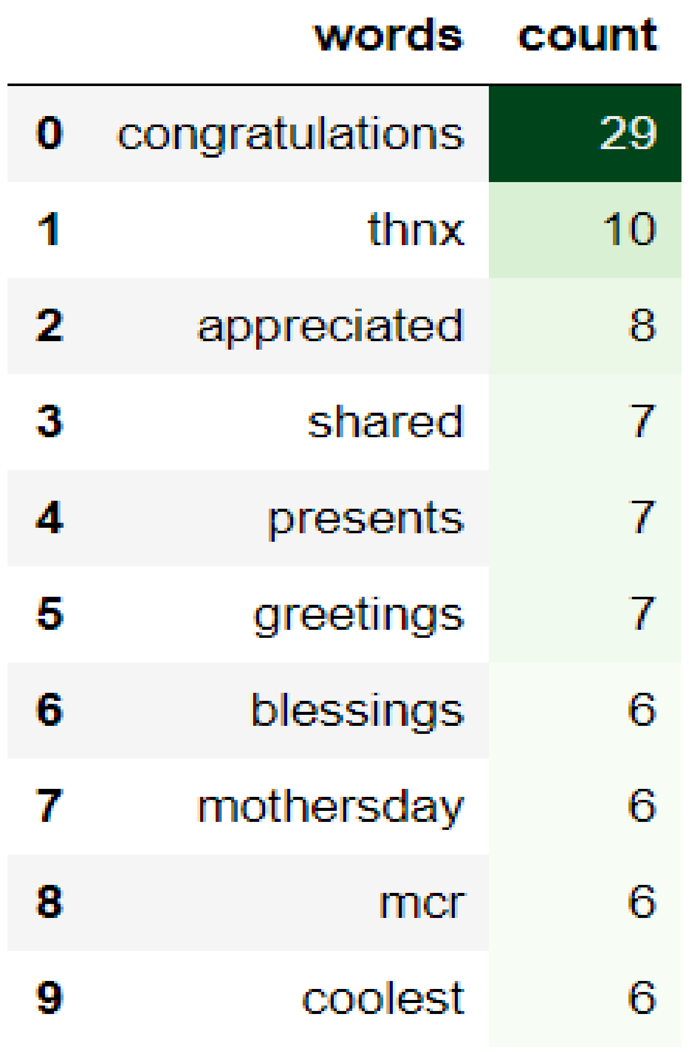
Data Feature Analysis

- Unique Words in each Segment



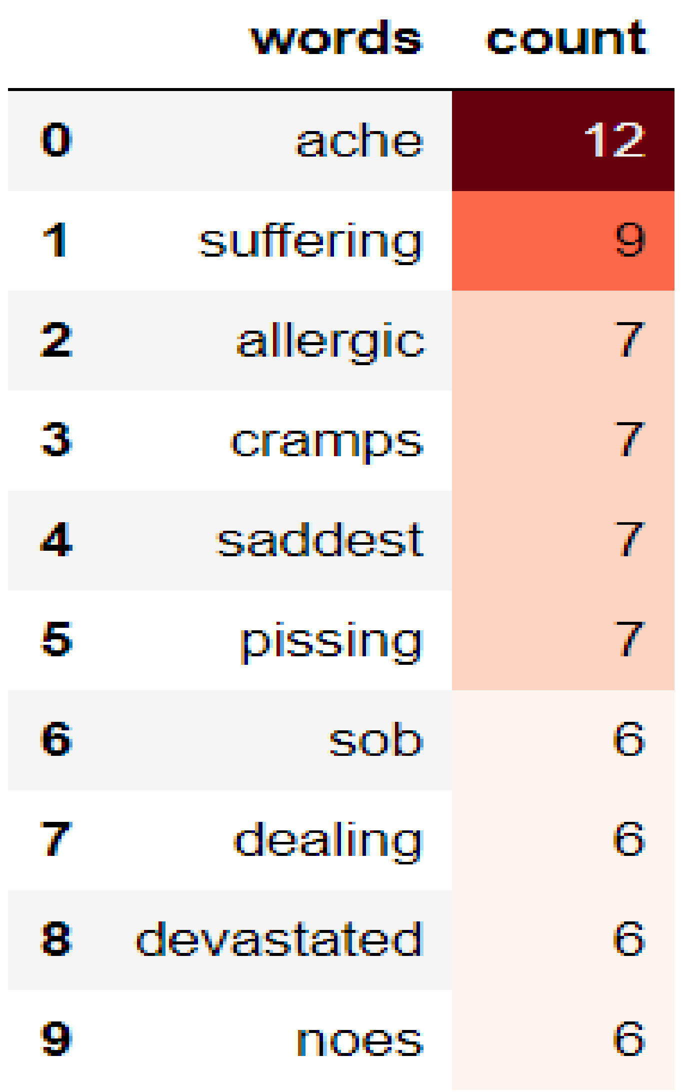
Data Feature Analysis

- Unique Words in each Segment
- By Looking at the Unique Words of each sentiment,we now have much more clarity about the data,these unique words are very strong determiners of Sentiment of tweets



