Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:
V Rupesh Kumar Patro: (varnasipatro@gmail.com) Upload dataset to Google Colab · Data Cleanup. · Data cleaning. . EDA · Preprocessing · Model Implementation · Technical Write up . Power Point Presentation **Shashank Maindola: (shashank.ddun@gmail.com)** · Upload dataset to Google Colab · EDA · Correction of data types · Data Visualizations . Feature Engineering · Technical Write up · PowerPoint presentation · Project summary
Please paste the GitHub Repo link.

Github Link:- https://github.com/shashankm10/Covid-19-tweet-sentiment-analysis.git

Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

- In year 2020 the covid-19 caused by SARS-CoV-2 caused pandemic suddenly came out of no where and surged throughout the globe and affected everyday life. Starting with Wuhan and being a highly contagious new disease, the governments across the globe followed the lockdown procedures by imposing travel related curbs and shutting all the economic activities which required travelling and gathering. According to the International Telecommunication Union (ITU), during the COVID-19 pandemic, Internet user ranks grew to 4.9 billion in 2021, from 4.1 billion in 2019. Remote education, remote work and remote health services came within the reach of many. Social media platforms largely became one of the methods for expression and connecting. They saw a huge surge in usage and as per a report, Twitter saw 24% surge(from 152 million to 166 million) in usage during the lockdown (**source:-bloomberg news**). Twitter took the internet by storm by becoming a key platform for some of the world's top experts to contribute to real-time knowledge-sharing and provide input on policymaking. Twitter became a method of expression for people by expressing their feelings regarding the vaccinations' safety and effectiveness. The proposed approach analyzes collected tweets' sentiments for sentiment classification using various feature sets and classifiers. The early detection of COVID-19 sentiments from collected tweets allow for a better understanding and handling of the pandemic. In our dataset we have tweets which are categorized into positive, negative, and neutral sentiment classes. We evaluate the performance of machine learning (ML) classifiers using evaluation metrics (i.e., accuracy, precision, recall, and F1-score).
  Salient features of our dataset:-
- The data contains covid related tweets from 16 March 2020 to 14 April 2020
- There are 6 columns and 41157 rows
- There are 5 sentiments based on the type of the tweet posted by the users.
- While the average characters in tweet are 204.2, the maximum is 355 and minimum is 11 characters.
- Out of the two months data, march has almost 60% of the total tweets.
- People tweeted more on weekdays than the weekends as people are more online for work on weekdays and they check the social media frequently then.
- London has the highest number of tweets while India ranks 8th in terms of tweets.

We are provided with a dataset namely Coronavirus Tweets.. There are 5 features:-

'UserName', 'ScreenName', 'Location', 'TweetAt', 'OriginalTweet', 'Sentiment'

Firstly, we imported the library which were required to process our data, then we mounted the data from the drive link folder. We used following steps for appliance energy prediction using machine learning.

- Exploratory Data Analysis
- Feature Engineering
- Pre Processing
- Model Implementation

Firstly, we conducted Exploratory Data analysis to get meaning full insight. We Broadly categorized our observations into following categories:-

1. Tweets by length
2. Tweets by Time
3. Tweets by Location
4. Tweets by sentiments
5. Most Common Words and Hashtags in tweets

While there were no duplicates, there were almost 20% missing values in the location field hence we imputed them with unknown.

As part of preprocessing, we only selected sentiment of people and analyze User sentiment tweet according to the scenario of Covid-19.
We used TF-IDF method as classic approach of converting input data from its raw format (i.e. text) into vectors of real numbers which is the format that ML models support TF-IDF is calculated as the product of two values: term frequency (TF) and Inverse document frequency (IDF). The term frequency (TF) measures how frequently a term appears in a document. The inverse document frequency (IDF) measures how important a term is in the entire corpus, by penalizing the terms that appear in many documents.

Then we trained our model using various regression models and checked their performance. The using Grid Search we did hyperparameter tuning to have a more better model performance.

Based on the model performance we selected Logistic Regression and we summarised our observations.