Reddit Flair Detection

Team Members:

Nivedita Singhal, Sunaina Chaudhry, Kiran Chawla, Shivika Gupta, Ena Motwani

Objective

Data Acquisition

- To collect the posts from the r/india subreddit making sure to have enough posts belonging to each flair on the subreddit.
- Trying to collect as much data associated with the post as possible (comments, link, timestamp, user handle, comments etc.)

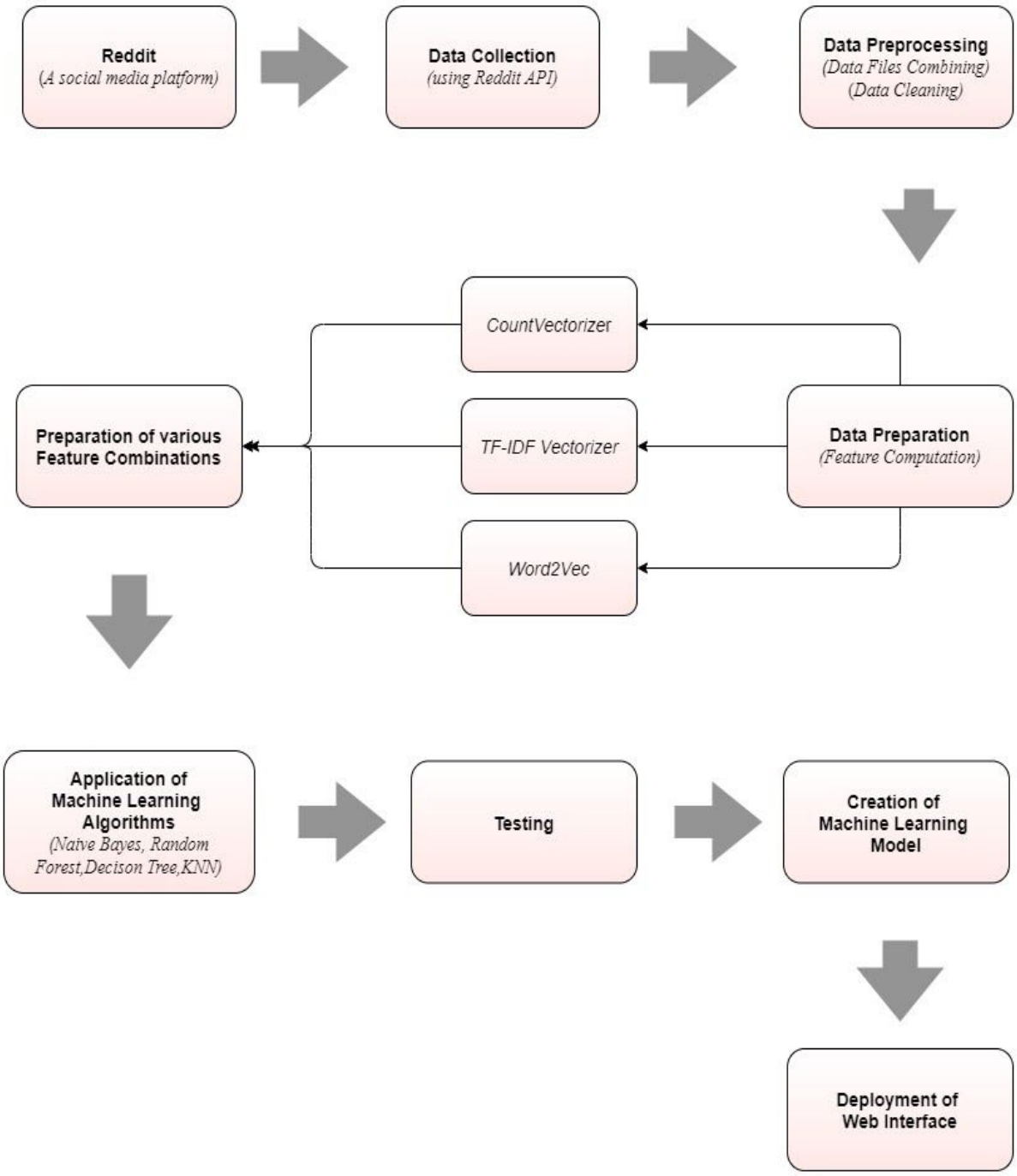
Flair Detection

- We are training a machine learning model to learn the flair prediction from the features of the posts. We will train multiple models using different sets of features (title, comment, link) and report the test set accuracy on each one of them.
