Explanations
25 June 2020

1 Libraries to be Imported/Installed

- os

- re

- pandas

- numpy

- sklearn 0.24 version

- matplotlib.pyplot

- collections

- codecs

- string

- datetime

- lexical_diversity

- spacy

- seaborn

- textstat

- xgboost

- warnings

- imblearn

    - imblearn.over_sampling
    - imblearn.under_sampling

- nltk

    - nltk.corpus
    - nltk.tokenize

- nltk.stem
- nltk.util
- nltk.sentiment.vader

- sklearn

  - sklearn.feature_extraction.text
  - sklearn.preprocessing
  - sklearn.naive_bayes import
  - sklearn.model_selection
  - sklearn.metrics
  - sklearn.svm

2 Steps for Execution

We have used Google Colab to extract features, run models, evaluate and visualize. Some of the libraries for example: imblearn, textstat are not available in Colab, so we have installed them before running the code. For installing packages in Colab only pip can be used as Conda is not preinstalled in Colab. For installing any package we have used command like:

!pip install package name

- Go to directory Colab Notebooks to connect notebook with the drive

- Import and Install packages

- Order of file execution:

  * Feature Extraction.ipynb
  * File final_features_without_null_values.csv is generated
  * Models Execution and Evaluation
    - Naive Bayes Grid Search.ipynb
    - SVM on Original Features.ipynb
    - Random Forest Original Features.ipynb
    - BagofWords.ipynb
  * Dealing with Imbalanced Data.ipynb
  * Gutenberg Visualization.ipynb