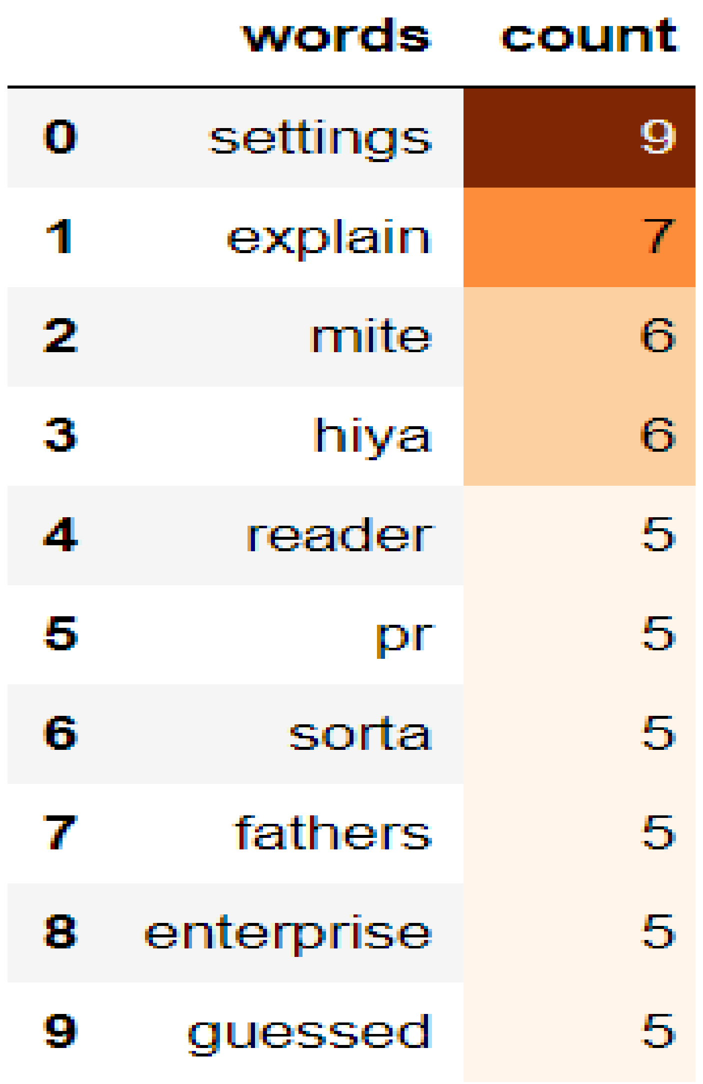
	words	count
0	congratulations	29
1	thnx	10
2	appreciated	8
3	shared	7
4	presents	7
5	greetings	7
6	blessings	6
7	mothersday	6
8	mcr	6
9	coolest	6

(g) Positive



	words	count
0	ache	12
1	suffering	9
2	allergic	7
3	cramps	7
4	saddest	7
5	pissing	7
6	sob	6
7	dealing	6
8	devastated	6
9	noes	6

