**Project Tittle:**

**“jigsaw Multilingual Toxic Comment Classification" Kaggle competition**

**Personal Details:**

This project is successfully developed by team of 2.

**Team Name: Error-404**

The details of the two members:

**Member 1: Member 2:**

Name: E. Vamsi Krishna Name: Pulugu Yamini

Email: [vamsievk8@gmail.com](mailto:vamsievk8@gmail.com) Email: [myself.yamu@gmail.com](mailto:myself.yamu@gmail.com)

Mobile no: 7989713510 Mobile no: 7358360558

College: Vit University Chennai Campus College: Vit University Chennai Campus

Semester: Six Semester: Six

**Project Overview:**

People tend to leave online conversations due to people posting toxic or disrespectful comments. You need to make a machine learning model to recognize if a comment is normal or toxic. If we can recognize such harmful contributions, we will have a healthier, more open internet.

**Models Used:**

Logistic Regression of bag of words model and Term FrequencyInverse Document Frequency model.

* A bag-of-words representation is simple to generate but far from perfect. If we count all words equally, then some words end up being emphasized more than we need.
* Tf-idf is a simple twist on the bag-of-words approach. It stands for term frequency–inverse document frequency. Instead of looking at the raw counts of each word in each document in a dataset, tf-idf looks at a normalized count where each word count is divided by the number of documents this word appears in. That is:

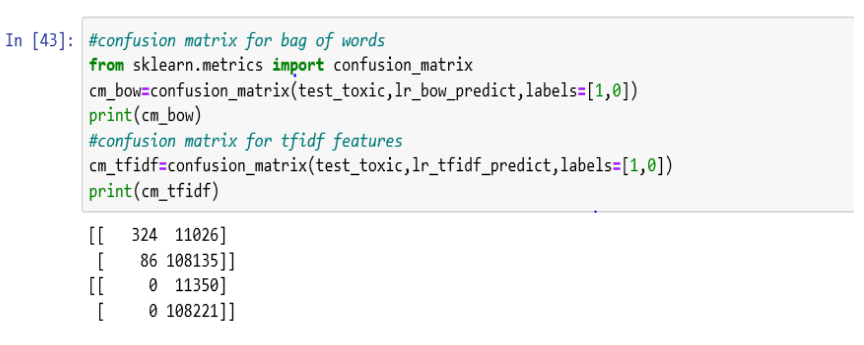
bow(w, d) = # times word w appears in document d

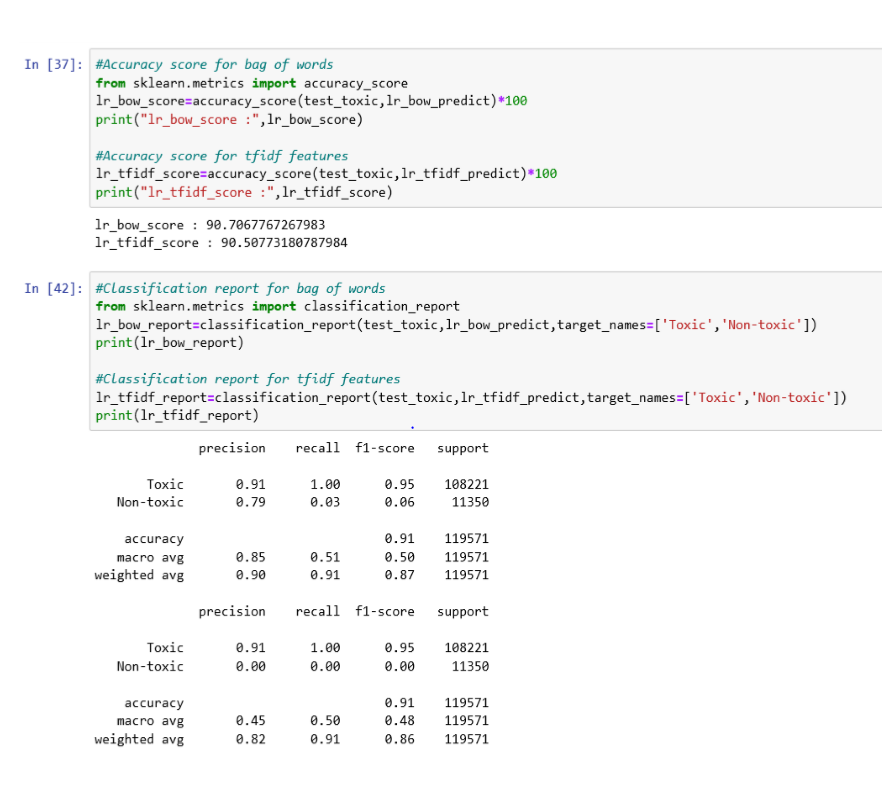
tf-idf(w, d) = bow(w, d) \* N / (# documents in which word w appears)

N is the total number of documents in the dataset. The fraction N / (# documents ...) is what’s known as the inverse document frequency. If a word appears in many documents, then its inverse document frequency is close to 1. If a word appears in just a few documents, then the inverse document frequency is much higher.

Alternatively, we can take a log transform instead using the raw inverse document frequency. Logarithm turns 1 into 0, and makes large numbers (those much greater than 1) smaller. (More on this later.)

**Accuracy Report Screenshot:**





**Conclusion:**

In this project of toxic data classification, using the logistic regression model of machine learning algorithms for Bag of Words model the accuracy is 90.7% and for the Term Frequency-Inverse Document Frequency model the accuracy is 90.5%.

**We hope that you are satisfied with our team work.**

**Thank you!**