

Faculty of Engineering & Technology

Electrical & Computer Engineering Department

ARTIFICIAL INTELLIGENCE

ENCS3340

Automatic Tweet Spam Detection

*Prepared by:*

*ShahEd Jamhour 1180654, Yasmeena Assi 1180899 & Rand Farhoud 1181567*

*Instructor: Dr. Aziz Qaroush*

*Date: January 2022.*

# Abstract

The aim of this project is to learn about machine learning techniques and apply them to real-life applications.

The main task is to develop an automated tweet spam detector by pre-processing the data, extracting features, and building different classifiers (Decision tree, Naive Bays, and Random Forest ). Testing of data will be carried out using a Python program with the help of the Scikit-Learn and using Google Collaboratory.

## 

**Table of Contents**

[Abstract 1](#_heading=h.30j0zll)

[Aim of the experiment 1](#_heading=h.1fob9te)

[Table of Figures 2](#_heading=h.3znysh7)

[Theory 3](#_heading=h.2et92p0)

[Conclusion 4](#_heading=h.tyjcwt)

[Reverences 5](#_heading=h.3dy6vkm)

[Appendix 6](#_heading=h.1t3h5sf)

Introduction

1. Machine learning

The field of machine learning is an extension of artificial intelligence and computer science which uses data and algorithms to imitate human learning and gradually improve its accuracy.

There are three primary categories of machine learning classifiers:

### Supervised machine learning

### Supervised Machine learning is defined by its use of labeled datasets to train algorithms to classify data or predict outcomes accurately. As input data is fed into the model, it adjusts its weights until the model has been fitted appropriately. This type of algorithm is used for classification problems e.g. Decision Tree, Artificial neural network

### Unsupervised machine learning

It uses machine learning algorithms to analyze and cluster unlabeled datasets. These algorithms discover hidden patterns or data groupings without the need for human intervention.

### Semi-supervised learning

### Semi-supervised learning offers a happy medium between supervised and unsupervised learning. During training, it uses a smaller labeled data set to guide classification and feature extraction from a larger, unlabeled data set

2. Automatic Spam Tweet Detection

Recent years have seen an increase in the popularity of social networking sites. Twitter is considered to be the fastest-growing site so spammers take advantage of this popularity by using Twitter for their malicious purposes.

The term spam refers to tweets posted by known fake Twitter accounts that are politically motivated, automatically generated, or clickbait in nature.

With both manual and automated tools, Twitter provides an environment free of spam and According to Twitter, spam detection features fall into three categories: (1) account-based features, (2) tweet-based features, and (3) relationship between the tweet's sender and receiver.

These features include number of hashtags,mentions and URLS the tweet has, the ratio of the number of following and the number of followers, average tweets per day,, etc.

3. Scikit-Learn Library

The Sklearn library is a collection of machine learning algorithms for data mining tasks. With Sklearn, you can preprocess data, classify, predict, cluster, and create association rules. Sklearn supports data preprocessing, classification, and feature selection, which are all standard data mining tasks.

Related work:

# Methodology:

# In this part , we are going to summarize briefly , what we did to the tweets , in order to detect the spam in the tweet.

to do that,these steps were done:

1. raw data was read , then we implemented the pre-processing .

2. feature extraction techniques were implemented .

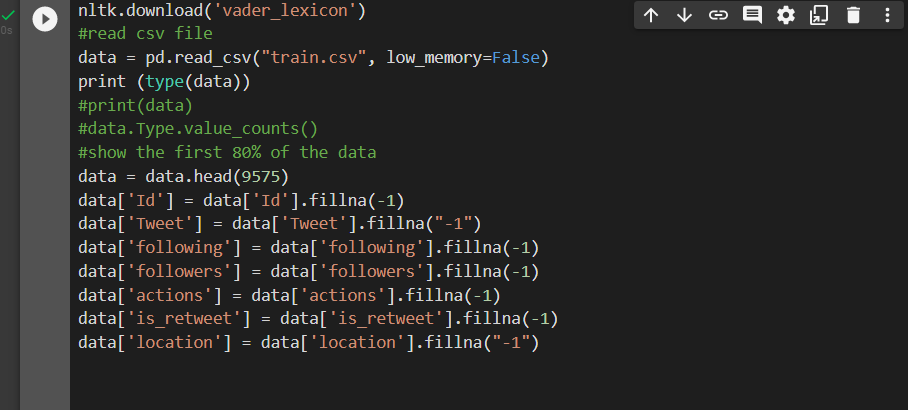
3.we have chosen the model to train our data , and made sure to balance our data , then trained it .

To do the above steps , we used many python libraries , and we used Google colaboratory , because here we have a large number of data, so it provides us a suitable size memory to do the feature extraction . Running python script takes a lot af computing power and a lot of time , but here we allowed to use the google’s server computing power, so we don't need to be worry about time .

