D Y PATIL GROUP

"Empowerment through quality technical education"

Dr D Y Patil School of Engineering

D Y Patil Knowledge City, Charholi (Bk), Via Lohegaon, Pune 412105

Affiliated to S. P. Pune University Pune, MSBTE, Approved by AICTE & DTE

(Accredited by NAAC, NABL & ISO 9001:2015 & 21001:2018 Certified Institute)

Department of Computer Engineering

Mini Project on DSBDA Laboratory

Members:

TE_A_81 Omkar Nandode (PRN No. 72034728K)
TE_A_76 Muskan Sawant (PRN No. 72034784L)

Title

Use the following dataset and classify tweets into positive and negative tweets.
https://www.kaggle.com/ruchi798/data-science-tweets

Problem Definition:

Classification of tweets into positive and negative from the given csv file.

Prerequisite:

Basic concepts of Python Programming

Software Requirements:

Python

Outcomes:

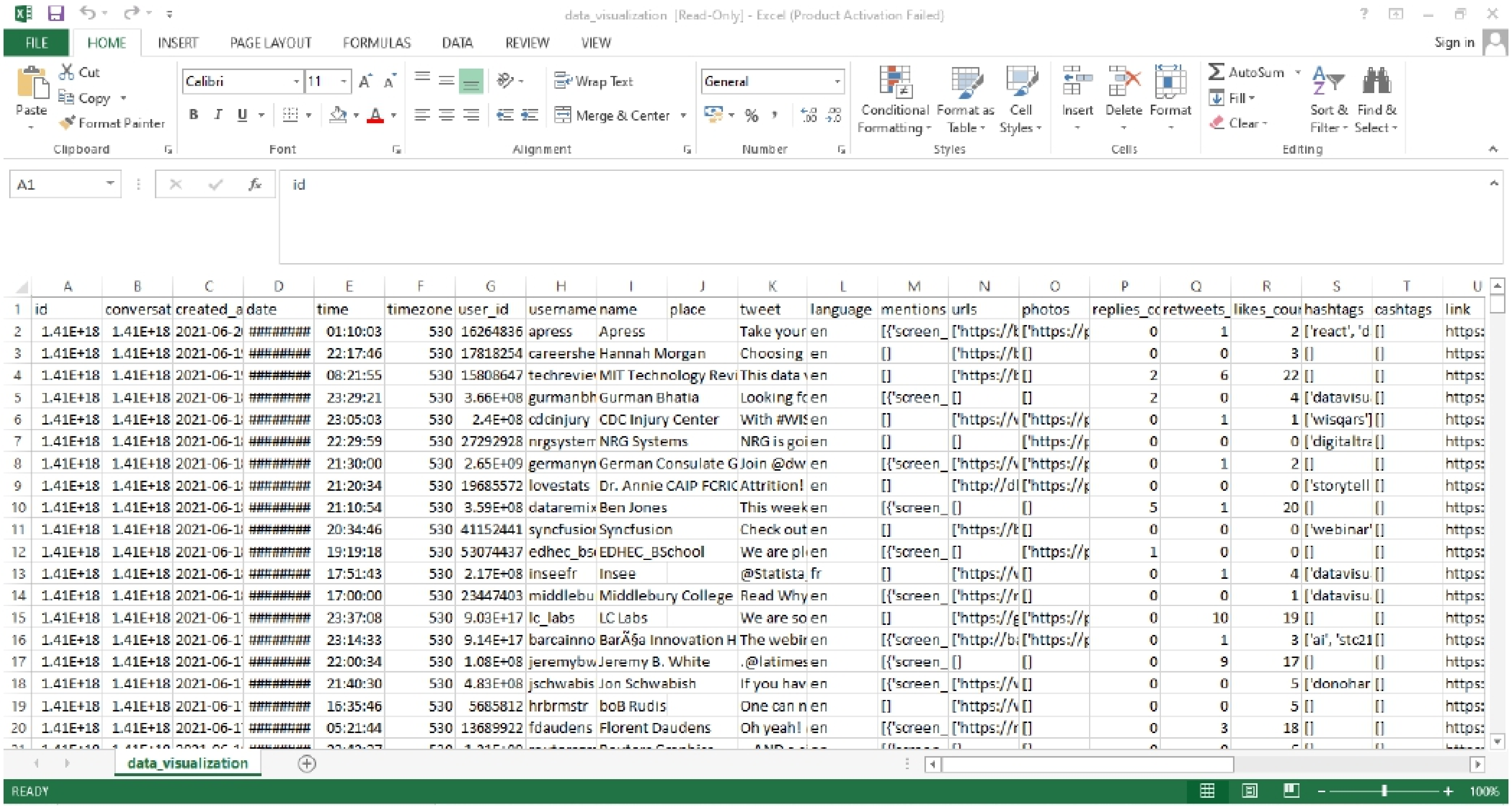
Rightly segregated tweets

Data Set Description:

- By using the given dataset we are going to classify the tweets into positive or negative.
- Perform Sentiment analysis.

Dataset

Following is the Snapshot of Dataset :-



A1 id

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	id	conversat	created_a	date	time	timezone	user_id	username	name	place	tweet	language	mentions	urls	photos	replies_cc	retweets_	likes_cou	hashtags	cashtags	link
2	1.41E+18	1.41E+18	2021-06-2	########	01:10:03	530	16264836	apress	Apress		Take your	en	[{'screen_	['https://t	['https://p	0	1	2	['react', 'd	[]	https:
3	1.41E+18	1.41E+18	2021-06-1	########	22:17:46	530	17818254	careershe	Hannah Morgan		Choosing	en	[]	['https://t	[]	0	0	3	[]	[]	https:
4	1.41E+18	1.41E+18	2021-06-1	########	08:21:55	530	15808647	techrevie	MIT Technology Revi		This data	en	[]	['https://t	[]	2	6	22	[]	[]	https:
5	1.41E+18	1.41E+18	2021-06-1	########	23:29:21	530	3.66E+08	gurmanbh	Gurman Bhatia		Looking fc	en	[{'screen_	[]	[]	2	0	4	['datavisu	[]	https:
6	1.41E+18	1.41E+18	2021-06-1	########	23:05:03	530	2.4E+08	cdcinjury	CDC Injury Center		With #WIS	en	[]	['https://v	['https://p	0	1	1	['wisqars']	[]	https:
7	1.41E+18	1.41E+18	2021-06-1	########	22:29:59	530	27292928	nrgsysterr	NRG Systems		NRG is goi	en	[]	[]	['https://p	0	0	0	['digitaltra	[]	https:
8	1.41E+18	1.41E+18	2021-06-1	########	21:30:00	530	2.65E+09	germanyn	German Consulate G		Join @dw	en	[{'screen_	['https://v	['https://p	0	1	2	[]	[]	https:
9	1.41E+18	1.41E+18	2021-06-1	########	21:20:34	530	19685572	lovestats	Dr. Annie CAIP FCRIC		Attrition!	en	[]	['http://dl	['https://p	0	0	0	['storytell	[]	https:
10	1.41E+18	1.41E+18	2021-06-1	########	21:10:54	530	3.59E+08	dataremix	Ben Jones		This week	en	[{'screen_	[]	[]	5	1	20	[]	[]	https:
11	1.41E+18	1.41E+18	2021-06-1	########	20:34:46	530	41152441	syncfusior	Syncfusion		Check out	en	[]	['https://t	[]	0	0	0	['webinar'	[]	https:
12	1.41E+18	1.41E+18	2021-06-1	########	19:19:18	530	53074437	edhec_bs	EDHEC_BSchool		We are pl	en	[{'screen_	[]	['https://p	1	0	0	[]	[]	https:
13	1.41E+18	1.41E+18	2021-06-1	########	17:51:43	530	2.17E+08	inseefr	Insee		@Statista	fr	[]	['https://v	[]	0	1	4	['datavisu	[]	https:
14	1.41E+18	1.41E+18	2021-06-1	########	17:00:00	530	23447403	middlebu	Middlebury College		Read Why	en	[{'screen_	['https://r	[]	0	0	1	['datavisu	[]	https:
15	1.41E+18	1.41E+18	2021-06-1	########	23:37:08	530	9.03E+17	lc_labs	LC Labs		We are so	en	[]	['https://g	['https://p	0	10	19	[]	[]	https:
16	1.41E+18	1.41E+18	2021-06-1	########	23:14:33	530	9.14E+17	barcainno	BarÃ§a Innovation H		The webir	en	[{'screen_	['http://ba	['https://p	0	1	3	['ai', 'stc21	[]	https:
17	1.41E+18	1.41E+18	2021-06-1	########	22:00:34	530	1.08E+08	jeremybw	Jeremy B. White		.@latimes	en	[{'screen_	[]	[]	0	9	17	[]	[]	https:
18	1.41E+18	1.41E+18	2021-06-1	########	21:40:30	530	4.83E+08	jschwabis	Jon Schwabish		If you hav	en	[{'screen_	['https://v	[]	0	0	5	['donohar	[]	https:
19	1.41E+18	1.41E+18	2021-06-1	########	16:35:46	530	5685812	hrbrmstr	boB Rudis		One can n	en	[]	['https://v	[]	0	0	5	[]	[]	https:
20	1.41E+18	1.41E+18	2021-06-1	########	05:21:44	530	13689922	fdaudens	Florent Daudens		Oh yeah!	en	[{'screen_	['https://r	[]	0	3	18	[]	[]	https:

Theory

1. Sentiment Analysis

Sentiment analysis could be performed in Python using 2 methods — i) Calculating polarity and subjectivity using the library **Textblob.** ii)
Using **sentimentIntensityAnalyzer** from the library **vader**.

The SentimentIntensityAnalyzer function relies on a dictionary that maps lexical features to emotion intensities known as sentiment scores. The sentiment score of a text can be obtained by summing up the intensity of each word in the text.

This function analyzes the text and returns the score in the form of a dictionary with the following components :negative, neutral, positive and compound. Based on the scores assigned to each component, we can define the overall sentiment of the text to be positive, negative or neutral.

2. Visualize the sentiment Counts

Once the sentiments are identified, we can create 3 lists for different sentiments. We can then calculate the overall percentage of each sentiment in the dataset.

Input & Output

Input:

```
import nltk
nltk.download('vader_lexicon')
from nltk.sentiment.vader import SentimentIntensityAnalyzer
for index, row in tweets_df['Text'].iteritems():
    score = SentimentIntensityAnalyzer().polarity_scores(row)
    if score['neg'] > score['pos']:
        tweets_df.loc[index, "Sentiment"] = "negative"
    elif score['pos'] > score['neg']:
        tweets_df.loc[index, "Sentiment"] = "positive"
    else:
        tweets_df.loc[index, "Sentiment"] = "neutral"

    tweets_df.loc[index, 'neg'] = score['neg']
    tweets_df.loc[index, 'neu'] = score['neu']
    tweets_df.loc[index, 'pos'] = score['pos']
    tweets_df.loc[index, 'compound'] = score['compound']

tweets_df.head(10)
#create new data frames for all sentiments
tweet_neg = tweets_df[tweets_df["Sentiment"] == "negative"]
tweet_neu = tweets_df[tweets_df["Sentiment"] == "neutral"]
tweet_pos = tweets_df[tweets_df["Sentiment"] == "positive"]
#function for calculating the percentage of all the sentiments
def calc_percentage(x,y):
    return x/y * 100
pos_per = calc_percentage(len(tweet_pos), len(tweets_df))
neg_per = calc_percentage(len(tweet_neg), len(tweets_df))
neu_per = calc_percentage(len(tweet_neu), len(tweets_df))
print("positive: {} {}%".format(len(tweet_pos),  format(pos_per, '.1f')))
print("negative: {} {}%".format(len(tweet_neg), format(neg_per, '.1f')))
print("neutral: {} {}%".format(len(tweet_neu), format(neu_per,
'.1f')))format(calc_percentage(len(tweet_neu), len(tweets_df)), '.1f')))
```

Output:

```
positive: 1788 35.8%
negative: 1795 35.9%
neutral: 1417 28.3%
```

Conclusion

Thus we have performed the twitter sentiment analysis. There are many more analysis techniques that you can apply on the twitter data like topic modelling. Other techniques are geospatial analysis, text similarity and knowledge graphs. The possibilities are endless.