(h) Negative

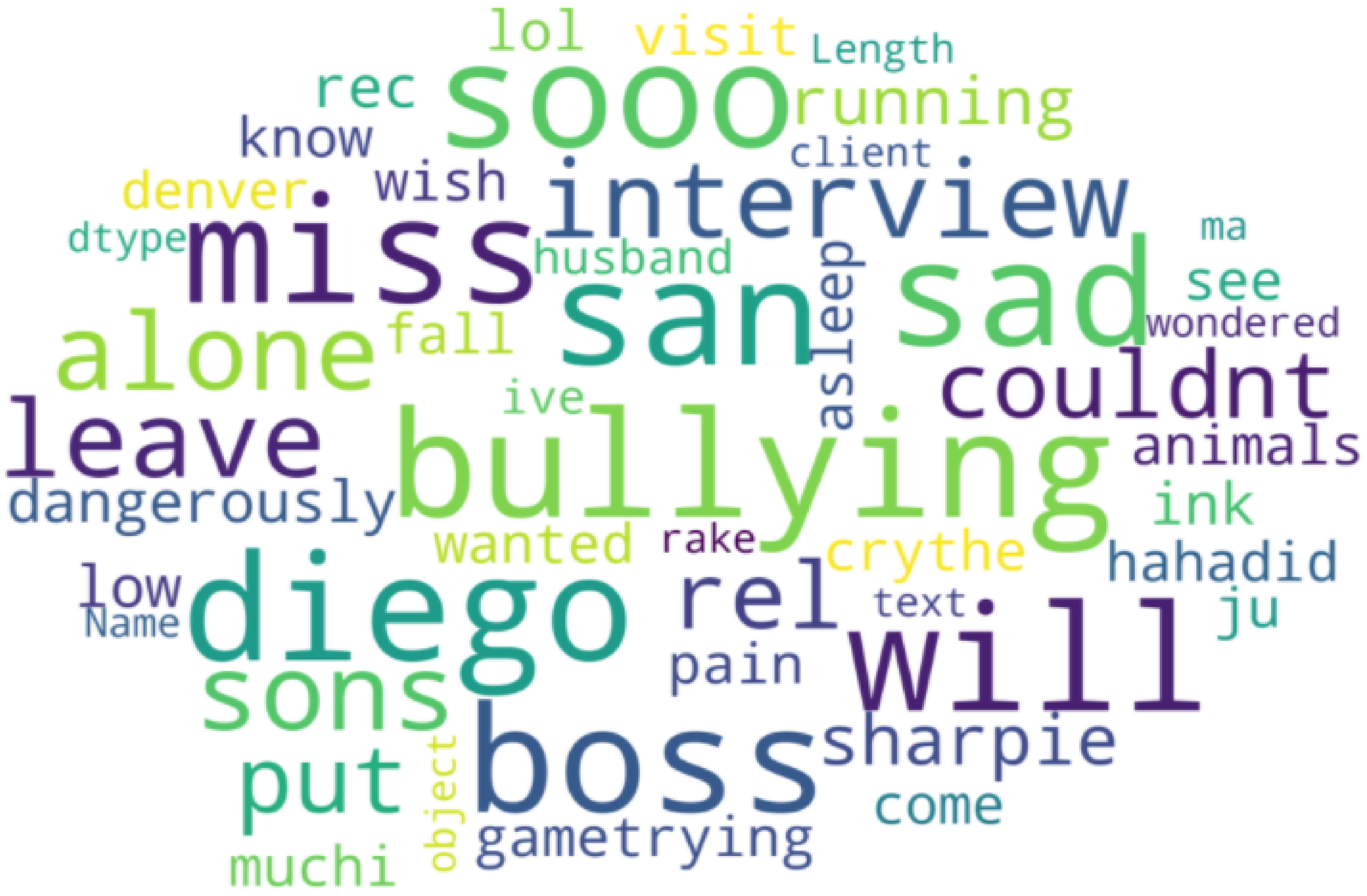


	words	count
0	settings	9
1	explain	7
2	mite	6
3	hiya	6
4	reader	5
5	pr	5
6	sorta	5
7	fathers	5
8	enterprise	5
9	guessed	5

