Great Learning

AIML | MODULE PROJECT

Statistical NLP

TOTAL SCORE 60

- **DOMAIN:** Media
- **CONTEXT:** The Social Dilemma, a documentary-drama hybrid explores the dangerous human impact of social networking, with tech experts sounding the alarm on their own creations regarding the dangerous human impact of social networking. This dataset brings you the Twitter responses made with the #TheSocialDilemma hashtag after watching the eye-opening documentary "The Social Dilemma", released on Netflix on September 9th, 2020.
- **DATA DESCRIPTION:** The dataset was extracted using TwitterAPI, consisting of nearly 20068 tweets from Twitter users all over the globe.
- **PROJECT OBJECTIVE :** The film's argument suggests that algorithms and artificial intelligence are increasingly adept at understanding who we are, and are leveraging this knowledge to curate our reality as well as influence our thoughts and decisions. The task is to analyse the tweets with the trending #TheSocialDilemma hashtag made by the users of twitter and identify their sentiments about the documentary

- **STEPS AND TASKS**: **[Total Score: 60 Marks]**

1. **Read and Analyse Dataset. Clearly write outcome of data analysis (20 marks)**
   A. Clean the Structured Data (2 marks)
      I. Missing value analysis and imputation. After checking for missing values, Write a function to plot the missing values in each column. [1 mark]
      II. Eliminate Non-English textual data. [1 mark]
   B. Write a custom function to plot the count of unique functions in every column. (1 mark)
   C. plot for Social Dilemma Sentiment Labels (1 mark)
   D. Plot and identify the top 20 users, user sources, user locations by number of tweets (3 marks)
   E. Take the top 50 user locations based on no of tweets and try to make the format into city, country for these locations. Incase if only city is present we try to map it to the country from the previous data available (3 marks)
   F. Plot the count of tweets from every place identified above. (1 mark)
   G. Get the number of Hashtags present in each tweet and plot the distribution of number of hashtags in tweet. (2 marks)
   H. Plot the daily and hourly distribution of tweets. (1 marks)
   I. Identify the number of users created every year and plot the distribution. (1 marks)
   J. Find the top 10 hashtags used in the tweet (1 marks)
   K. Get the number of words in each text and plot the distribution of number of words for each class. (2 marks)
   L. Plot the word cloud for negative and positive tweets and write your inferences. (2 marks)

2. **Data preparation. (15 marks)**

   Pre-process the data using various techniques and libraries

   A. Eliminate All special Characters and Numbers (2 marks)
   B. Remove html tags (1 mark)
   C. Replace contractions in strings e.g. replace I'm --> I am) and so on. (2 marks)
   D. Remove the URL's (2 marks)
   E. Remove the mentions in the tweets ('@'). (2 marks)
   F. Remove all Stopwords (2 marks)
   G. Lowercase all textual data (1 mark)
   H. Perform tokenization, lemmatization, normalization appropriately. (2 marks)
   I. Remove the hashtags (1 mark)

3. **Build a base Classification model (25 marks)**

   A. Create dependent and independent variables (1 mark)

      Hint: Treat 'airline sentiment' as a Target variable.

   B. Split data into train and test. (2 mark)
   C. Vectorize data using any one vectorizer, so that we can feed the data in the model. (2 mark)
   D. Build a base model for Supervised Learning - Classification. (3 marks)
   E. Clearly print Performance Metrics. (3 marks)
   F. Improve performance of the model and mention which parameter/hyperparameter significantly helped to improve performance and its probable reason (5 marks)
   G. Try at least three different models and identify which model performs best. Print and plot Confusion matirx to get an idea of how the distribution of the prediction is among all the classes. Write your inferences on the same. (5 marks)
   H. Wordcloud of top 40 important features from the final model chosen. 4 marks)