Here a description for each step in our project:

1. Reading data

The CSV file which holds the data was read using Pandas Library and stored in the variable “data” as shown in fig below.

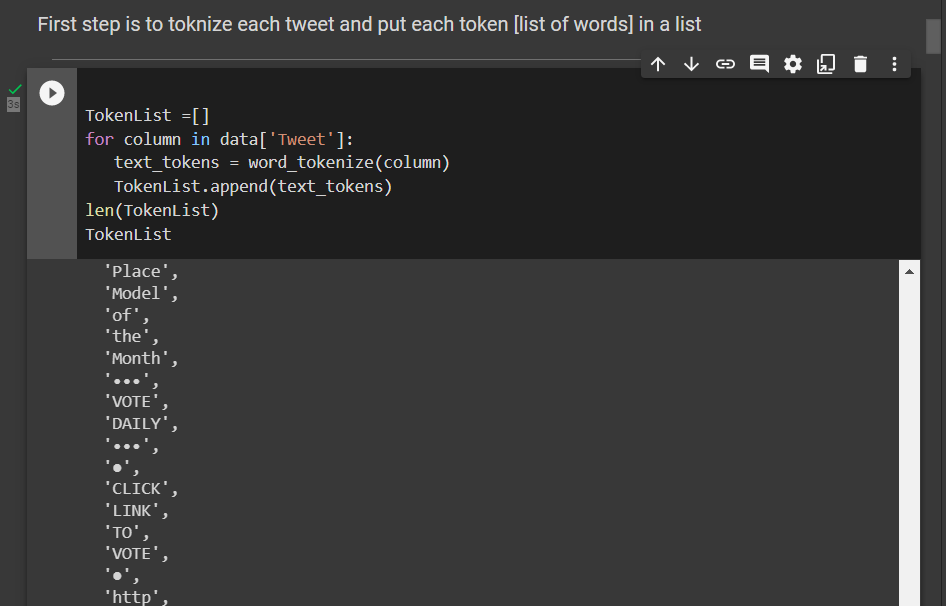


Only 70% of the data( which equals 8378Rows) was trained using data.head function. Then, Nan values were handled by Substituting the Nan value with (-1) using fillna() function as shown above. This step was done because The sklearn implemented algorithms can’t perform on datasets that have such values.

3.1 step 1: Pre-Processing

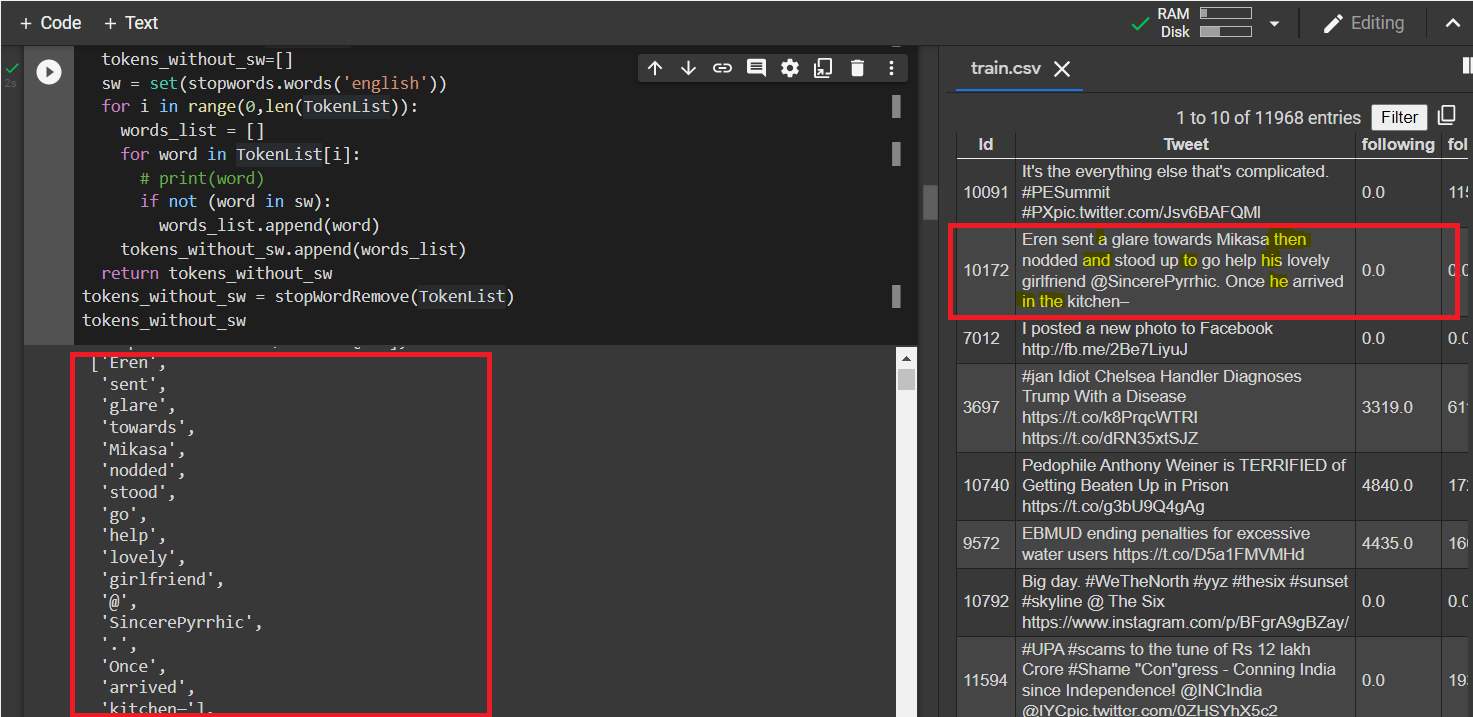
Here in this step , raw data was organized before building the trained model , this step makes us more comfortable with data , also it enhances the quality of the data .So , here are the steps of the pre-processing :