(i) Neutral

WordClouds

- Word Cloud of Negative Tweets



Model

Model-NER

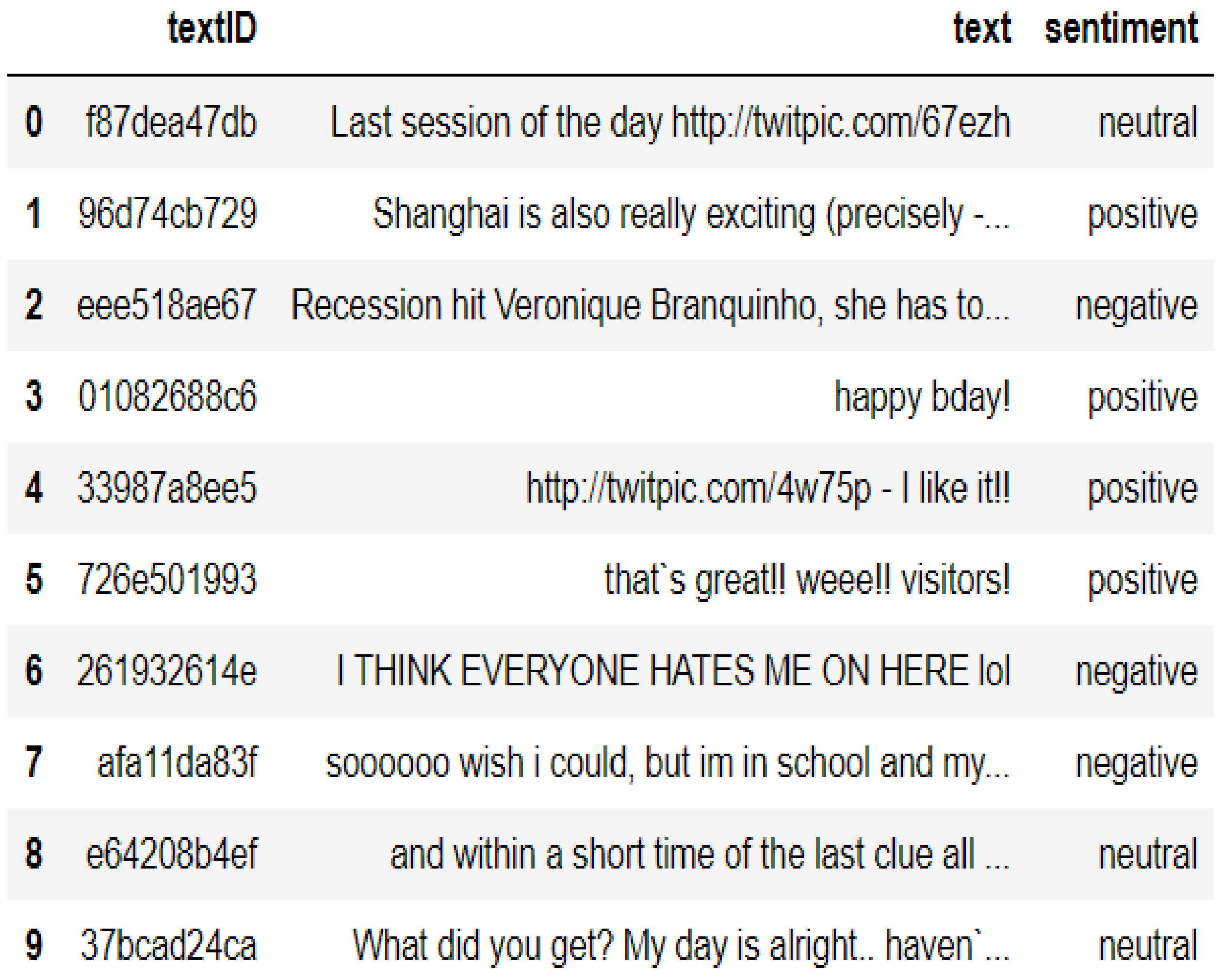
- To use text as selected_text for all neutral tweets due to their high jaccard similarity.
- To use text as selected_text for all tweets having number of words less than 3 in text.
- Train two different models for Positive and Negtive tweets.
- To use spacy for creating our own customised NER model or models (seperate for each Sentiment).

