COMSATS UNIVERSITY

WAH CAMPUS

Submitted By

Name:

Shayan Babar	M.Arsal	Ali Raza	Saad Rashid

Registration No:

Fa19bcs051	Fa19-BCS-056	Fa19-BCS-047	Fa19-BCS-194

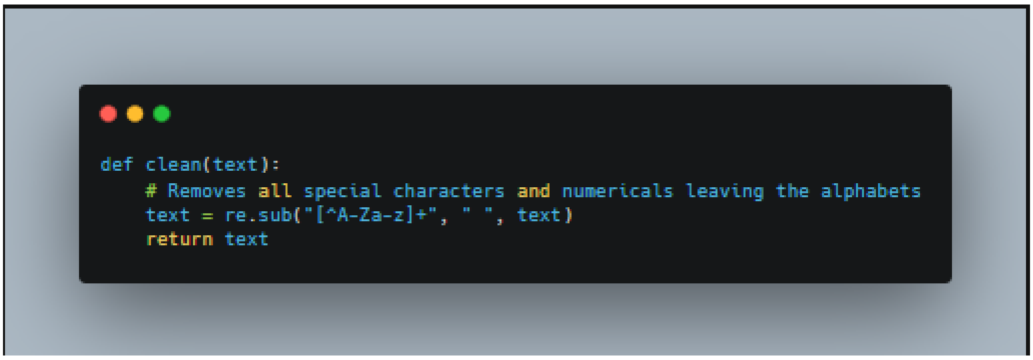
Class/Section: *BSCS/6d*

Submitted To

Teacher Name: *Ammara Zamir*

Date of Submission: *14/06/2022*

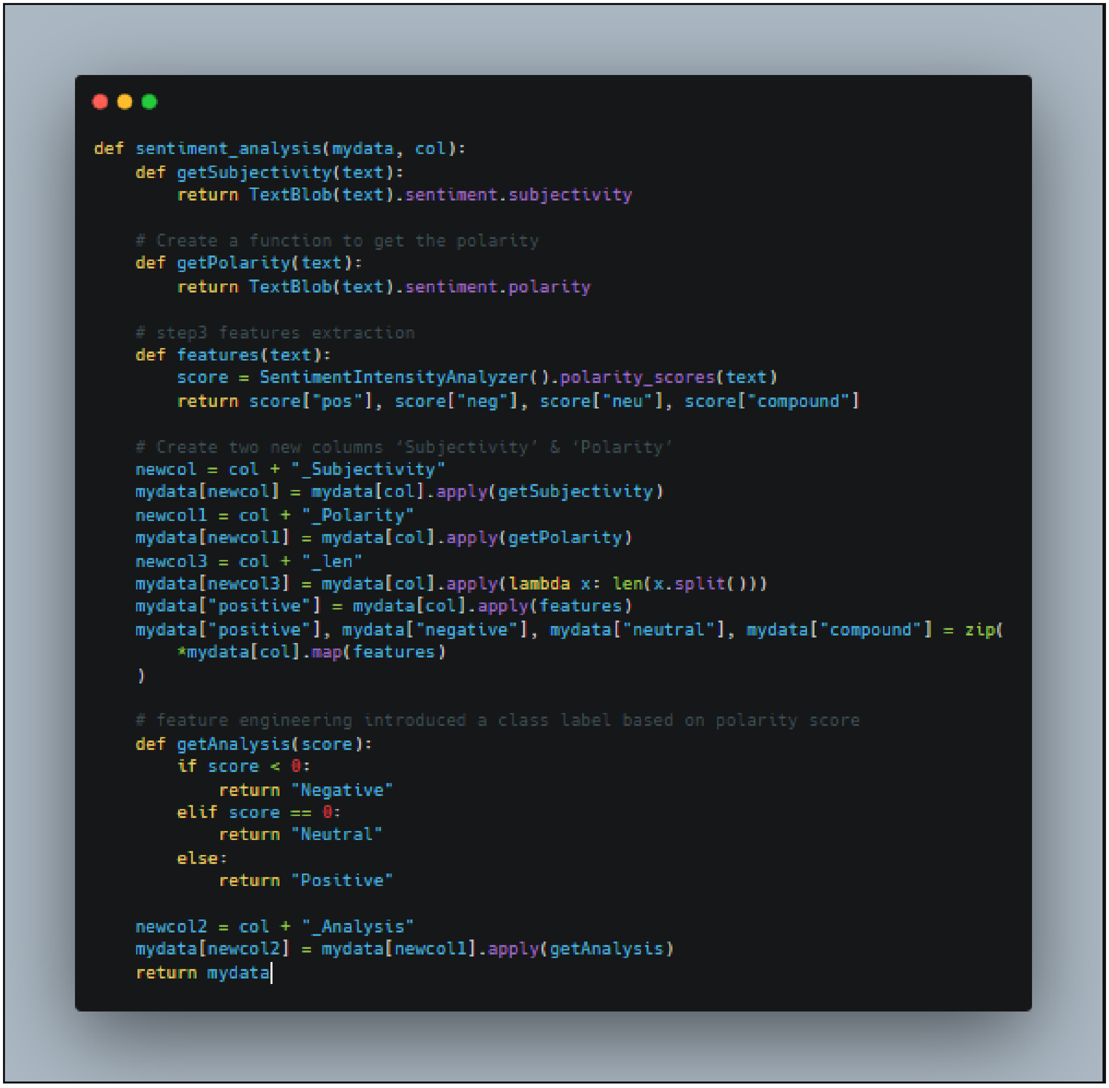
1. Cleaning data



```
def clean(text):
    # Removes all special characters and numericals leaving the alphabets
    text = re.sub("[^A-Za-z]+", " ", text)
    return text
```

Because out data set was small we were able to analyze that the only irregularities it had were specials characters present in it hence using regular expressions we removed than using the above code which is called later in the actual code its self.

2. Text blob sentiment analysis and feature extraction



```
def sentiment_analysis(mydata, col):
    def getSubjectivity(text):
        return TextBlob(text).sentiment.subjectivity

    # Create a function to get the polarity
    def getPolarity(text):
        return TextBlob(text).sentiment.polarity

    # step3 features extraction
    def features(text):
        score = SentimentIntensityAnalyzer().polarity_scores(text)
        return score["pos"], score["neg"], score["neu"], score["compound"]

    # Create two new columns 'Subjectivity' & 'Polarity'
    newcol = col + "_Subjectivity"
    mydata[newcol] = mydata[col].apply(getSubjectivity)
    newcol1 = col + "_Polarity"
    mydata[newcol1] = mydata[col].apply(getPolarity)
    newcol3 = col + "_len"
    mydata[newcol3] = mydata[col].apply(lambda x: len(x.split()))
    mydata["positive"] = mydata[col].apply(features)
    mydata["positive"], mydata["negative"], mydata["neutral"], mydata["compound"] = zip(
        *mydata[col].map(features)
    )

    # feature engineering introduced a class label based on polarity score
    def getAnalysis(score):
        if score < 0:
            return "Negative"
        elif score == 0:
            return "Neutral"
        else:
            return "Positive"

    newcol2 = col + "_Analysis"
    mydata[newcol2] = mydata[newcol1].apply(getAnalysis)
    return mydata
```

- **getSubjectivity** – finds the subjectivity of the text passed using text blob function extBlob(text).sentiment.subjectivity

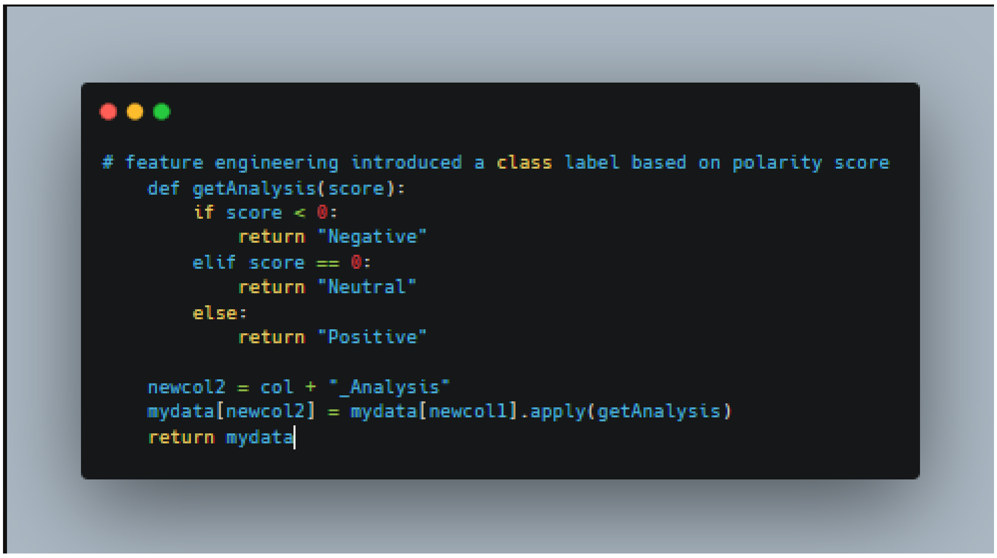
- **getPolarity** --finds the Polarity of the text passed using text blob function extBlob(text).sentiment. Polarity

- **features(text)** – ectracts features(positive,negative,neutral,compound) using ntlk function SentimentIntensityAnalyzer().polarity_scores(text)

- **len—**feature is founded using string.split()
- **data is maped using dataframe[column].apply(function_name)**
- **in feature extraction method .map() function is used to map 4 columns in data frame.**

3. Label class

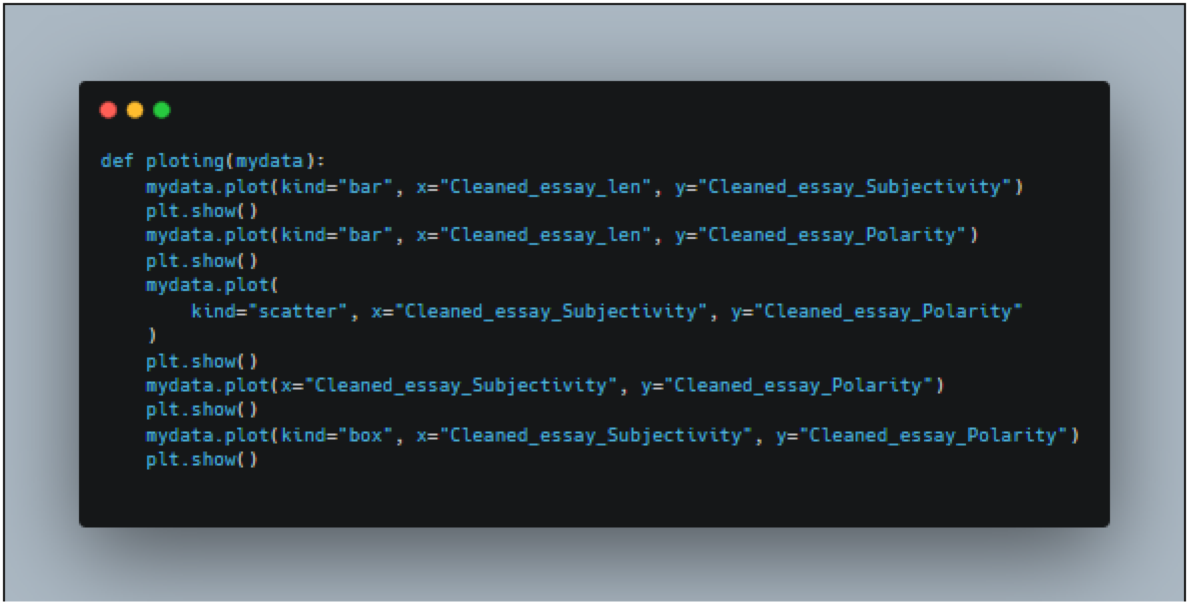
Label class is introduced using following code (cols named is” Cleaned_essay_Analysis”)



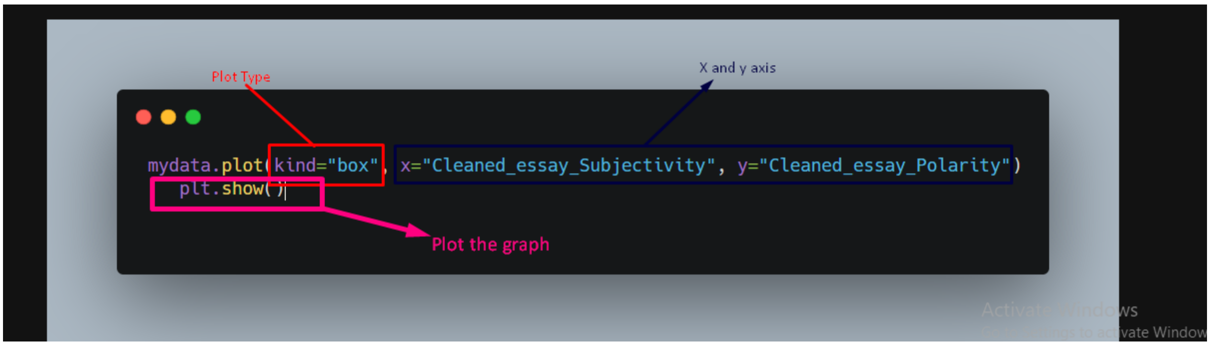
```
# feature engineering introduced a class label based on polarity score
    def getAnalysis(score):
        if score < 0:
            return "Negative"
        elif score == 0:
            return "Neutral"
        else:
            return "Positive"

    newcol2 = col + "_Analysis"
    mydata[newcol2] = mydata[newcol1].apply(getAnalysis)
    return mydata
```

