THE TWITTER BEHAVIOURAL ANALYTICS

Dr C K Gomathy[1], Ms. Bairagi Lalitha[2], Ms. Chilukuri Lakshmi Sowjanya[3], Pavan Sampath[4]

***

ABSTARCT: Social media is basically used to convey opinions. So mostly the communications are done through Facebook, Instagram, whatsapp and twitter. So among all these the best communicator of public opinions could be Twitter. Twitter is the fast and easy way to understand the users perspectives. The major idea of this project is unlike existing systems that analyse positive, negative and neutral tweets we wanted to add some other features and effective solution for all the problems in twitter behavioural analytics. Sentiment analysis comes under behaviour analytics. The aim of our project is we have to develop a system that is efficient and accurate in classifying the tweets. Analysing tweets helped in wide variety of applications like stock market prediction, climatic conditions, predicting prime minister (politics) etc. So for this we need to build a perfect classifier with as many techniques as possible.

INTRODUCTION:

Social media has come an essential part of life, it's a comparatively cheap and extensively accessible medium that enables anyone to broadcast and pierce information/ news/ knowledge, and to make new relations. It's a tool which is used to partake different opinions that can belong to different subjects; similar as philanthropic causes, environmental irregularities, profitable issues, or political controversies. Some of the social media allows the druggies to interact with other who are far down from them. Due to simple and easy sequestration programs and easy availability of the social media the peoples are choices to use further and further. Twitter is one of the social media people used to partake their opinion through their post. Twitter is a social network which allows its druggies to post and partake short dispatches (up to 140 characters) called tweets. Over the once decades, Twitter has spread worldwide and has come one of the major social networks. Behavioral analysis is the type of the analysis where the data from the set of data stored in the data depository are taken and relating the geste or the opinion of the stoner grounded on that data.

PROBLEM STATEMENT:

Now-a-days we have many systems on Twitter data sentiment analysis but none are the perfect classifiers according to our research. And there is no efficient model that could predict the exact output with a good accuracy.

Data extraction is also a difficult task. So as the people are more dependent on the social, edia accounts rather seeing news or newspapers the information need to be accurate.so we should detect the sentiment of the people in a perfect way.

Twitter has a big volume data so analysing this big data is a huge task. So we need to extract each sentences from each tweet. And then the sentiments need to be translated by removing some symbols like #, @ etc. amd then we need to summarize the content of each tweet and organize tweet by tweet according to the class they belong to and make it in an understandable format.

ENGLISH SENTENCES IN AN IMPROPER WAY: Some people does not know proper

English. They text the sentences in whatever way they need. and some cannot express their opinions properly even that sentences need to be analysed which is a difficult task and also there is no system that has a good accuracy. So we need to use as many algorithms that are possible and bring out a efficient solution in classifying tweets word by word, sentence by sentence.

LITERATURE REVIEW:

Now-a-days twitter is used in many applications like health, crime, disasters etc. one among them is Geo-tweets which is proposed by T. Hu, B. She, L. Duan, H. Yue and J. Clunis [2020].Here the users geographical location is detected from the tweet that is created by the user which is called geotagging .For each tweet location is detected on hourly, weekly and monthly base of each user and then detects what is the home town of that particular user based on the higher number of tweets in one particular location and it also detects a nonhuman users like if a user send tweets more than 50 in a single day he/she is detected as a non-human. The methodology used here is Latent Dirichlet Allocation model. The limitations here are if a person is staying in a place not his local place some other place and doing the tweets for longer time it will detect that as his local place[1] .

A novel algorithm is proposed called SentDiff proposed by L. Wang, J. Niu and S. Yu[2020].The tweet data is taken from the Beijing Intelligent Star Shine Information Technology Corporation it is a leading service provider for big data in

China .Using SentDiff algorithm they classified all the tweets as positive negative and neutral. Here each tweet is classified according to the sentiment label like positive is given as +1 ,negative is given as 1,neutral is given as 0.and based on the words the algorithm is trained if positive words are there it will be given +1 for that positive words accordingly negative and neutral.so based on that if there are more number of +1 for tweet that tweets are all positive so they can come to a conclusion that that tweet is real or fake[2].

PROPOSED WORK:

The aim of our project is we will be implementing an integrated system which comprises the functionalities of the previous systems but here we will be integrating all the systems to overcome the challenges in each one .Some of the features that we integrated include word cloud plot ,some supervised learning algorithms like xgboost classifier ,logistic regression, decision tree and compare the accuracy of each algorithm for our system and finally we show the accuracy of each algorithm and which algorithm is efficient, robust, effective, accurate to our system. We will also use some of the visualization techniques like word cloud plot, bar plots and we will also explain the various data pre-processing stages like stemming, tokenization etc. and we will also integrate the unsupervised learning (Lexicon based analysis) with this.

ABOUT DATASET:

DATA LINK Test.csv
https// drive.google.com/file/d/1M0xz5bPjvMP7rmTW4aQt1hVCvIqsl8KN/view?usp=sh aring

Train.csv

https// drive.google.com/file/d/10UbXkCuIgorzHPrCd6mrgUQC8qGfpdqN/view?usp=sh aring

- SIZE OF THE DATASET 4629KB

- NUMBER OF ATTRIBUTES 3

- ATTRIBUTES INFORMATION

i. id The id associated with the tweets in the given dataset

ii. tweets the
tweets collected from colorful sources and having either positive or negative sentiments associated with it iii. marker A tweet with marker'0'is of positive sentiment while a tweet with marker'1'is of negative sentiment.