Train Model-NER

- Set up the pipeline and train the entity recognizer,create the built-in pipeline components and add them to the pipeline
- Add labels: entities
- get names of other pipes to disable them during training
- The sample can be divided into equal subsets, one gradient descent can be made for each subset, then the values of parameters W and B can be updated, and then the gradient descent can be continued in the next subset.
- Training models for Positive and Negative tweets.

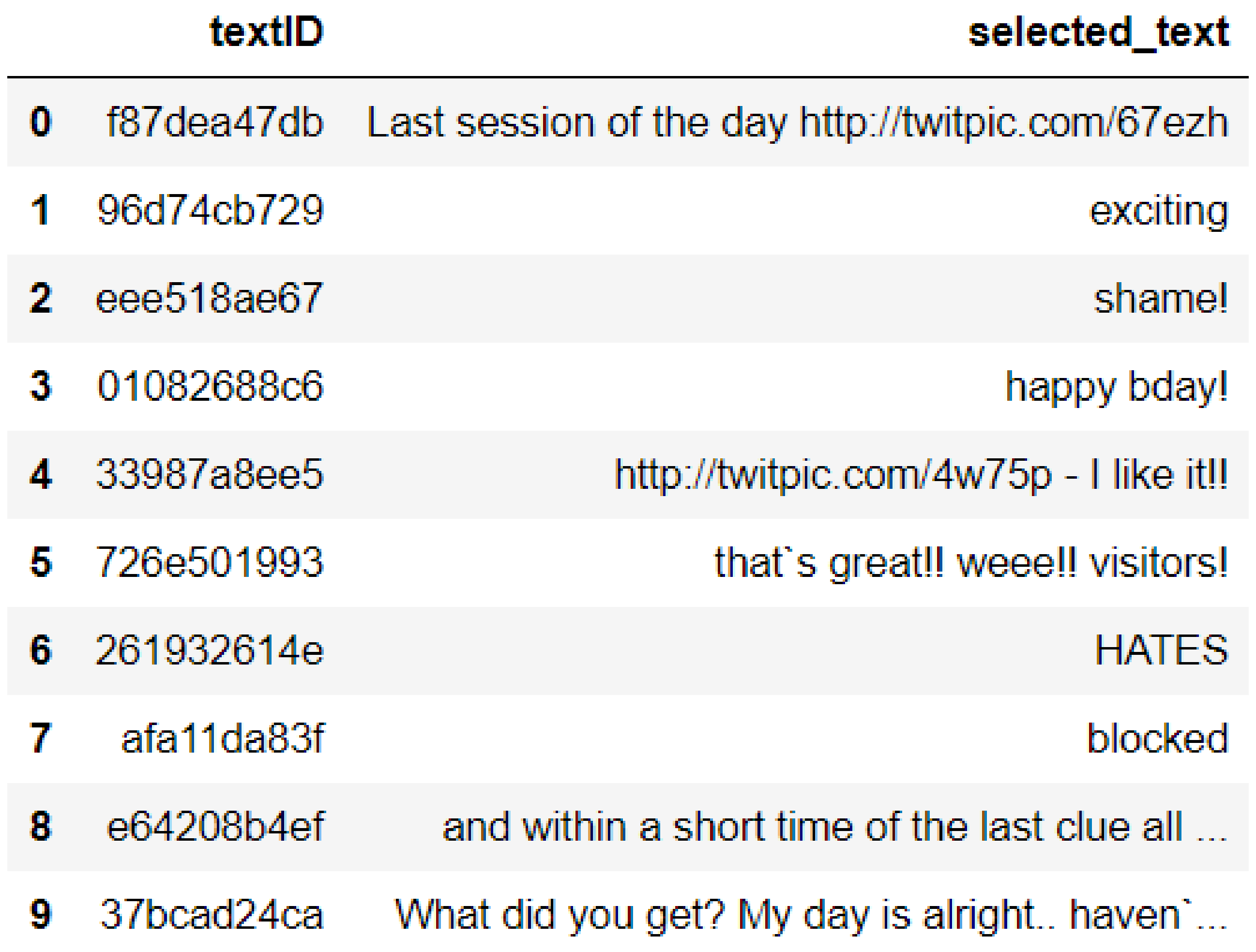
Test Model-NER

- Unique Words in each Segment



	textID	text	sentiment
0	f87dea47db	Last session of the day http://twitpic.com/67ezh	neutral
1	96d74cb729	Shanghai is also really exciting (precisely -...	positive
2	eee518ae67	Recession hit Veronique Branquinho, she has to...	negative
3	01082688c6	happy bday!	positive
4	33987a8ee5	http://twitpic.com/4w75p - I like it!!	positive
5	726e501993	that`s great!! weee!! visitors!	positive
6	261932614e	I THINK EVERYONE HATES ME ON HERE lol	negative
7	afa11da83f	soooooo wish i could, but im in school and my...	negative
8	e64208b4ef	and within a short time of the last clue all ...	neutral
9	37bcad24ca	What did you get? My day is alright.. haven`...	neutral

(j)



	textID	selected_text
0	f87dea47db	Last session of the day http://twitpic.com/67ezh
1	96d74cb729	exciting
2	eee518ae67	shame!
3	01082688c6	happy bday!
4	33987a8ee5	http://twitpic.com/4w75p - I like it!!
5	726e501993	that`s great!! weee!! visitors!