Polarity is the output that lies between [-1,1], where -1 refers to negative sentiment and +1 refers to positive sentiment

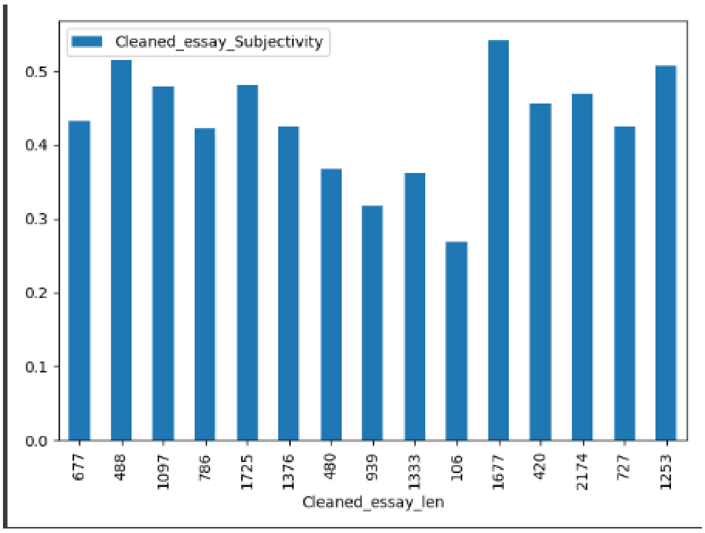
4. Mathplotlib



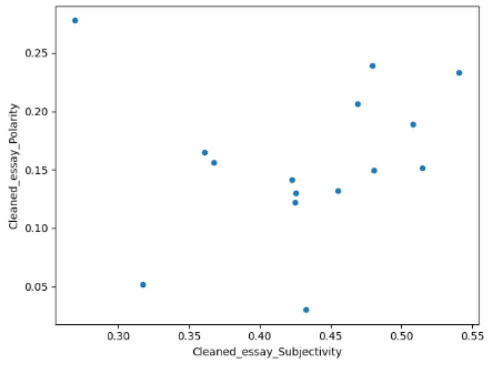
```
def ploting(mydata):
    mydata.plot(kind="bar", x="Cleaned_essay_len", y="Cleaned_essay_Subjectivity")
    plt.show()
    mydata.plot(kind="bar", x="Cleaned_essay_len", y="Cleaned_essay_Polarity")
    plt.show()
    mydata.plot(
        kind="scatter", x="Cleaned_essay_Subjectivity", y="Cleaned_essay_Polarity"
    )
    plt.show()
    mydata.plot(x="Cleaned_essay_Subjectivity", y="Cleaned_essay_Polarity")
    plt.show()
    mydata.plot(kind="box", x="Cleaned_essay_Subjectivity", y="Cleaned_essay_Polarity")
    plt.show()
```



```
mydata.plot(kind="box", x="Cleaned_essay_Subjectivity", y="Cleaned_essay_Polarity")
    plt.show()
```