- Predicting the flair of the post based on classified flairs on reddit like AskIndia, Food, Business, Political and many more.

Deployment through Web Application

As a final step, we would be deploying the code and the trained model as a web application with a user interface where the user will input a link to a Reddit post and the model would be used to predict the flair for that post.

Proposed Approach



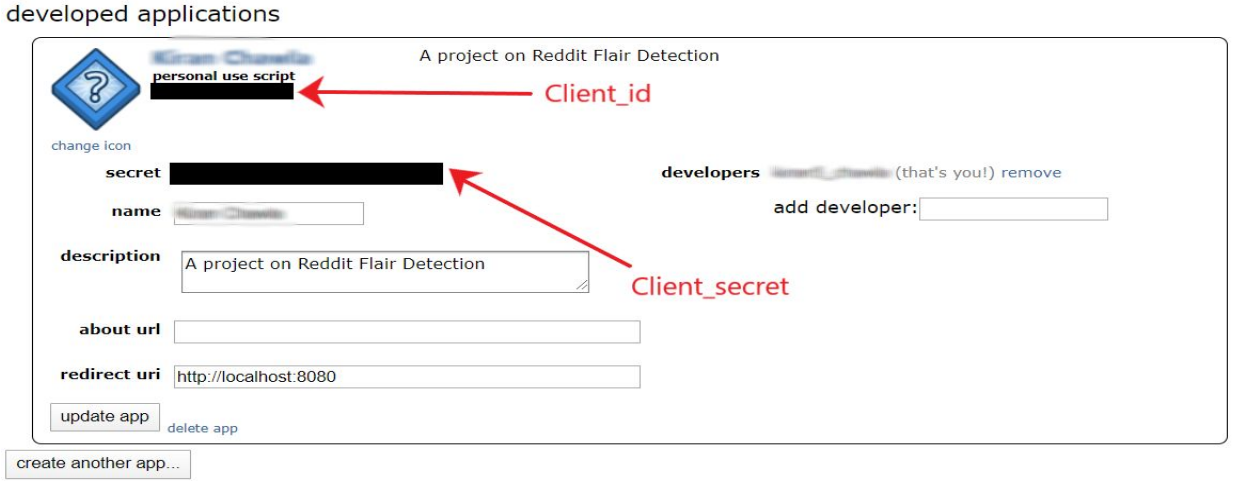
Directory Structure

The description of files and folders can be found below:

Data Collection

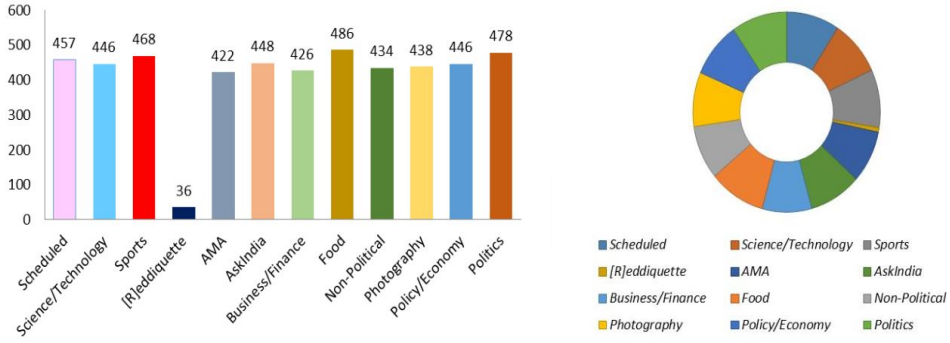
Data is collected using the Reddit API. The application is developed by the programmer, using this link: https://www.reddit.com/login/?dest=https%3A%2F%2Fwww.reddit.com%2Fprefs%2Fapps