6	261932614e	HATES
7	afa11da83f	blocked
8	e64208b4ef	and within a short time of the last clue all ...
9	37bcad24ca	What did you get? My day is alright.. haven`...

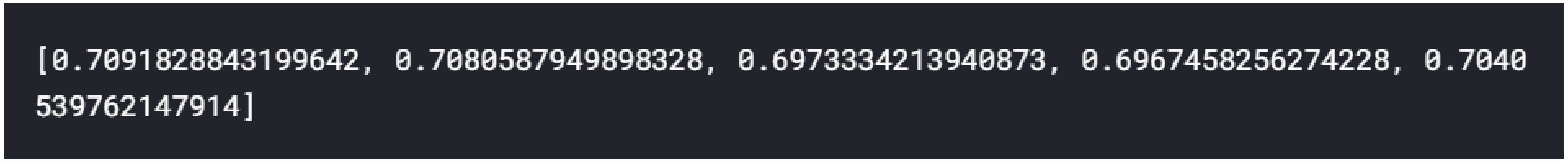
(k)

Model-BERT

- Data analysis, data visualization and data cleaning are all consistent with the above. Bert model is used for prediction, and the accuracy of this model is higher than that of NER model Z in natural language processing.

Train Model-BERT

- The Jaccard socre of training iteration 5 times.



```
[0.7091828843199642, 0.7080587949898328, 0.6973334213940873, 0.6967458256274228, 0.7040
539762147914]
```

Test Model-BERT

Table 4: There are 10 Random Comparison Results in The Test Results

	textID	text	selected_text
0	12005b65fc	Waitingformyturnonwiifitgymclosed	waitingformyturnonwiifitgymclosed
1	bcf13877f7	Goodmorningeveryone	good
2	575e4a89fe	ttsridiculouslysweetofyou	ridiculouslysweetofyou
3	a0b1828b67	Bridesalamode'powwowfirstthingthismorningTh...	lovely
4	472c3e2c41	Gettingsomewherewithmyfirst'real'KiokuDBandcatal...	gettingsomewherewithmyfirst'real'kiokudbandcata...
5	ce71d002ec	Mommasdayismay10th!Don‘tforgettodosomethingnice...	nice
6	8db4aaef4a	watchingthenotebook	watchingthenotebook
7	895de1648c	reallytired.andhavetoworkthewholedaytomorrow,t...	reallytired.
8	78d89e7c64	YeahprblypickinupsongsforSingStar.Haven‘tchecke...	yeahprblypickinupsongsforsingstar.haven‘tchecke...
9	756d255e40	isathomewithapukeyboy!Poorlittlebaby	poorlittlebaby

Thanks