Name: V Hanvighna
Roll Number: 103120123

REPORT

Sentiment analysis on Donald Trump Tweets

LIBRARIES USED:

- Numpy: Used for working on arrays. Useful in domains of algebra and matrices.
- Pandas: Used for analysis of data and manipulation of dataframes.
- Seaborn: Used to plot graphs and visualize data.
- Matplotlib: Used for plotting various types of visualizations.
- Textblob: Used to perform analysis and operations of text data and provides to divide text into NLP.
- Nltk: Used for processing text data, tokenization of data, classification, stemming etc.
- Wordclouds: Used to visualize frequent words in an unstructured data in a visual and expansive format.
- Sklearn: Used to perform regression, classification, clustering and dimensionality reduction of dataset.
- Re: Used to check if a particular string matches the given regular expression.

DATASET USED:

Donald Trump tweets dataset

CODE EXPLAINED:

1. First, I imported the required libraries.
2. Using pandas, I read the dataset which is in csv format and I used .head() to understand the pattern of the dataset.
3. Then, I checked for null values in all columns to determine the significant columns. I came to the conclusion that we had no null values, so the dataset was good to proceed.

4. Defined a function named Clean(). This function is used to remove @mentions,#'s,Re-Tweet mentions,links as the do not impact sentiment of the tweet. This function is applied to all rows using .apply()
5. Stemming function is declared. The function makes sure that reduces a word to it's stem data. This helps in data normalization and removes redundancy.
6. This stemming function is once again applied to all rows of dataframe. This time lambda is used inside apply function and it's called for each row.
7. Subjectivity and Polarity functions are defined. Subjectivity will quantify personal and factual. Polarity is used to evaluate the sentiment of the text and provide a -ve score for negative sentiment, +ve score for positive sentiment and 0 for neutral statements.
8. These functions are applied to the entire column and new columns subjectivity and Polarity are created to store the respective values for each corresponding row.
9. A Word cloud is created. Word cloud basically tells us the most used words in tweets. The larger the font of the word, the more common it is.
10. A sent function is declared that provides the output of “positive”, ”negative” or “neutral” based on polarity value.
11. A new column named Sentiment is added and the outputs obtained by applying sent function on each row in the polarity column are stored in it.
12. Pie chart and bar chart are drawn to show the distribution of various entries in Sentiment column. These help to easily visualize the sentiment distribution of the data.
13. The neutral, positive, and negative sentiment values data are filtered separately and sorted. The obtained 3 data frames are used to obtain word clouds for each activity.
14. Countvectorizer function is used to convert the text in a vector form based on the frequency of each data.
15. The data for regression analysis is prepared. X is the text column and Y is the Sentiment column.
16. The train and test data are separated from the data frame and the ratio of train to test data frame is 8:2 and randomized segregation is used for better results and to avoid overfitting.
17. Logistic Regression analysis is performed and the predicted Y, actual Y values are compared using an accuracy score. The accuracy of this model turns out to be 89.78%.
18. Confusion matrix is plotted to evaluate the model. A confusion matrix is a N x N table (where N is the number of classes) that contains the number

of correct and incorrect predictions of the classification model. The rows of the matrix represent the real classes, while the columns represent the predicted classes. The classification report is also evaluated which provides other constants which help for better understanding of accuracy of model.

19. Then implement a grid search approach to increase model’s accuracy. GridSearchCV uses a fit and search method. It is a cross-validation method that finds the optimal values for a given set of parameters.
20. Again, the confusion matrix and classification report is provided for the improved model. The accuracy is now 90.61%.
21. We apply the linearSVC class on the model to increase accuracy. LinearSVC applies a linear kernel function to do classification and tries to find the best fit model by returning the best fit hyperplane.
22. After applying linearSVC, accuracy of model is 91.12%. The confusion matrix and classification report are also documented.
23. Finally, we use GridSearchCV to again fine tune the results. We use degree and gamma in addition to provide hyper-parameter tuning to the model. The confusion matrix and classification report are computed.
24. Finally, the accuracy of the model is 91.12%.