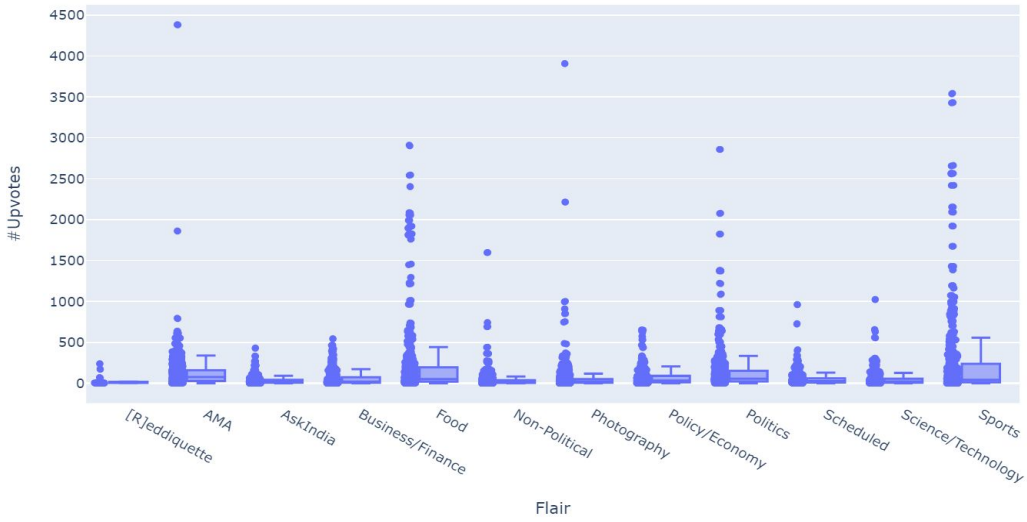
The code for Data Collection is in the folder named “Data Collection”. https://drive.google.com/drive/u/0/folders/1ThoeQTmAg2BrOy1-VH6BJmMZv_QQ83xW

