Assignment

Problem Statement:

StumbleUpon is a user-curated web content discovery engine that recommends relevant, high quality pages and media to its users, based on their interests. While some pages we recommend, such as news articles or seasonal recipes, are only relevant for a short period of time, others maintain a timeless quality and can be recommended to users long after they are discovered. In other words, pages can either be classified as "ephemeral" or "evergreen".

Data:

There are two components to the data provided for this challenge:

The first component is two files: train.tsv and test.tsv. Each is a tab-delimited text file containing the fields outlined below for 10,566 urls total. Fields for which no data is available are indicated with a question mark.

train.tsv is the training set and contains 7,395 urls. Binary evergreen labels (either evergreen (1) or non-evergreen (0)) are provided for this set. test.tsv is the test/evaluation set and contains 3,171 urls.

Approach:

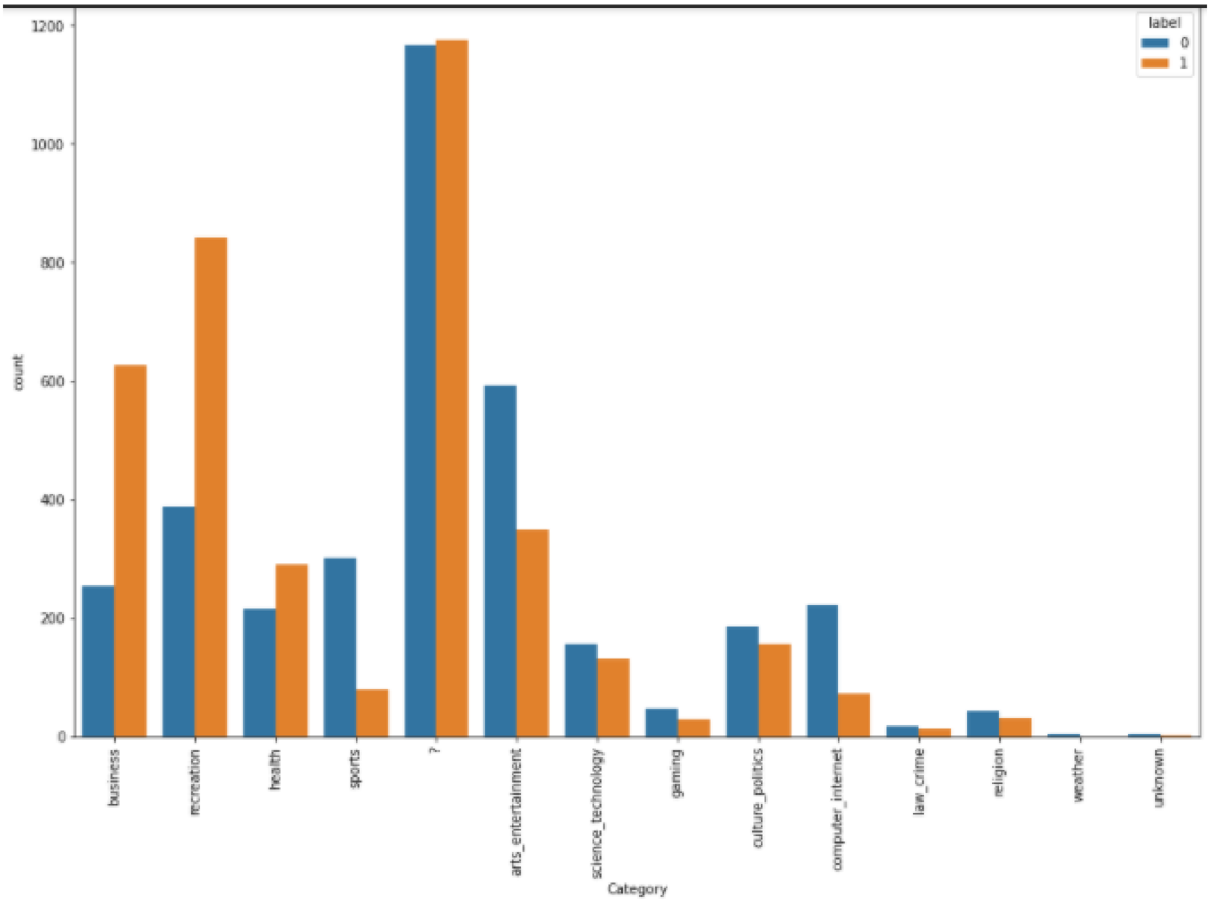
Instead of using transformers, pre-trained models and deep learning I have used traditional and a basic approach for solving this classification problem with the help of Pretrained Word Embeddings and decision tree classifier. This a very basic yet effective way of solving this problem accurately.

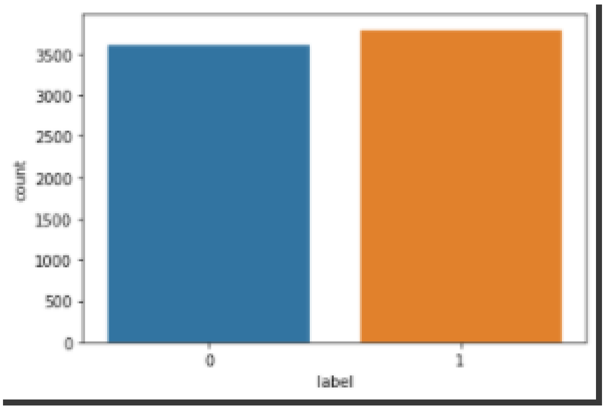
Steps involved –

1. Visualising the data with the help of graphs with seaborn library.
2. Data Cleaning with basic methodologies.
3. Splitting the data
4. Training the model for calculating embeddings with Word2Vec
5. Calculating word embeddings and saving it.
6. Fitting the model
7. Testing with decision tree classifier and calculating precision recall.

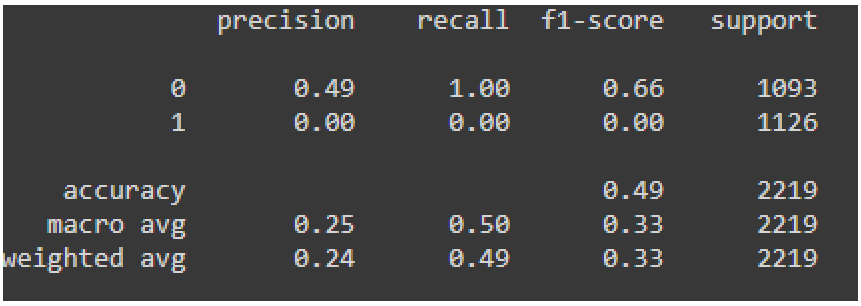
Outcome:

Data Visualisation –





Final Results –



```
              precision    recall  f1-score   support

           0       0.49      1.00      0.66      1093
           1       0.00      0.00      0.00      1126

    accuracy                           0.49      2219
   macro avg       0.25      0.50      0.33      2219
eighted avg       0.24      0.49      0.33      2219
```

References:

1. https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html
2. https://seaborn.pydata.org/tutorial.html
3. https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa