Capstone Project Submission

**Instructions:**

1. Please fill in all the required information.
2. Avoid grammatical errors.

| **Team Member’s Name, Email and Contribution:** |
| --- |
| Name: Vikramaditya Sah  Email: [vsah625@gmail.com](mailto:vsah625@gmail.com)  Contribution:   1. Vikramaditya Sah    * Data overview    * EDA    * TEXT PREPROCESSING    * REGRESSION ANALYSIS    * EVALUATION METRICS COMPARISON |
| **Please paste the GitHub Repo link.** |
| Github Link:-  https://github.com/vikram0050/Coronavirus-Tweet-Sentiment-Analysis |
| **Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)** |
| We all know Covid-19 have greatly impacted our life in many ways. There have been good news and bad news all over the world and people have been responding to it on twitter. So my project is about classifying the tweets into positive, negative and neutral (sentimental analysis) using different classification models. At the end we compare their performance and find which one is better.  I started off with data overview, just to understand what’s in the dataset and plan out the steps to get the final result.  Next step was to perform EDA and data preprocessing which had the following steps-   * Removing links/URLs * Removing usernames/@user * Removing punctuations, numbers and special characters * Removing stop words and tokenizing * Removing short words (less than 3 letters) * Stemming * Data encoding -   Changing neutral labels as 0  Combining Positive and extremely positive labels as 1 Combining Negative and extremely negative labels as -1   * Creating word clouds * Vectorization and data split |

| Then, I ran the following classification models-   1. Logistic regression 2. RANDOM FOREST CLASSIFIER 3. XGBoost CLASSIFIER 4. KNN CLASSIFIER 5. SVM CLASSIFIER   I also saved all the evaluation metrics in a data frame.  In the end, I compared all the metrics and came to the following conclusions – MODEL F1 SCORE  Logistic regression 0.8109  SVM CLASSIFIER 0.7680  RANDOM FOREST CLASSIFIER 0.7328  XGBOOST CLASSIFIER 0.6224  KNN CLASSIFIER 0.4079  We can conclude that Logistic regression is the best model for our dataset, followed closely by SVM classifier and Random Forest classifier. XGboost and KNN classifiers did not give a good result compared to others. |
| --- |