PREPROCESSING:

Data Preprocessing:

a. Data integration:

For ML to be powerful and investigation to be exhaustive, ventures should use information from the best conceivable assortment of sources. An ML calculation is just pretty much as great as the information used to prepare it. Despite the fact that there is a wealth of big business information, quite a bit of it is as yet difficult to track down or use. This kind of information is called dim information. Undertakings are battling to illuminate this dull information and utilize it.

Information joining frameworks have adapted to the situation, bringing about a rise of a few information lists and information lake the board items. These items try to be the "Google for big business information" and offer a basic inquiry-based interface to discover and investigate all the information in an endeavor while regarding existing access control systems.

Information mix frameworks are progressively hoping to utilize ML based methodologies for finding and featuring the islands of helpful information in the huge expanse of dull information (and in this manner improve investigation). Metadata is acquiring a more grounded accentuation and is being caught unequivocally or induced with assistance of ML. A few models are the utilization of ML in the derivation of diagram, information appropriation, and normal worth examples.

ML calculations are utilized on metadata, social setting, and operational attributes to recognize exact, clean, and significant information for different investigation works out. For instance, with regards to information indexes, bunching calculations can be utilized to bunch comparable informational collections, and afterward community sifting calculations can be utilized to suggest the more valuable ones among them in every unique situation.

Likewise, with regards to information security, characterization calculations can be utilized to consequently distinguish delicate information and ensure them utilizing a fitting plan in a strategy driven way. These a few instances of how information incorporation frameworks are applying ML to improve investigation. Each element of information the executives is advancing to account for applying ML to improve the entire cycle.

ML and information reconciliation making each other more powerful is a genuine illustration of an advantageous framework. This is only the start of what vows to be an energizing excursion.

b. Data Transformation

Data Transformation is the interaction where you take information from its crude, siloed and standardized source state and change it into information that is consolidated, dimensionally demonstrated, de-standardized, and prepared for examination. Without the correct innovation stack set up, information change can be tedious, costly, and monotonous. By the by, changing your information will guarantee greatest information quality which is basic to acquiring precise investigation, prompting important bits of knowledge that will in the long run enable information driven choices.

Building and preparing models to deal with information is a splendid idea, and more undertakings have received, or plan to convey, AI to deal with numerous pragmatic applications. However, for models to gain from information to make important forecasts, the actual information should be coordinated to guarantee its examination yield significant bits of knowledge.

c. Data reduction

At the point when you gather information from various information distribution centers for investigation, it brings about an immense measure of information. It is hard for an information examiner to manage this enormous volume of information.

It is even hard to run the mind-boggling questions on the enormous measure of information as it requires some investment and now and then it even gets difficult to follow the ideal information.

This is the reason diminishing information gets significant. Information decrease strategy lessens the volume of information yet keeps up the uprightness of the information.

Information decrease doesn't influence the outcome acquired from information mining that implies the outcome got from information mining before information decrease and after information decrease is something very similar (or practically the equivalent).

The solitary distinction happens in the proficiency of information mining. Information decrease builds the effectiveness of information mining. In the accompanying segment, we will examine the procedures of information decrease.

d. Data cleaning

Data Cleaning implies the way toward distinguishing the off base, fragmented, incorrect, superfluous or missing piece of the information and afterward altering, supplanting or erasing them as per the need. Information cleaning is considered an essential component of the fundamental information science.

Information is the most significant thing for Analytics and Machine learning. In figuring or Business information is required all over. With regards to this present reality information, it isn't unrealistic that information may contain fragmented, conflicting or missing qualities. Assuming the information is defiled, it might block the interaction or give mistaken outcomes. We should see a few instances of the significance of information cleaning.

- **Stemming:** Stemming is the way toward creating morphological variations of a root/base word. Stemming programs are usually alluded to as stemming calculations or stemmers. A stemming calculation diminishes the words "chocolates", "chocolatey", "choco" to the root word, "chocolate" and "recovery", "recovered", "recovers" lessen to the stem "recover". Stemming is a significant piece of the pipelining cycle in Natural language preparing. The contribution to the stemmer is tokenized words.

- **Tokenization:** Tokenization is the way toward separating text into a bunch of significant pieces. These pieces are called tokens. For instance, we can partition a lump of text into words, or we can isolate it into sentences. Contingent upon the main job, we can characterize our own conditions to partition the information text into significant tokens.

METHODOLOGY:

1. Machine Learning Models:

i. **Logistic Regression:** It utilizes a condition as the portrayal, especially like direct relapse. Info esteems (x) are consolidated straightly utilizing loads or coefficient esteems (alluded to as the Greek capital letter Beta) to anticipate a yield esteem (y). A critical contrast from straight relapse is that the yield esteem being displayed is a paired quality (0 or 1) as opposed to a numeric worth.