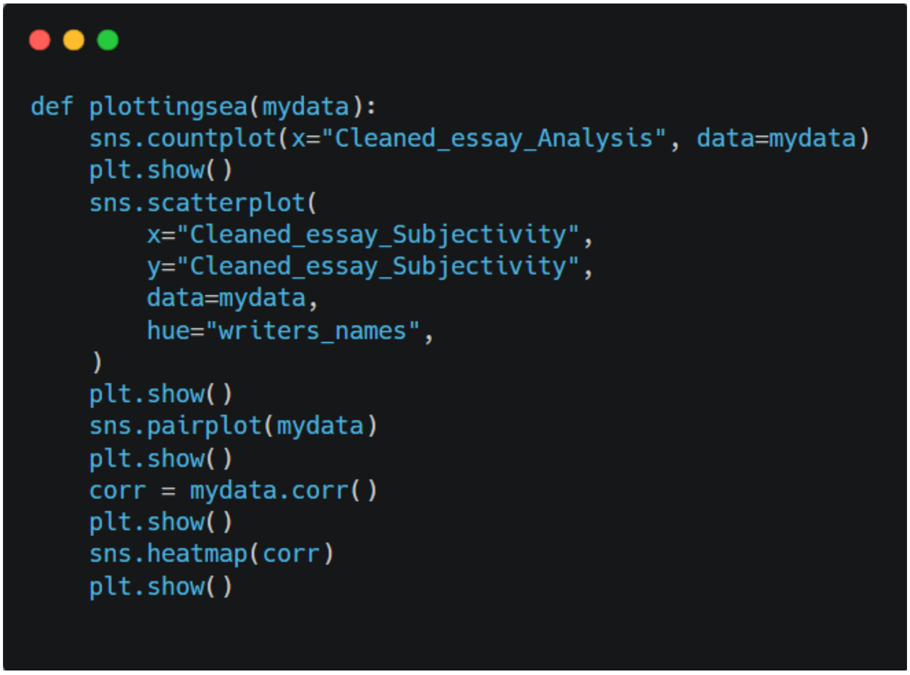


Simple bar chart of subjectivity and length of the essay



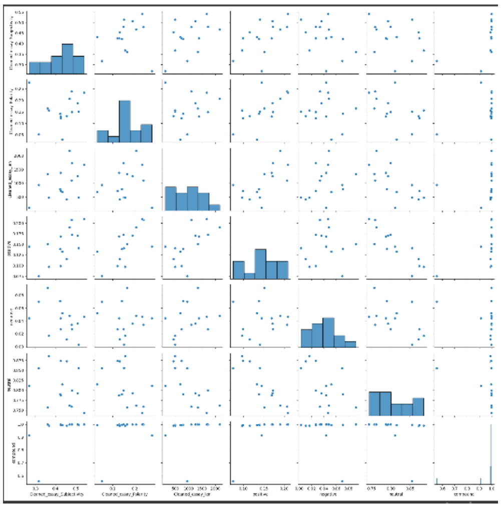
Scatter plot subjectivity and polarity

5. Seaborn

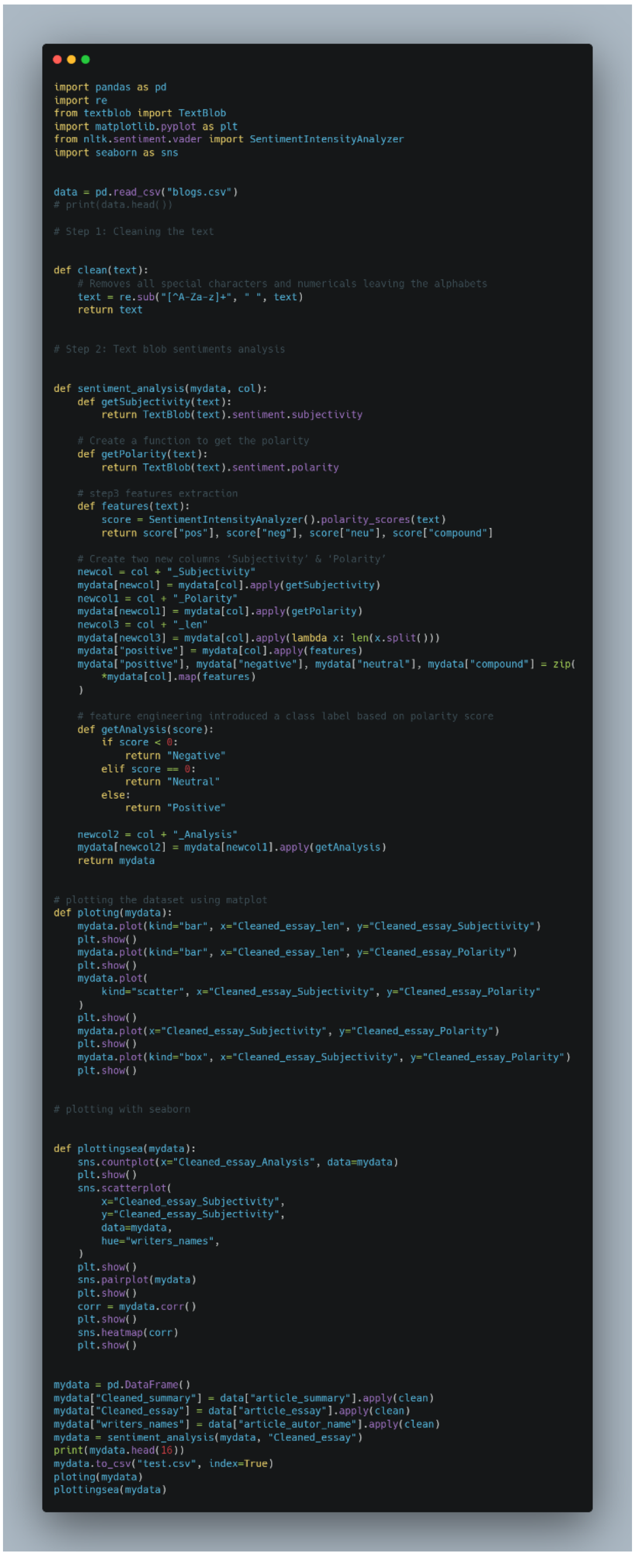


```
def plottingsea(mydata):
    sns.countplot(x="Cleaned_essay_Analysis", data=mydata)
    plt.show()
    sns.scatterplot(
        x="Cleaned_essay_Subjectivity",
        y="Cleaned_essay_Subjectivity",
        data=mydata,
        hue="writers_names",
    )
    plt.show()
    sns.pairplot(mydata)
    plt.show()
    corr = mydata.corr()
    plt.show()
    sns.heatmap(corr)
    plt.show()
```

Same as mathplotlip except the plot time is defined like this sns.plottype(,,,,,)

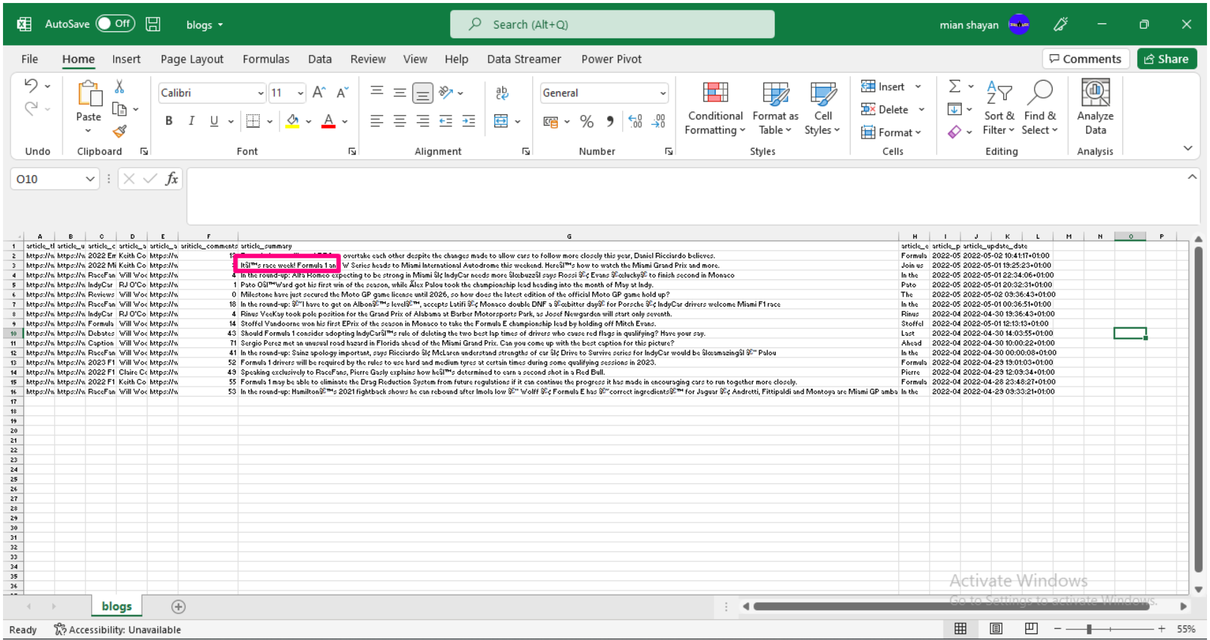


6. Main code



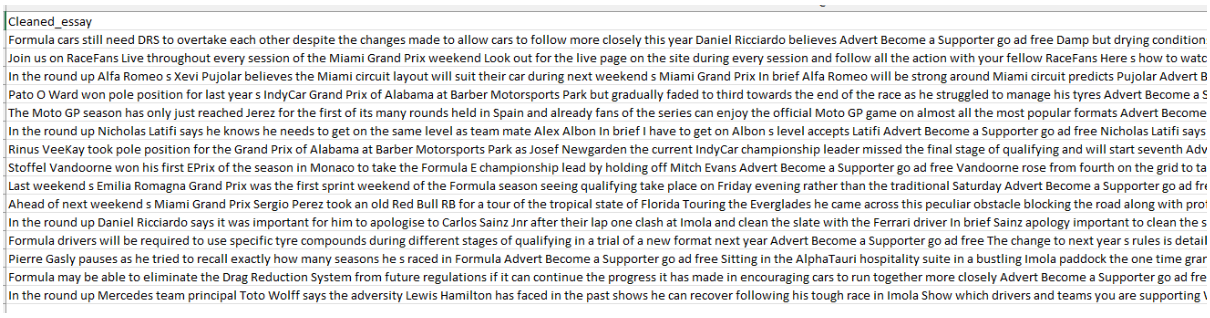
```python
import pandas as pd
import re
from textblob import TextBlob
import matplotlib.pyplot as plt
from nltk.sentiment.vader import SentimentIntensityAnalyzer
import seaborn as sns


data = pd.read_csv("blogs.csv")
# print(data.head())

# Step 1: Cleaning the text


def clean(text):
    # Removes all special characters and numericals leaving the alphabets
    text = re.sub("[^A-Za-z]+", " ", text)
    return text


# Step 2: Text blob sentiments analysis


def sentiment_analysis(mydata, col):
    def getSubjectivity(text):
        return TextBlob(text).sentiment.subjectivity

    # Create a function to get the polarity
    def getPolarity(text):
        return TextBlob(text).sentiment.polarity

    # step3 features extraction
    def features(text):
        score = SentimentIntensityAnalyzer().polarity_scores(text)
        return score["pos"], score["neg"], score["neu"], score["compound"]

    # Create two new columns ‘Subjectivity’ & ‘Polarity’
    newcol = col + "_Subjectivity"
    mydata[newcol] = mydata[col].apply(getSubjectivity)
    newcol1 = col + "_Polarity"
    mydata[newcol1] = mydata[col].apply(getPolarity)
    newcol3 = col + "_len"
    mydata[newcol3] = mydata[col].apply(lambda x: len(x.split()))
    mydata["positive"] = mydata[col].apply(features)
    mydata["positive"], mydata["negative"], mydata["neutral"], mydata["compound"] = zip(
        *mydata[col].map(features)
    )

    # feature engineering introduced a class label based on polarity score
    def getAnalysis(score):
        if score < 0:
            return "Negative"
        elif score == 0:
            return "Neutral"
        else:
            return "Positive"

    newcol2 = col + "_Analysis"
    mydata[newcol2] = mydata[newcol1].apply(getAnalysis)
    return mydata


# plotting the dataset using matplot
def ploting(mydata):
    mydata.plot(kind="bar", x="Cleaned_essay_len", y="Cleaned_essay_Subjectivity")
    plt.show()
    mydata.plot(kind="bar", x="Cleaned_essay_len", y="Cleaned_essay_Polarity")
    plt.show()
    mydata.plot(
        kind="scatter", x="Cleaned_essay_Subjectivity", y="Cleaned_essay_Polarity"
    )
    plt.show()
    mydata.plot(x="Cleaned_essay_Subjectivity", y="Cleaned_essay_Polarity")
    plt.show()
    mydata.plot(kind="box", x="Cleaned_essay_Subjectivity", y="Cleaned_essay_Polarity")
    plt.show()


# plotting with seaborn


def plottingsea(mydata):
    sns.countplot(x="Cleaned_essay_Analysis", data=mydata)
    plt.show()
    sns.scatterplot(
        x="Cleaned_essay_Subjectivity",
        y="Cleaned_essay_Subjectivity",
        data=mydata,
        hue="writers_names",
    )
    plt.show()
    sns.pairplot(mydata)
    plt.show()
    corr = mydata.corr()
    plt.show()
    sns.heatmap(corr)
    plt.show()


mydata = pd.DataFrame()
mydata["Cleaned_summary"] = data["article_summary"].apply(clean)
mydata["Cleaned_essay"] = data["article_essay"].apply(clean)
mydata["writers_names"] = data["article_autor_name"].apply(clean)
mydata = sentiment_analysis(mydata, "Cleaned_essay")
print(mydata.head(16))
mydata.to_csv("test.csv", index=True)
ploting(mydata)
plottingsea(mydata)
```

7. Csv results


AutoSave Off
blogs
Search (Alt+Q)
mian shayan
File Home Insert Page Layout Formulas Data Review View Help Data Streamer Power Pivot
Comments
Share
O10
Ready
Accessibility: Unavailable
55%

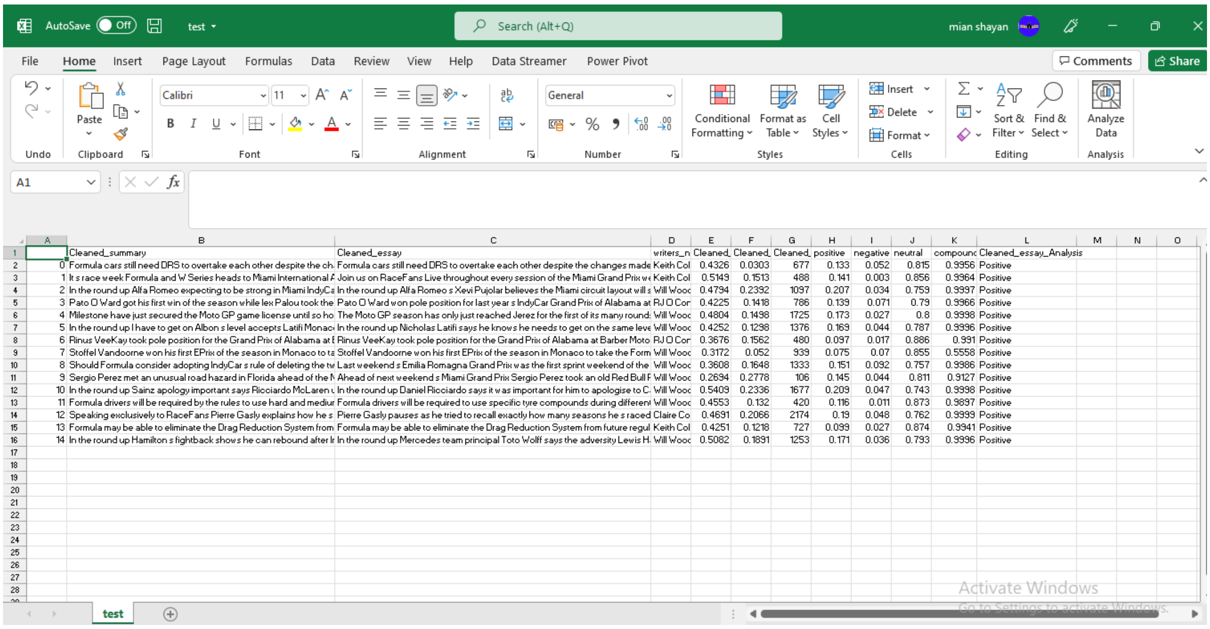
	A	B	C	D	E	F	G	H	I	J
1	article_tl	article_u	article_c	article_a	article_a	ariticle_comments	article_summary	article_e	article_p	article_update_date
2	https://w	https://w	2022 Em	Keith Co	https://w	12	Formula 1 cars still need DRS to overtake each other despite the changes made to allow cars to follow more closely this year, Daniel Ricciardo believes.	Formula	2022-05	2022-05-02 10:41:17+01:00
3	https://w	https://w	2022 Mi	Keith Co	https://w	[illegible]	Itâ€™s race week! Formula 1 an W Series heads to Miami International Autodrome this weekend. Hereâ€™s how to watch the Miami Grand Prix and more.	Join us	2022-05	2022-05-01 19:25:23+01:00
4	https://w	https://w	RaceFan	Will Woc	https://w	4	In the round-up: Alfa Romeo expecting to be strong in Miami â€¢ IndyCar needs more â€œbuzzâ€ says Rossi â€¢ Evans â€œluckyâ€ to finish second in Monaco	In the	2022-05	2022-05-01 22:34:06+01:00
5	https://w	https://w	IndyCar	RJ O'Co	https://w	1	Pato Oâ€™Ward got his first win of the season, while Ãlex Palou took the championship lead heading into the month of May at Indy.	Pato	2022-05	2022-05-01 20:32:31+01:00
6	https://w	https://w	Reviews	Will Woc	https://w	0	Milestone have just secured the Moto GP game license until 2026, so how does the latest edition of the official Moto GP game hold up?	The	2022-05	2022-05-02 09:36:43+01:00
7	https://w	https://w	RaceFan	Will Woc	https://w	18	In the round-up: â€˜I have to get on Albonâ€™s levelâ€™, accepts Latifi â€¢ Monaco double DNF a â€œbitter dayâ€ for Porsche â€¢ IndyCar drivers welcome Miami F1 race	In the	2022-05	2022-05-01 00:36:51+01:00
8	https://w	https://w	IndyCar	RJ O'Co	https://w	4	Rinus VeeKay took pole position for the Grand Prix of Alabama at Barber Motorsports Park, as Josef Newgarden will start only seventh.	Rinus	2022-04	2022-04-30 19:36:43+01:00
9	https://w	https://w	Formula	Will Woc	https://w	14	Stoffel Vandoorne won his first EPrix of the season in Monaco to take the Formula E championship lead by holding off Mitch Evans.	Stoffel	2022-04	2022-05-01 12:13:13+01:00
10	https://w	https://w	Debates	Will Woc	https://w	43	Should Formula 1 consider adopting IndyCarâ€™s rule of deleting the two best lap times of drivers who cause red flags in qualifying? Have your say.	Last	2022-04	2022-04-30 14:03:55+01:00
11	https://w	https://w	Caption	Will Woc	https://w	71	Sergio Perez met an unusual road hazard in Florida ahead of the Miami Grand Prix. Can you come up with the best caption for this picture?	Ahead	2022-04	2022-04-30 10:00:22+01:00
12	https://w	https://w	RaceFan	Will Woc	https://w	41	In the round-up: Sainz apology important, says Ricciardo â€¢ McLaren understand strengths of car â€¢ Drive to Survive series for IndyCar would be â€œamazingâ€ â€“ Palou	In the	2022-04	2022-04-30 00:00:08+01:00
13	https://w	https://w	2023 F1	Will Woc	https://w	52	Formula 1 drivers will be required by the rules to use hard and medium tyres at certain times during some qualifying sessions in 2023.	Formula	2022-04	2022-04-29 19:01:03+01:00
14	https://w	https://w	2022 F1	Claire Cc	https://w	49	Speaking exclusively to RaceFans, Pierre Gasly explains how heâ€™s determined to earn a second shot in a Red Bull.	Pierre	2022-04	2022-04-29 12:09:34+01:00
15	https://w	https://w	2022 F1	Keith Co	https://w	55	Formula 1 may be able to eliminate the Drag Reduction System from future regulations if it can continue the progress it has made in encouraging cars to run together more closely.	Formula	2022-04	2022-04-28 23:48:27+01:00
16	https://w	https://w	RaceFan	Will Woc	https://w	53	In the round-up: Hamiltonâ€™s 2021 fightback shows he can rebound after Imola low â€“ Wolff â€¢ Formula E has â€˜correct ingredientsâ€™ for Jaguar â€¢ Andretti, Fittipaldi and Montoya are Miami GP amba	In the	2022-04	2022-04-29 09:33:21+01:00

Uncleaned scaped data



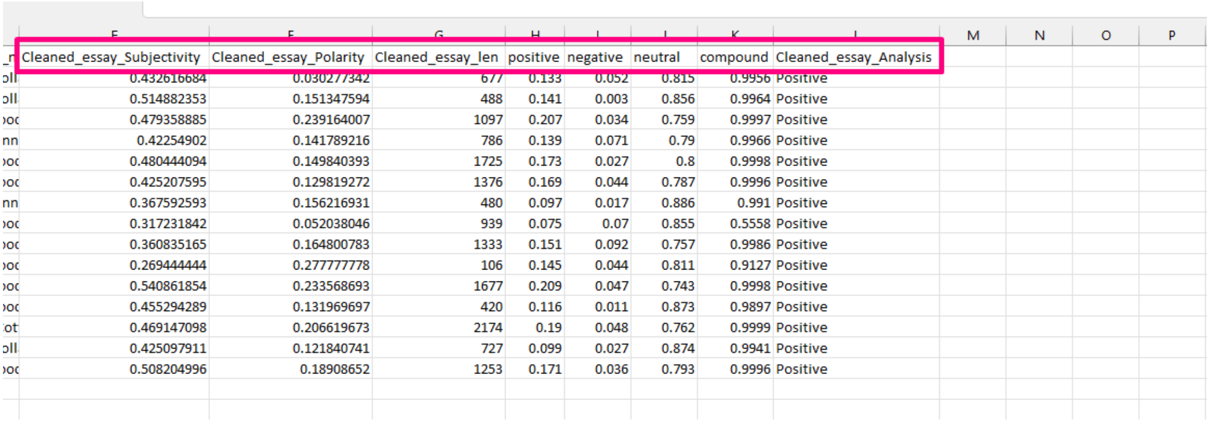
Cleaned_essay
Formula cars still need DRS to overtake each other despite the changes made to allow cars to follow more closely this year Daniel Ricciardo believes Advert Become a Supporter go ad free Damp but drying condition
Join us on RaceFans Live throughout every session of the Miami Grand Prix weekend Look out for the live page on the site during every session and follow all the action with your fellow RaceFans Here s how to watc
In the round up Alfa Romeo s Xevi Pujolar believes the Miami circuit layout will suit their car during next weekend s Miami Grand Prix In brief Alfa Romeo will be strong around Miami circuit predicts Pujolar Advert B
Pato O Ward won pole position for last year s IndyCar Grand Prix of Alabama at Barber Motorsports Park but gradually faded to third towards the end of the race as he struggled to manage his tyres Advert Become a S
The Moto GP season has only just reached Jerez for the first of its many rounds held in Spain and already fans of the series can enjoy the official Moto GP game on almost all the most popular formats Advert Become
In the round up Nicholas Latifi says he knows he needs to get on the same level as team mate Alex Albon In brief I have to get on Albon s level accepts Latifi Advert Become a Supporter go ad free Nicholas Latifi says
Rinus VeeKay took pole position for the Grand Prix of Alabama at Barber Motorsports Park as Josef Newgarden the current IndyCar championship leader missed the final stage of qualifying and will start seventh Adv
Stoffel Vandoorne won his first EPrix of the season in Monaco to take the Formula E championship lead by holding off Mitch Evans Advert Become a Supporter go ad free Vandoorne rose from fourth on the grid to ta
Last weekend s Emilia Romagna Grand Prix was the first sprint weekend of the Formula season seeing qualifying take place on Friday evening rather than the traditional Saturday Advert Become a Supporter go ad fr
Ahead of next weekend s Miami Grand Prix Sergio Perez took an old Red Bull RB for a tour of the tropical state of Florida Touring the Everglades he came across this peculiar obstacle blocking the road along with pro
In the round up Daniel Ricciardo says it was important for him to apologise to Carlos Sainz Jnr after their lap one clash at Imola and clean the slate with the Ferrari driver In brief Sainz apology important to clean the s
Formula drivers will be required to use specific tyre compounds during different stages of qualifying in a trial of a new format next year Advert Become a Supporter go ad free The change to next year s rules is detail
Pierre Gasly pauses as he tried to recall exactly how many seasons he s raced in Formula Advert Become a Supporter go ad free Sitting in the AlphaTauri hospitality suite in a bustling Imola paddock the one time grar
Formula may be able to eliminate the Drag Reduction System from future regulations if it can continue the progress it has made in encouraging cars to run together more closely Advert Become a Supporter go ad fre
In the round up Mercedes team principal Toto Wolff says the adversity Lewis Hamilton has faced in the past shows he can recover following his tough race in Imola Show which drivers and teams you are supporting \