1.Tekonization: in this step, Sensitive data was exchanged for nonsensitive data,Here we are going to divide the text into chunks, called tokens.In order to use it to remove the stop words .



The tweet was tokenized using **word\_tokenized** function. Each word was tokenized as a list element as can be seen in the fig above.

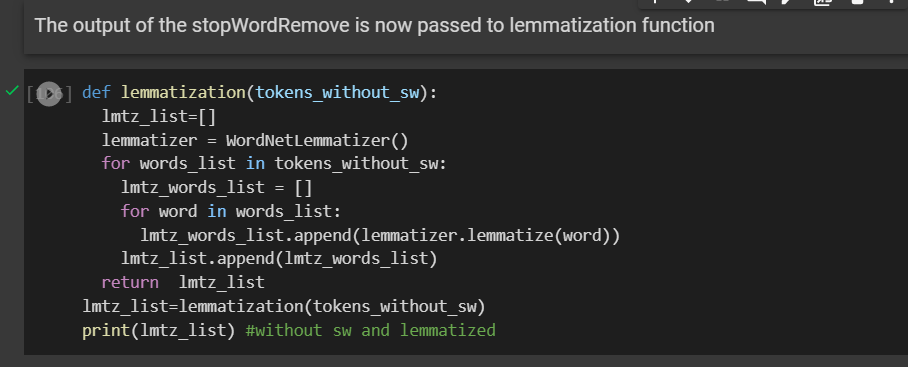
2. Removing stop words from the tweet content: Here in this step we are going to remove the stop words like (are,the,is,am, then , a… etc), from the tweet content , this step is important ,we are going to use it in lemmatization, and when we get the similarity afterwords.



Stop words which are (a, then, and, to, his ,he,in) were removed from the tokenized list using (stopWordRemove) function as can be seen in the above fig.

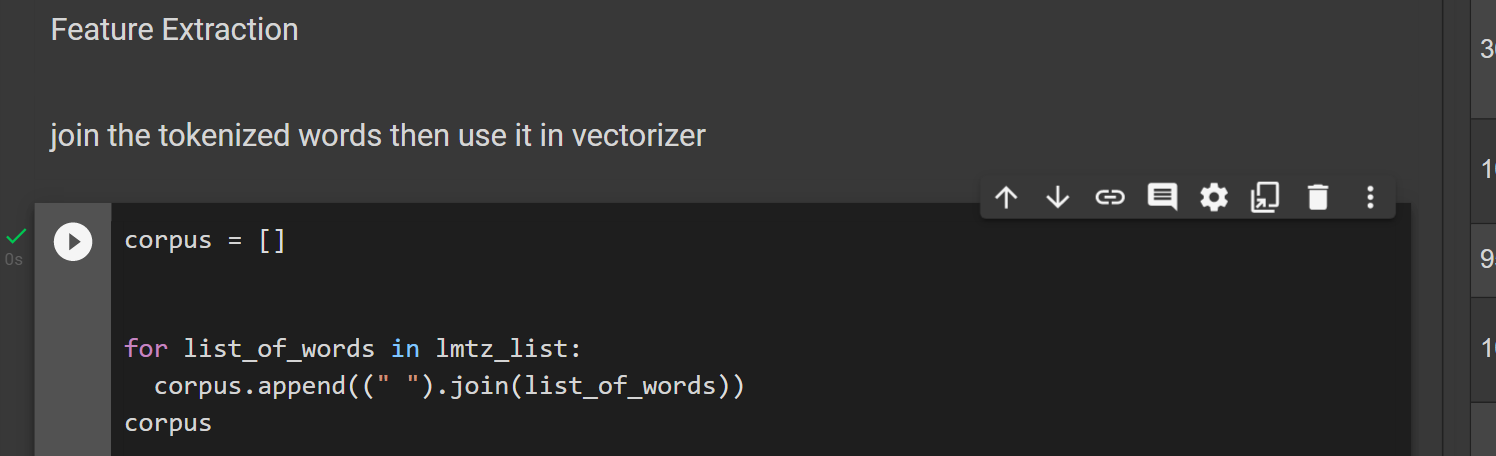
Then it was stored in (tokens\_without\_sw), then passed it to the Lemmatization function.