The following is a model calculated condition: y = e^(b0 + b1*x)/(1 + e^(b0 + b1*x))

Where y is the anticipated yield, b0 is the predisposition or capture term and b1 is the coefficient for the single info esteem (x). Every segment in your information has a related b coefficient (a steady genuine worth) that should be gained from your preparation information.

ii. **Decision Tree**: Decision Tree is a Supervised learning procedure that can be utilized for both characterization and Regression issues, yet for the most part it is liked for taking care of Classification issues. In a Decision tree, there are two hubs, which are the Decision Node and Leaf Node. Choice hubs are utilized to settle on any choice and have numerous branches, while Leaf hubs are the yield of those choices and don't contain any further branches. The choices or the test are performed based on highlights of the given dataset. It is a graphical portrayal for getting every one of the potential answers for an issue/choice dependent on given conditions. It is known as a choice tree on the grounds that, like a tree, it begins with the root hub, which develops further branches and builds a tree-like design. To construct a tree, we utilize the CART calculation, which represents Classification and Regression Tree calculation. A choice tree essentially poses an inquiry, and dependent on the appropriate response (Yes/No), it further split the tree into subtrees.

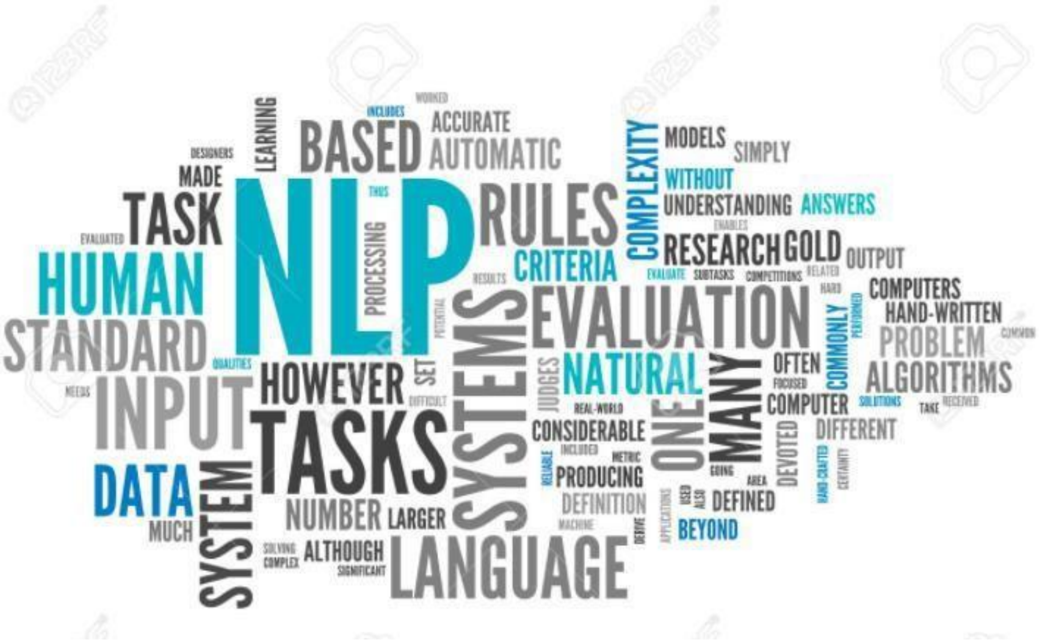
iii. **XGBoost Algorithm:** It is a decision tree-based outfit Machine Learning calculation that utilizes an angle boosting system.In forecast issues including unstructured information (pictures, text, and so forth) counterfeit neural organizations will in general beat any remaining calculations or systems. Notwithstanding, with regards to little to-medium organized/even information, choice tree based calculations are viewed as top tier at the present time. If it's not too much trouble, see the diagram underneath for the development of tree-based calculations over the years. The algorithm separates itself in the accompanying manners: A wide scope of utilizations: Can be utilized to tackle relapse, arrangement, positioning, and client characterized forecast issues. Versatility: Runs easily on Windows, Linux, and OS X. Dialects: Supports all significant programming dialects including C++, Python, R, Java, Scala, and Julia. Cloud Integration: Supports AWS, Azure, and Yarn bunches and functions admirably with Flink, Spark, and different environments.

