PESRONALITY PREDICTION

Data Extraction: We have used the **google search** package to extract twitter id of subject by passing a query based on name and key words given in input file and appended **twitter** key word to it.

Ex: For - Chetan Bhagat, author -> the query is “Chetan bhagat author twitter”.

From the search results we have considered the first result and extracted the twitter user id of the subject. We have used the twitter API and extracted at most 500 tweets tweeted by the subject. As some of the tweets are in different languages, we have used the **google translator package** and converted them into English.

Multiple sentiments Classified Using the Data Extracted: Based on the tweets extracted we predicted

1. Whether the subject is bully or not by predicting each of the tweet is bullying or not by using **bagging classifier** and to train it we have used **Hate speech and offensive language detection dataset**.
2. Whether the subject tweets hate speech or not by predicting each of tweet is having hate speech or not using ‘Random Forest Classifier’. And to train it we have used **Hate Speech Twitter Annotations** dataset.
3. Whether the **subject** is tweeting any personal attack tweets to anyone using Logistic Regression And to train it we have used **Wikipedia talk labels: Personal attack dataset**.
4. Whether the subject is **sexist** or not by predicting each of the tweet using ‘Random Forest Classifier’. We have used **Hateful Symbols or Hateful People** dataset to train it.

Tagging Subject personality into the Big 5 Model basis the scanned content: The extracted tweets of the subject are cleaned by removing some noises like # symbol, RT tag, mentions and URLs. The cleaned data was preprocessed using **BERT** which was a pre trained using a large corpus of sentences. **BERT** gives contextualized token embeddings of the textual data. We used 5 deep learning models each consisting two dense layers with activation function sigmoid, Stochastic gradient descent (often abbreviated **SGD**) as optimizer and binary cross entropy as loss function. We used **myPersonality traits** dataset to tarin the models. The 5 models named O_model, C_model, E_model,A_model, N_model predicts whether the subject's openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism is high or low respectively.

Citations:

- Hate speech dataset and Bullying dataset - https://github.com/zeerakw/hatespeech .
- Dataset to predict if the subject is sexist - https://www.kaggle.com/arkhoshghalb/twitter-sentiment-analysis-hatred-speech .
- Dataset to predict if the subject is attacking is - https://figshare.com/articles/dataset/Wikipedia_Talk_Labels_Personal_Attacks/4054689 .
- Dataset to predict the personality traits of the subject is - https://github.com/yeziyqf/Personality-Prediction-based-on-facebook-posts .