Cleaned data


AutoSave Off
test
Search (Alt+Q)
mian shayan
File Home Insert Page Layout Formulas Data Review View Help Data Streamer Power Pivot
Comments
Share
A1

	A	B	C	D	E	F	G	H	I	J	K	L
1		Cleaned_summary	Cleaned_essay	writers_n	Cleaned_	Cleaned_	Cleaned_	positive	negative	neutral	compounc	Cleaned_essay_Analysis
2	0	Formula cars still need DRS to overtake each other despite the ch	Formula cars still need DRS to overtake each other despite the changes made	Keith Col	0.4326	0.0303	677	0.133	0.052	0.815	0.9956	Positive
3	1	It s race week Formula and W Series heads to Miami International A	Join us on RaceFans Live throughout every session of the Miami Grand Prix we	Keith Col	0.5149	0.1513	488	0.141	0.003	0.856	0.9964	Positive
4	2	In the round up Alfa Romeo expecting to be strong in Miami IndyCa	In the round up Alfa Romeo s Xevi Pujolar believes the Miami circuit layout will s	Will Wooc	0.4794	0.2392	1097	0.207	0.034	0.759	0.9997	Positive
5	3	Pato O Ward got his first win of the season while lex Palou took the	Pato O Ward won pole position for last year s IndyCar Grand Prix of Alabama at	RJ O Cor	0.4225	0.1418	786	0.139	0.071	0.79	0.9966	Positive
6	4	Milestone have just secured the Moto GP game license until so ho	The Moto GP season has only just reached Jerez for the first of its many round:	Will Wooc	0.4804	0.1498	1725	0.173	0.027	0.8	0.9998	Positive
7	5	In the round up I have to get on Albon s level accepts Latifi Monac	In the round up Nicholas Latifi says he knows he needs to get on the same leve	Will Wooc	0.4252	0.1298	1376	0.169	0.044	0.787	0.9996	Positive
8	6	Rinus VeeKay took pole position for the Grand Prix of Alabama at B	Rinus VeeKay took pole position for the Grand Prix of Alabama at Barber Moto	RJ O Cor	0.3676	0.1562	480	0.097	0.017	0.886	0.991	Positive
9	7	Stoffel Vandoorne won his first EPrix of the season in Monaco to ta	Stoffel Vandoorne won his first EPrix of the season in Monaco to take the Form	Will Wooc	0.3172	0.052	939	0.075	0.07	0.855	0.5558	Positive
10	8	Should Formula consider adopting IndyCar s rule of deleting the tw	Last weekend s Emilia Romagna Grand Prix was the first sprint weekend of the	Will Wooc	0.3608	0.1648	1333	0.151	0.092	0.757	0.9986	Positive
11	9	Sergio Perez met an unusual road hazard in Florida ahead of the M	Ahead of next weekend s Miami Grand Prix Sergio Perez took an old Red Bull F	Will Wooc	0.2694	0.2778	106	0.145	0.044	0.811	0.9127	Positive
12	10	In the round up Sainz apology important says Ricciardo McLaren u	In the round up Daniel Ricciardo says it was important for him to apologise to C	Will Wooc	0.5409	0.2336	1677	0.209	0.047	0.743	0.9998	Positive
13	11	Formula drivers will be required by the rules to use hard and mediur	Formula drivers will be required to use specific tyre compounds during different	Will Wooc	0.4553	0.132	420	0.116	0.011	0.873	0.9897	Positive
14	12	Speaking exclusively to RaceFans Pierre Gasly explains how he s	Pierre Gasly pauses as he tried to recall exactly how many seasons he s raced	Claire Co	0.4691	0.2066	2174	0.19	0.048	0.762	0.9999	Positive
15	13	Formula may be able to eliminate the Drag Reduction System from	Formula may be able to eliminate the Drag Reduction System from future regul	Keith Col	0.4251	0.1218	727	0.099	0.027	0.874	0.9941	Positive
16	14	In the round up Hamilton s fightback shows he can rebound after Ir	In the round up Mercedes team principal Toto Wolff says the adversity Lewis H	Will Wooc	0.5082	0.1891	1253	0.171	0.036	0.793	0.9996	Positive

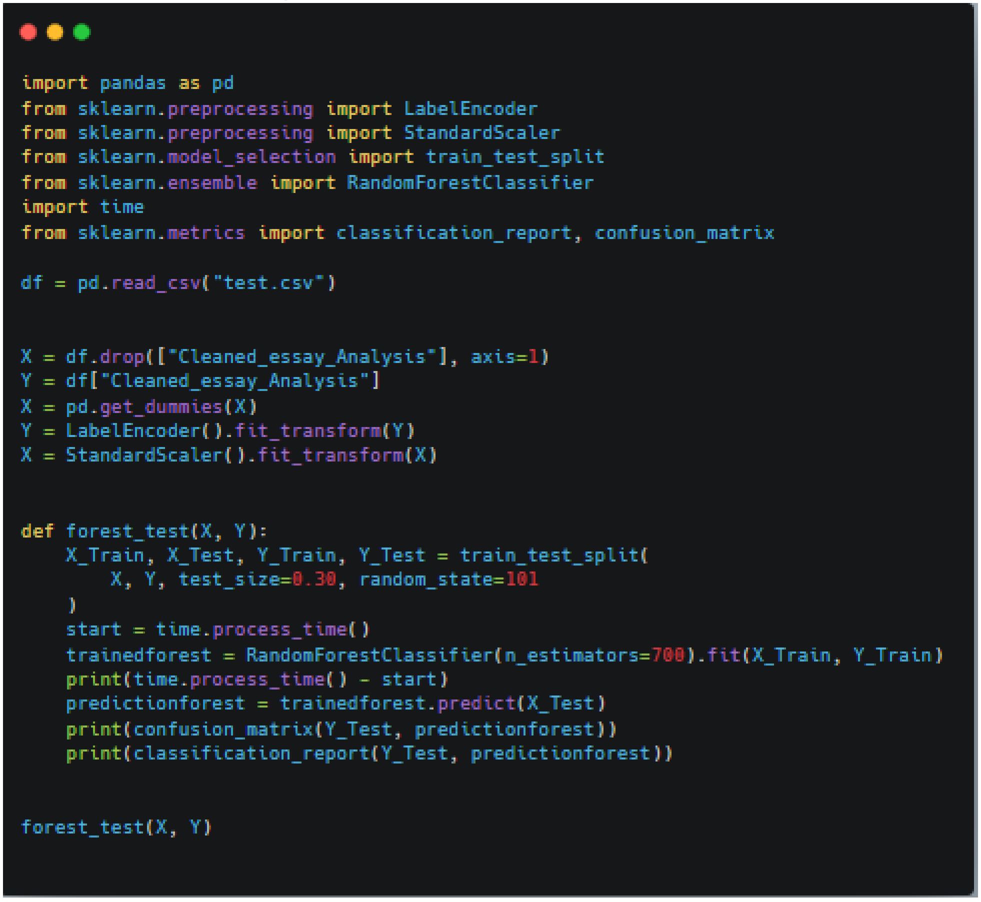
After running the code



_r	Cleaned_essay_Subjectivity	Cleaned_essay_Polarity	Cleaned_essay_len	positive	negative	neutral	compound	Cleaned_essay_Analysis
oll	0.432616684	0.030277342	677	0.133	0.052	0.815	0.9956	Positive
oll	0.514882353	0.151347594	488	0.141	0.003	0.856	0.9964	Positive
ooc	0.479358885	0.239164007	1097	0.207	0.034	0.759	0.9997	Positive
nn	0.42254902	0.141789216	786	0.139	0.071	0.79	0.9966	Positive
ooc	0.480444094	0.149840393	1725	0.173	0.027	0.8	0.9998	Positive
ooc	0.425207595	0.129819272	1376	0.169	0.044	0.787	0.9996	Positive
nn	0.367592593	0.156216931	480	0.097	0.017	0.886	0.991	Positive
ooc	0.317231842	0.052038046	939	0.075	0.07	0.855	0.5558	Positive
ooc	0.360835165	0.164800783	1333	0.151	0.092	0.757	0.9986	Positive
ooc	0.269444444	0.277777778	106	0.145	0.044	0.811	0.9127	Positive
ooc	0.540861854	0.233568693	1677	0.209	0.047	0.743	0.9998	Positive
ooc	0.455294289	0.131969697	420	0.116	0.011	0.873	0.9897	Positive
ot	0.469147098	0.206619673	2174	0.19	0.048	0.762	0.9999	Positive
oll	0.425097911	0.121840741	727	0.099	0.027	0.874	0.9941	Positive
ooc	0.508204996	0.18908652	1253	0.171	0.036	0.793	0.9996	Positive

Values

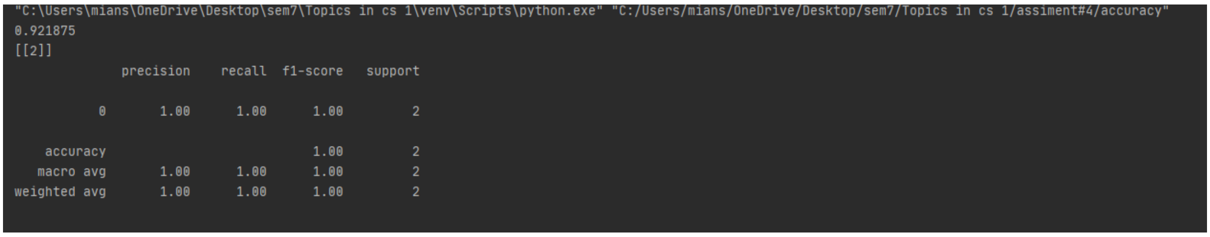
8. Classification using random and accuracy with confusion matrix



```
import pandas as pd
from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
import time
from sklearn.metrics import classification_report, confusion_matrix

df = pd.read_csv("test.csv")


X = df.drop(["Cleaned_essay_Analysis"], axis=1)
Y = df["Cleaned_essay_Analysis"]
X = pd.get_dummies(X)
Y = LabelEncoder().fit_transform(Y)
X = StandardScaler().fit_transform(X)


def forest_test(X, Y):
    X_Train, X_Test, Y_Train, Y_Test = train_test_split(
        X, Y, test_size=0.30, random_state=101
    )
    start = time.process_time()
    trainedforest = RandomForestClassifier(n_estimators=700).fit(X_Train, Y_Train)
    print(time.process_time() - start)
    predictionforest = trainedforest.predict(X_Test)
    print(confusion_matrix(Y_Test, predictionforest))
    print(classification_report(Y_Test, predictionforest))


forest_test(X, Y)
```

Before feeding this data into our Machine Learning models I decided to divide our data into features (X) and labels (Y)

a function ($forest\_test$) to divide the input data into train and test sets and then train and test a Random Forest Classifier



```
"C:\Users\mians\OneDrive\Desktop\sem7\Topics in cs 1\venv\Scripts\python.exe" "C:/Users/mians/OneDrive/Desktop/sem7/Topics in cs 1/assiment#4/accuracy"
0.921875
[[2]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00         2

    accuracy                           1.00         2
   macro avg       1.00      1.00      1.00         2
weighted avg       1.00      1.00      1.00         2
```

As shown below, training a Random Forest classifier using all the features, led to 100% Accuracy in about 0.9s of training time