2. **Visualization:**

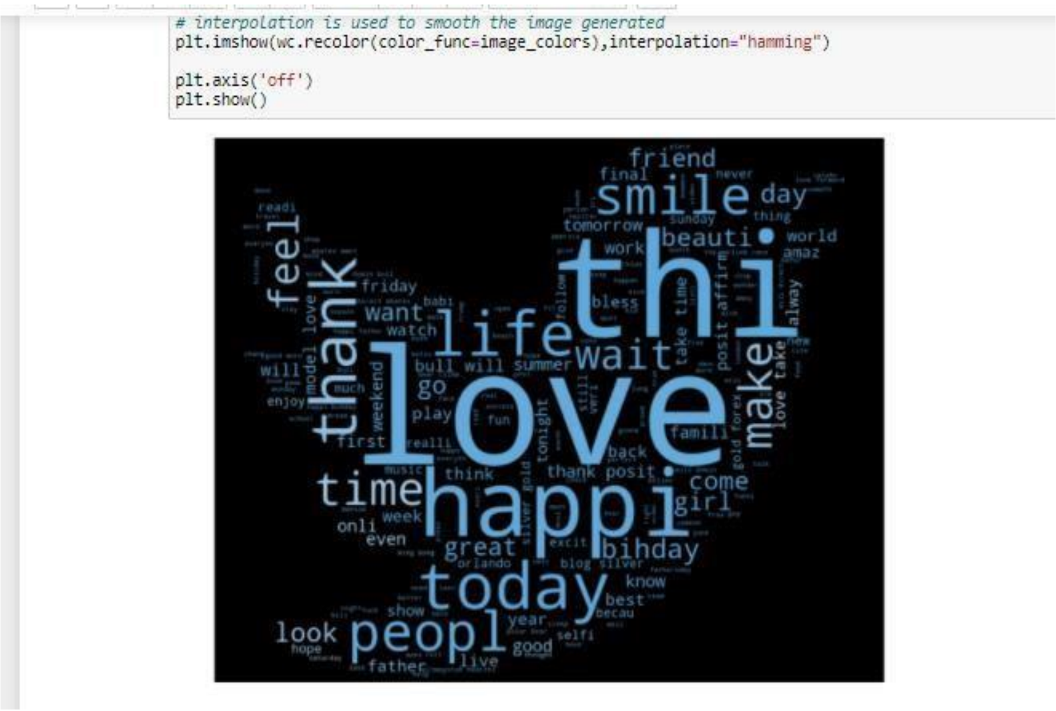
i. **World Cloud Plot:** Word Cloud is a strategy to show which words are the most successive among the given content. The primary thing you might need to do prior to utilizing any capacities is look at the docstring of the capacity, and see all required and discretionary contentions.

```
4   5   0.0   factsguide: society now #motivation        factsguid societi #motiv
```

Visualization from Tweets

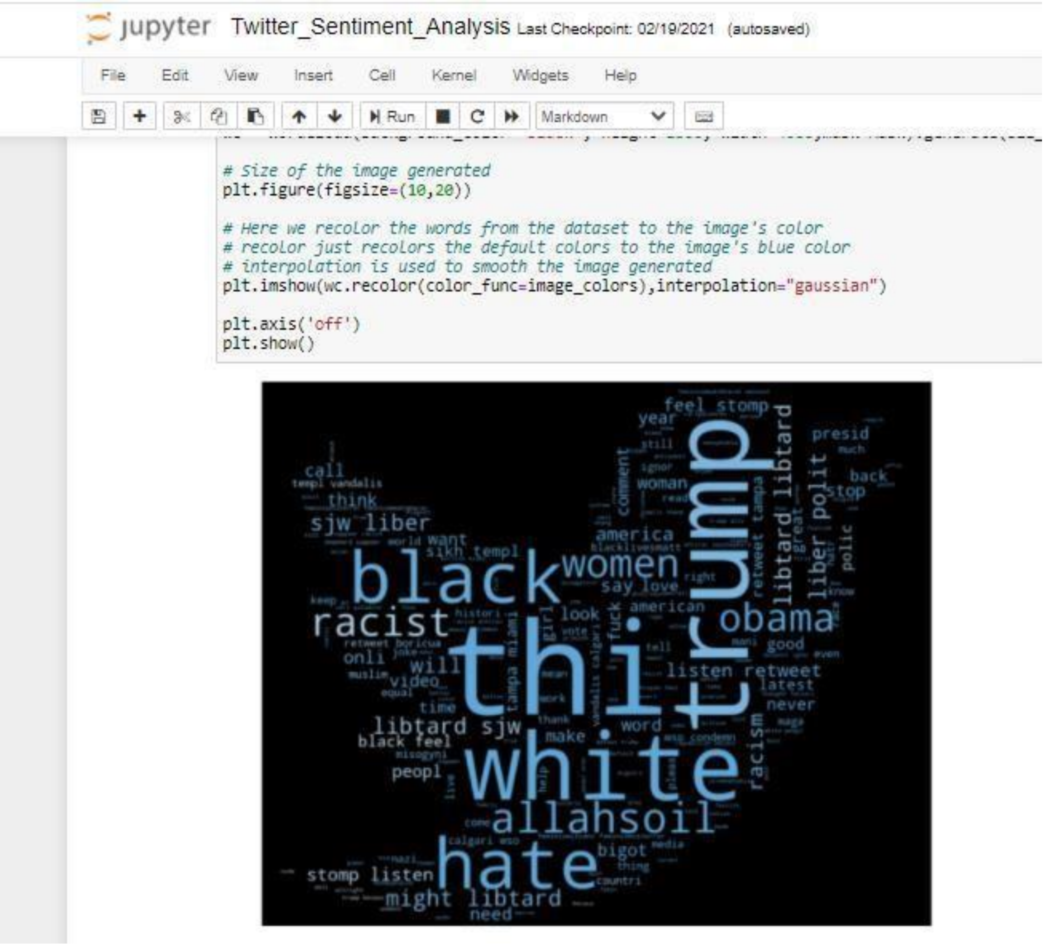


- Positive Plot:



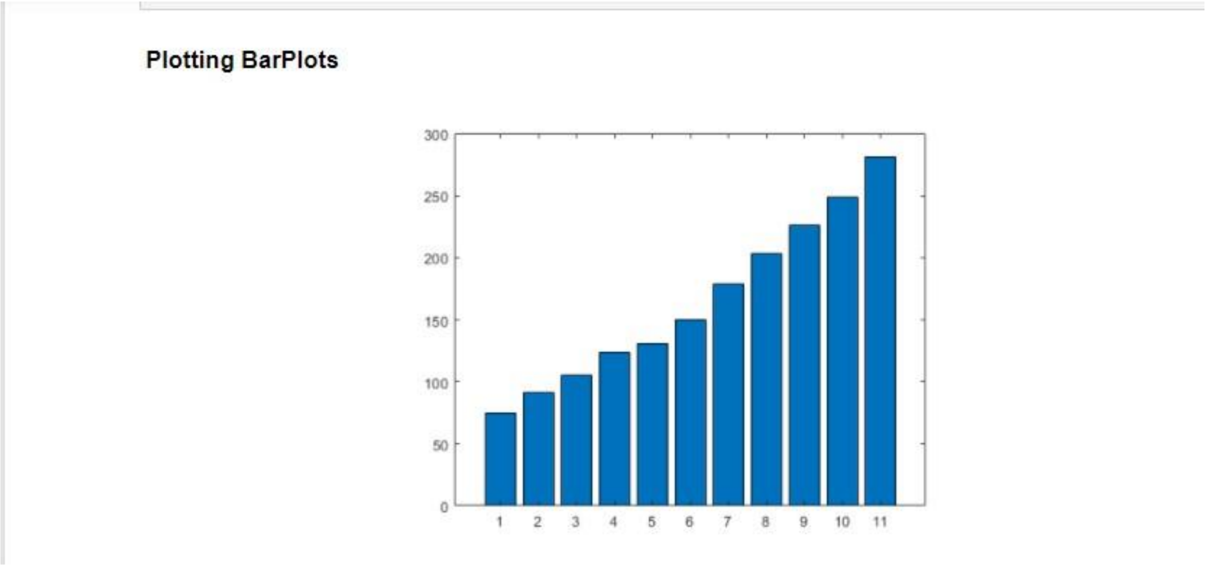
```
# interpolation is used to smooth the image generated
plt.imshow(wc.recolor(color_func=image_colors),interpolation="hamming")

plt.axis('off')
plt.show()
```

- Negative Plot:


jupyter Twitter_Sentiment_Analysis Last Checkpoint: 02/19/2021 (autosaved)
File Edit View Insert Cell Kernel Widgets Help
Run
Markdown
thi
trump
white
black
hate
women
obama
racist
allahsoil
libtard
sjw liber
feel stomp

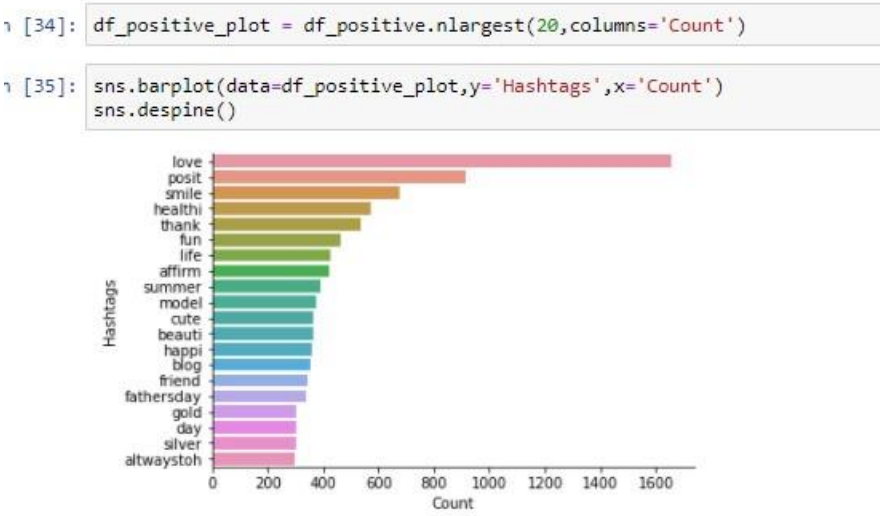
```
# Size of the image generated
plt.figure(figsize=(10,20))

# Here we recolor the words from the dataset to the image's color
# recolor just recolors the default colors to the image's blue color
# interpolation is used to smooth the image generated
plt.imshow(wc.recolor(color_func=image_colors),interpolation="gaussian")

plt.axis('off')
plt.show()
```

ii. **Bar plots**: A bar plot or bar diagram is a chart that addresses the class of information with rectangular bars with lengths and statures that is relative to the qualities which they address. The bar plots can be plotted evenly or in an upward direction. A bar graph depicts the examinations between the discrete classes.