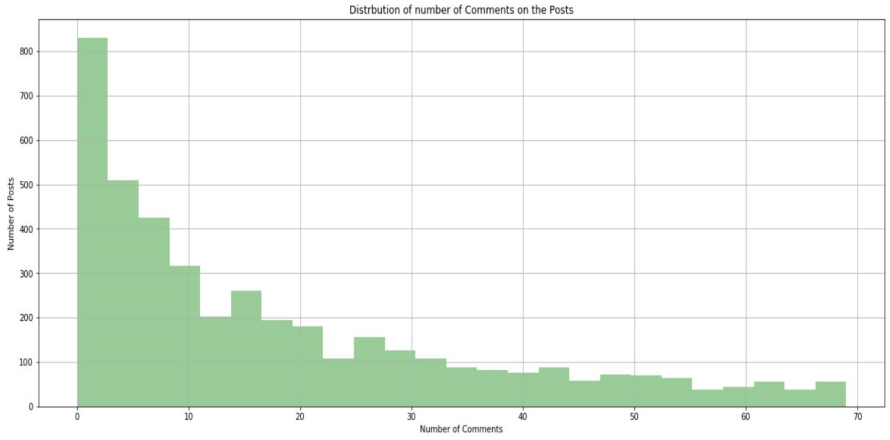


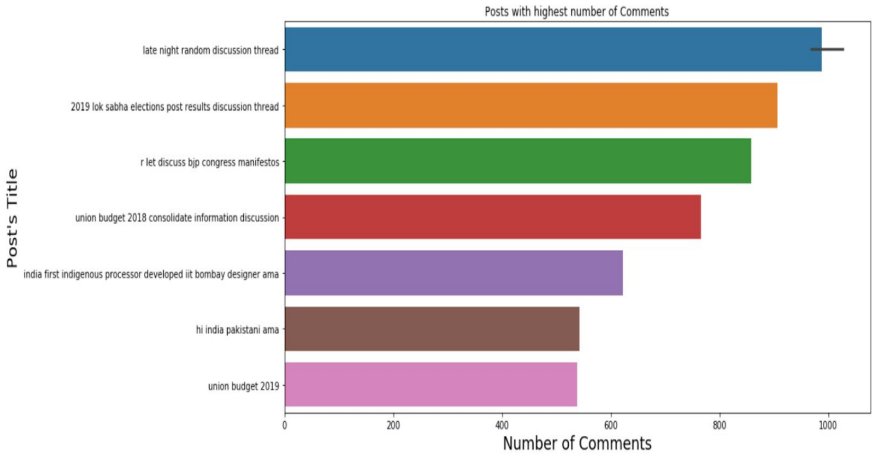
Data Visualization

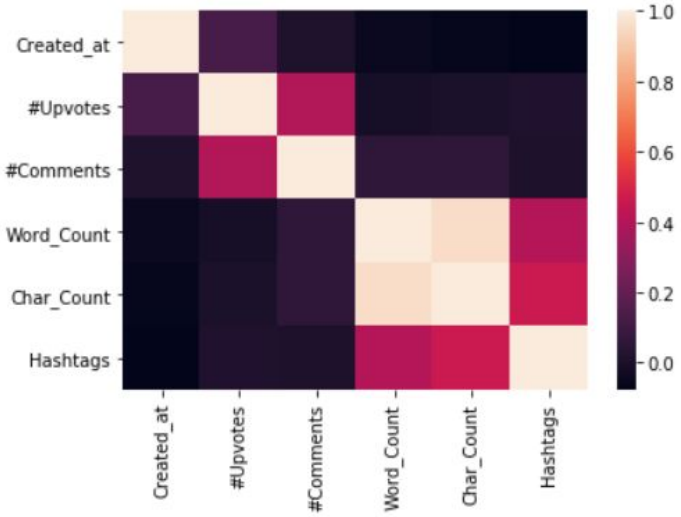
Plots for better understanding of relations among the data attributes. These plots for data visualisation are presented in the folder named “Data Visualisation” https://drive.google.com/drive/u/0/folders/1S1mrm-rJ-vCexX3mF2Rs_WK6-lI_82Wd



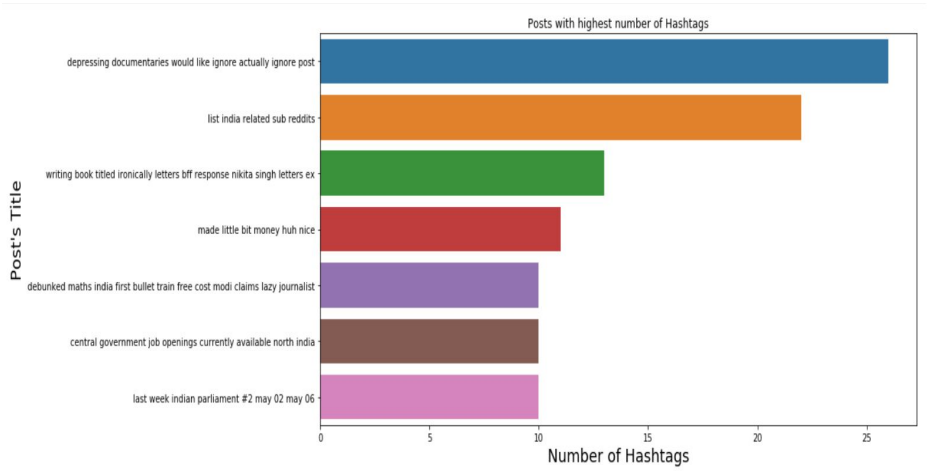








Correlation

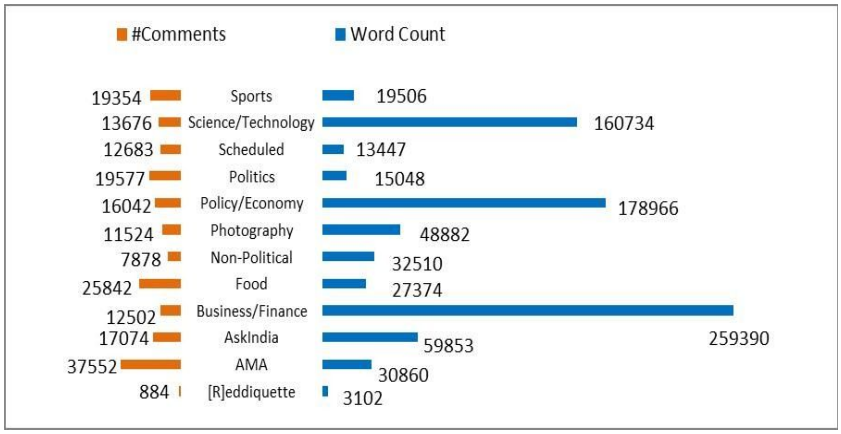


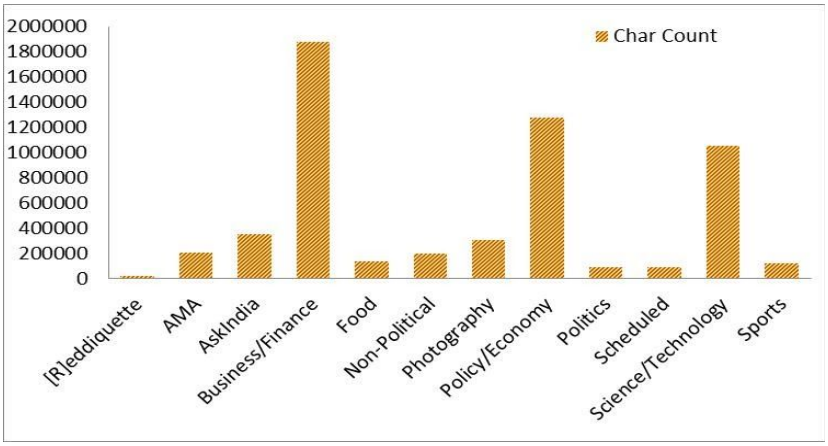
Algorithms

The “Algorithms” folder in the drive has three subfolders named after the vectorizers used for the study, that is, CountVectorizer, Word2Vec,TF-IDF.

The attribute combinations chosen are:

1. Body
2. Body+Comments+URL
3. Title+URL+Comments
4. Title+Body+URL
5. Title+Body+Comments
6. Title+Comments
7. Title+Body
8. Title+URL
9. Body+Comments
10. Body+URL
11. Comments+URL





In all these subfolders, the executed files with various above mentioned attribute combinations are stored in two types of file formats:

.html files, .ipynb files

Link to “Algorithms” folder:

https://drive.google.com/drive/folders/1wHhmR6A5JUNzejLtQ8P-E0gWTITviz-4

Finalised Model

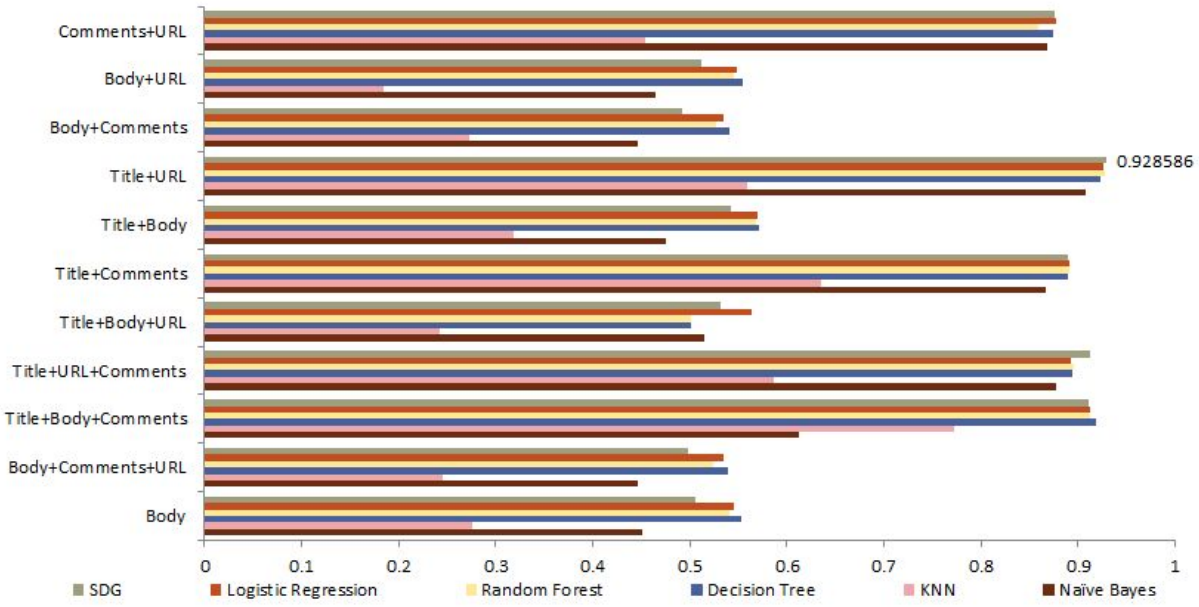
The algorithm, that provides the best accuracy that is SDG for the attribute combination Title+URL (93.18%), is deployed as the model for further process.

Link to"Finalised Model" folder:
https://drive.google.com/drive/folders/1di_N8v60wG6cMY96fzIr9djX5Y4AjIf-