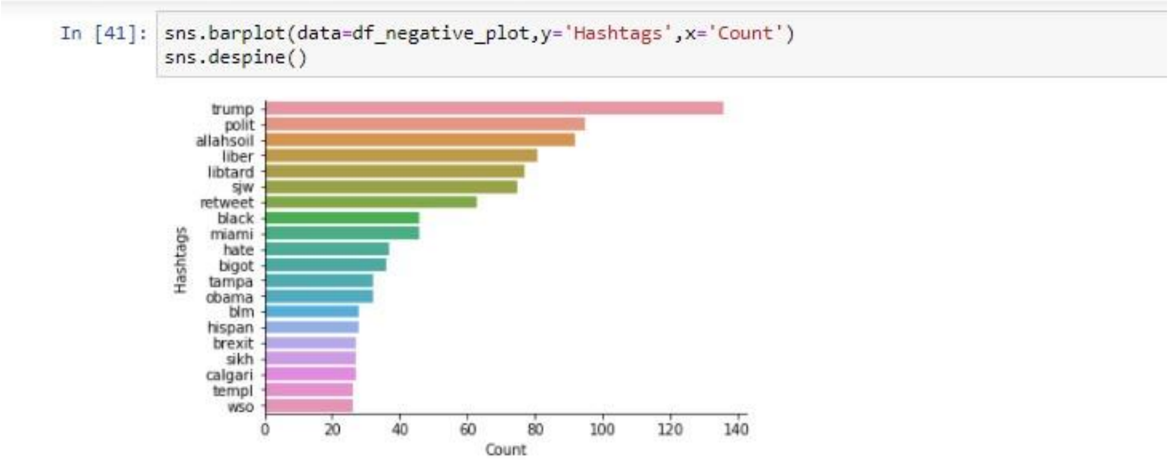
Plotting BarPlots

- Most frequently used positive words:



```
ı [34]: df_positive_plot = df_positive.nlargest(20,columns='Count')

ı [35]: sns.barplot(data=df_positive_plot,y='Hashtags',x='Count')
        sns.despine()
```

- Most frequently used negative words:

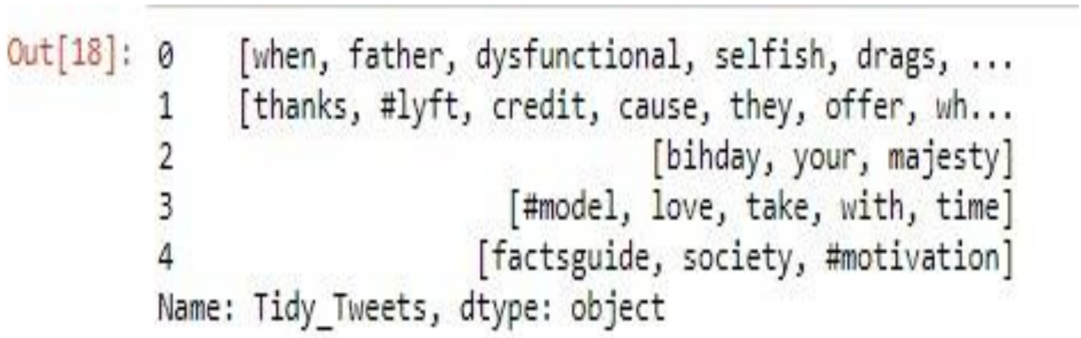


```
In [41]: sns.barplot(data=df_negative_plot,y='Hashtags',x='Count')
         sns.despine()
```

RESULTS AND DISCUSSIONS:

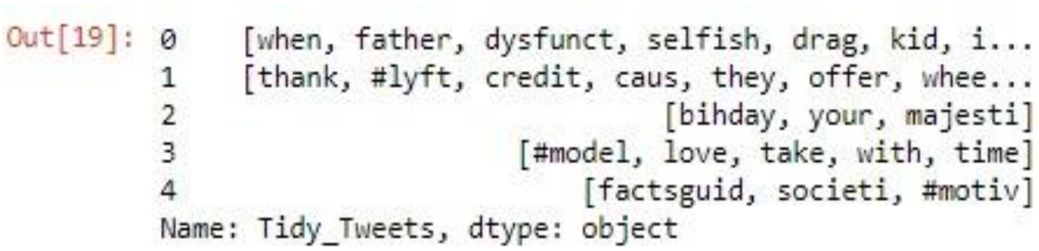
Experimental results:

- TOKENIZATION:



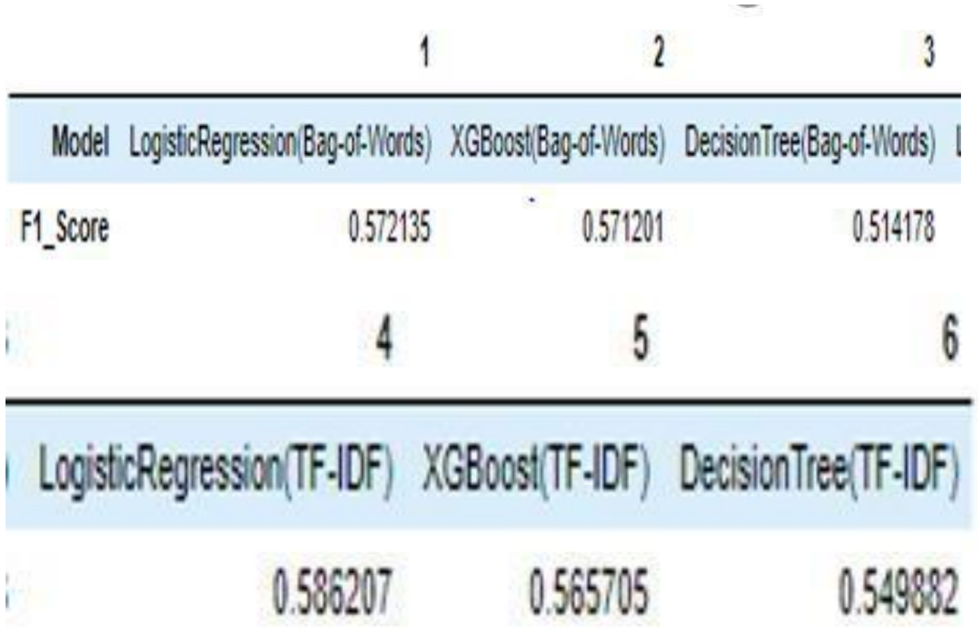
```
Out[18]: 0    [when, father, dysfunctional, selfish, drags, ...
         1    [thanks, #lyft, credit, cause, they, offer, wh...
         2                              [bihday, your, majesty]
         3                     [#model, love, take, with, time]
         4                 [factsguide, society, #motivation]
         Name: Tidy_Tweets, dtype: object
```