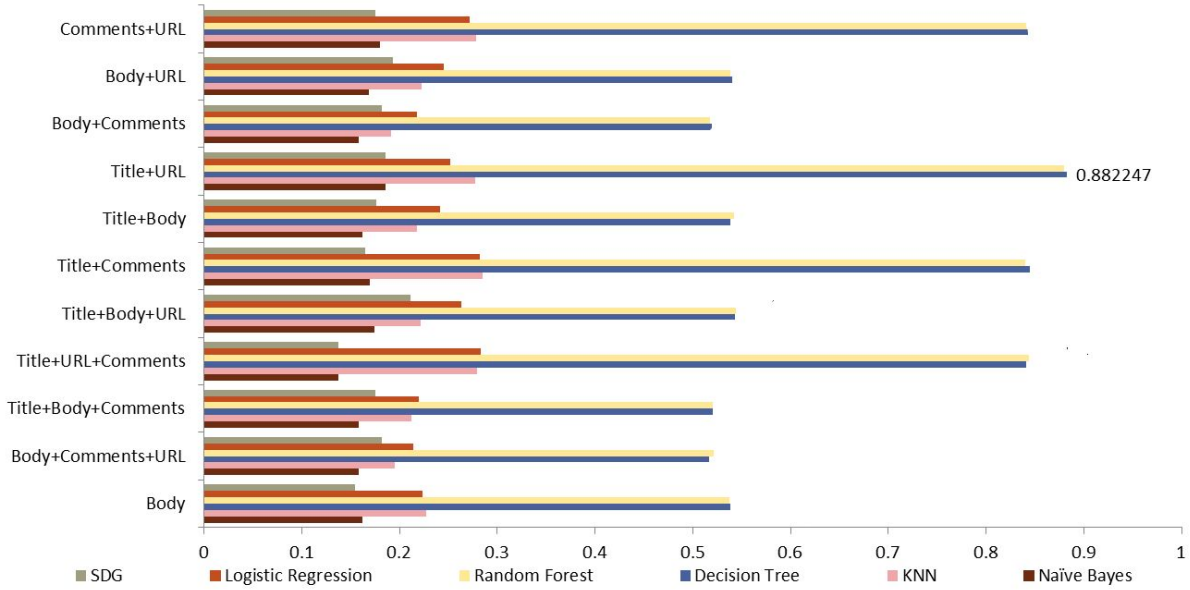
Results

Link to spreadsheet containing Results:
https://docs.google.com/spreadsheets/d/1DcNFGp4qSJMHb7mlkfh_zwrlNx4lWBLc/edit#gid=215286377

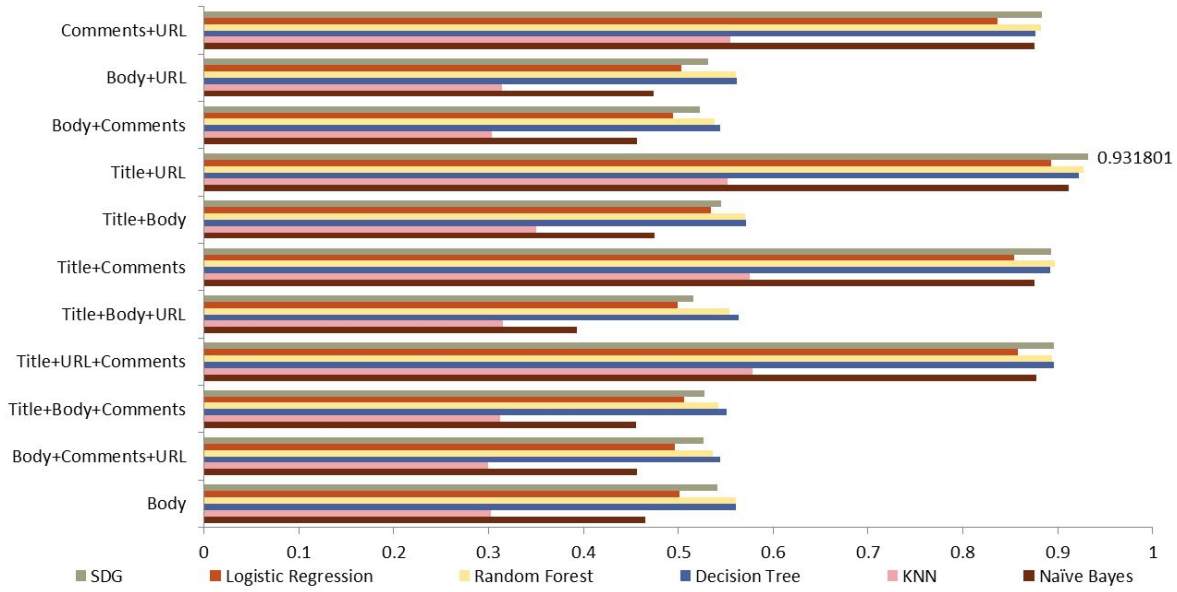
Results using CountVectorizer:



Results using Word2Vec:



Results using TF-IDF:



Web Deployment

Link to the folder “Web Deployment” which contains API and html files:
https://drive.google.com/drive/folders/1Pd3o5UdkGzy5mt3DmikATyKPEw7ET3e2

Report

The link to the report:

https://drive.google.com/drive/u/0/folders/18GWTmzfeA38MoQOIt0M_v-MLsrxv3AqW