- STEMMING:



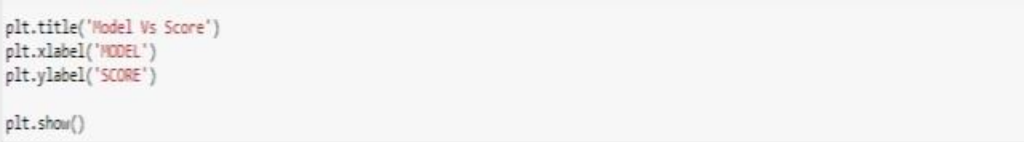
```
Out[19]: 0    [when, father, dysfunct, selfish, drag, kid, i...
         1    [thank, #lyft, credit, caus, they, offer, whee...
         2                              [bihday, your, majesti]
         3                     [#model, love, take, with, time]
         4                          [factsguid, societi, #motiv]
         Name: Tidy_Tweets, dtype: object
```

- IDENTIFYING THE BEST METHOD FOR PREDICTION:

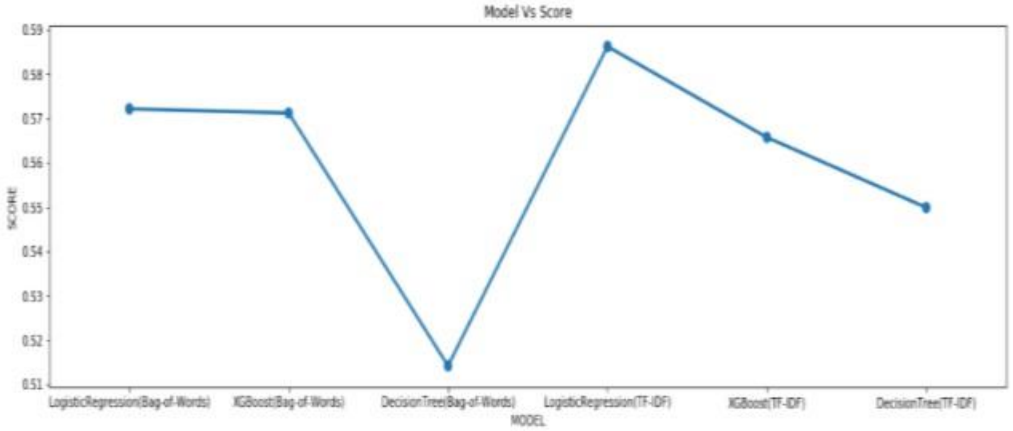


	1	2	3	4	5	6
Model	LogisticRegression(Bag-of-Words)	XGBoost(Bag-of-Words)	DecisionTree(Bag-of-Words)	LogisticRegression(TF-IDF)	XGBoost(TF-IDF)	DecisionTree(TF-IDF)
F1_Score	0.572135	0.571201	0.514178	0.586207	0.565705	0.549882

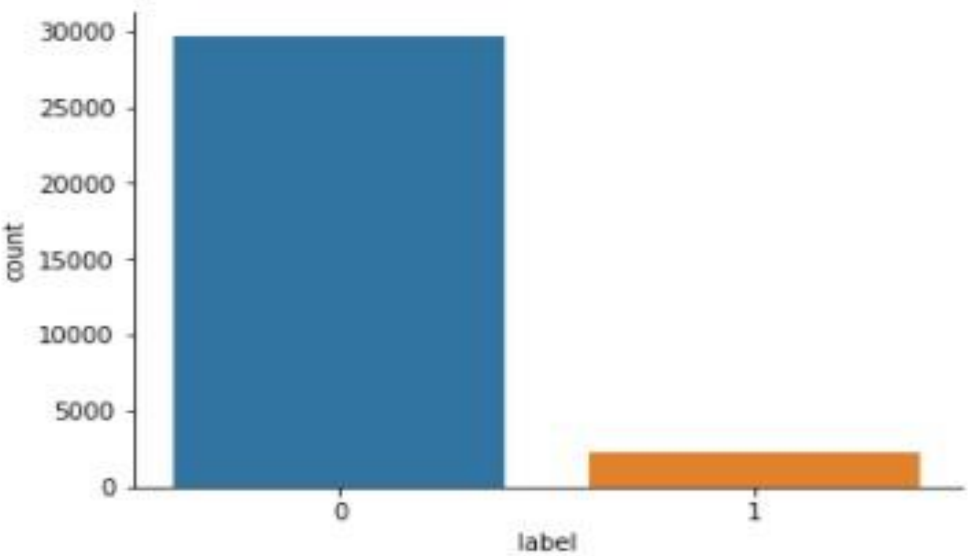
- MODELS COMPARISON:



```
plt.title('Model Vs Score')
plt.xlabel('MODEL')
plt.ylabel('SCORE')

plt.show()
```



- So logistic regression gives the best results and using logistic regression the output will be as follows: And the total count of the positive and negative words will be as follows:



From the given dataset we were able to predict on which class i.e., Positive or Negative does the given tweet fall into. and the various preprocessing stages that are performed are Removing Twitter Handles(@user), Removing punctuation, numbers, special characters Removing short words i.e., words with length them we have analyzed logistic regression could be the best one. And for the evaluation metrics instead of accuracy we used F1 score to predict which model could be best.

REFERENCES:

[1] T. Hu, B. She, L. Duan, H. Yue, and J. Clunis, “A systematic spatial and temporal sentiment analysis on geo-tweets,” IEEE Access, vol. 8, pp. 8658–8667, 2020, doi: 10.1109/ACCESS.2019.2961100.

[2] L. Wang, J. Niu, and S. Yu, “SentiDiff: Combining Textual Information and Sentiment Diffusion Patterns for Twitter Sentiment Analysis,” IEEE Trans. Knowl. Data Eng., vol. 32, no. 10, pp. 2026–2039, 2020, doi: 10.1109/TKDE.2019.2913641.

[3] S. E. Saad and J. Yang, “Twitter Sentiment Analysis Based on Ordinal Regression,” IEEE Access, vol. 7, pp. 163677–163685, 2019, doi: 10.1109/ACCESS.2019.2952127.

[4] Z. Jianqiang and G. Xiaolin, “Comparison research on text pre-processing methods on twitter sentiment analysis,” IEEE Access, vol. 5, no. c, pp. 2870–2879, 2017, doi: 10.1109/ACCESS.2017.2672677.

[5] D. Deng, L. Jing, J. Yu, and S. Sun, “Sparse Self-Attention LSTM for Sentiment Lexicon Construction,” IEEE/ACM Trans. Audio Speech Lang. Process., vol. 27, no. 11, pp. 1777–1790, 2019, doi: 10.1109/TASLP.2019.2933326.

[6] N. Chethan and R. Sangeetha, “Sentiment Analysis of Twitter Data to Examine the Movement of Exchange Rate and Sensex,” J. Comput. Theor. Nanosci., vol. 17, no. 8, pp. 3323–3327, 2020, doi: 10.1166/jctn.2020.9179.

[7] M. Alruily and O. R. Shahin, “Sentiment Analysis of Twitter Data for Saudi Universities,” Int. J. Mach. Learn. Comput., vol. 10, no. 1, pp. 18–24, 2020, doi: 10.18178/ijmlc.2020.10.1.892.

[8] R. S., P. V., S. S., and D. Nagpal, “Twitter Data Sentiment Analysis and Visualization,” Int. J. Comput. Appl., vol. 180, no. 20, pp. 14–16, 2018, doi: 10.5120/ijca2018916463.

[9] A. Amalanathan and P. Nikil, “67 Data Preprocessing in Sentiment Analysis Using Twitter Data,” Int. Educ. Appl. Res. J., vol. 03, no. July, pp. 89–92, 2019.

[10] A. Alsaeedi and M. Z. Khan, “A study on sentiment analysis techniques of Twitter data,” Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 2, pp. 361–374, 2019, doi: 10.14569/ijacsa.2019.0100248.

[11] C. Lin et al., “SenseMood: Depression detection on social media,” ICMR 2020 - Proc. 2020 Int. Conf. Multimed. Retr., pp. 407–411, 2020, doi: 10.1145/3372278.3391932.

[12] A. Baltas, A. Kanavos, and A. K. Tsakalidis, “An apache spark implementation for sentiment analysis on twitter data,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 10230 LNCS, pp. 15–25, 2017, doi: 10.1007/978-3-319-57045-7_2.

[13] R. Wagh and P. Punde, “Survey on Sentiment Analysis using Twitter Dataset,” Proc. 2nd Int. Conf. Electron. Commun. Aerosp. Technol. ICECA 2018, no. Iceca, pp. 208– 211, 2018, doi: 10.1109/ICECA.2018.8474783.

Author’s Profiles:

1. Ms.Chilukuri Lakshmi Sowjanya, , Student, B.E. Computer Science and Engineering, Sri Chandrasekharendra SaraswathiViswa Mahavidyalaya deemed to be university, Enathur, Kanchipuram, India.

2. Ms.Lalitha Bairagi, , Student, B.E. Computer Science and Engineering, Sri Chandrasekharendra SaraswathiViswa Mahavidyalaya deemed to be university, Enathur, Kanchipuram, India.

3. Mr.Pavan Sampath, Student, B.E. Computer Science and Engineering, Sri Chandrasekharendra SaraswathiViswa Mahavidyalaya deemed to be university, Enathur, Kanchipuram, India.

4. Dr. C.K. Gomathy is an Assistant Professor in Computer Science and Engineering at Sri Chandrasekharendra SaraswathiViswa Mahavidyalaya deemed to be university, Enathur, Kanchipuram, India. Her area of interest is Software Engineering, Web Services, Knowledge Management and IOT.