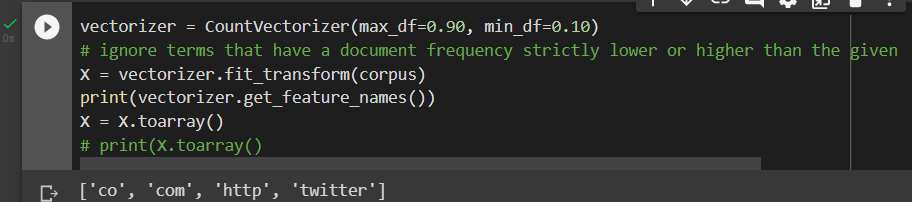
3.Lemmatization:In this step , we are going to group together the different inflected forms ,it normally aims to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma , Here we use the the vocabulary and morphological analysis of words.



2. Feature extraction :

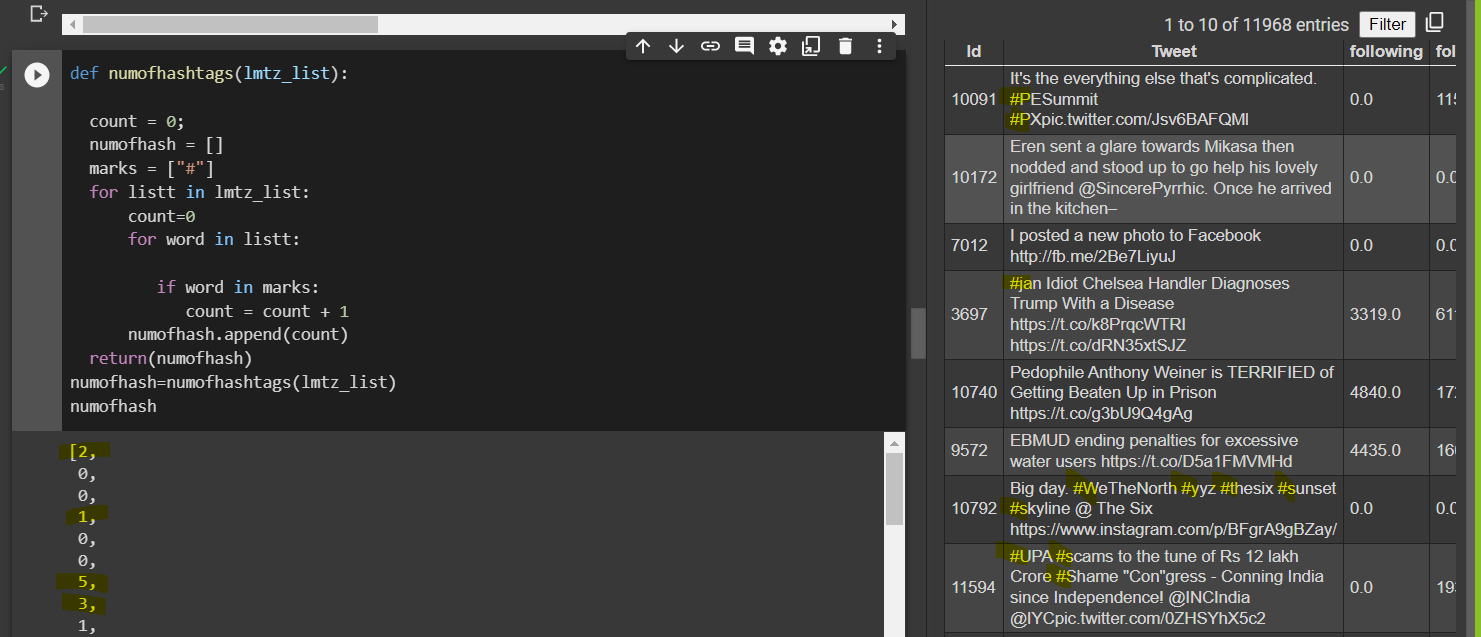
Raw Data includes redundant data, and as our goal is to build a good classifier, we need to get a complete and meaningful set of the needed features. Feature Extraction aims to create and deduce new features from existing ones. Our feature extraction process consists of the following:

1. Rejoin the tokenize tweets ****
2. *CountVectorizer* to create the bag of words array

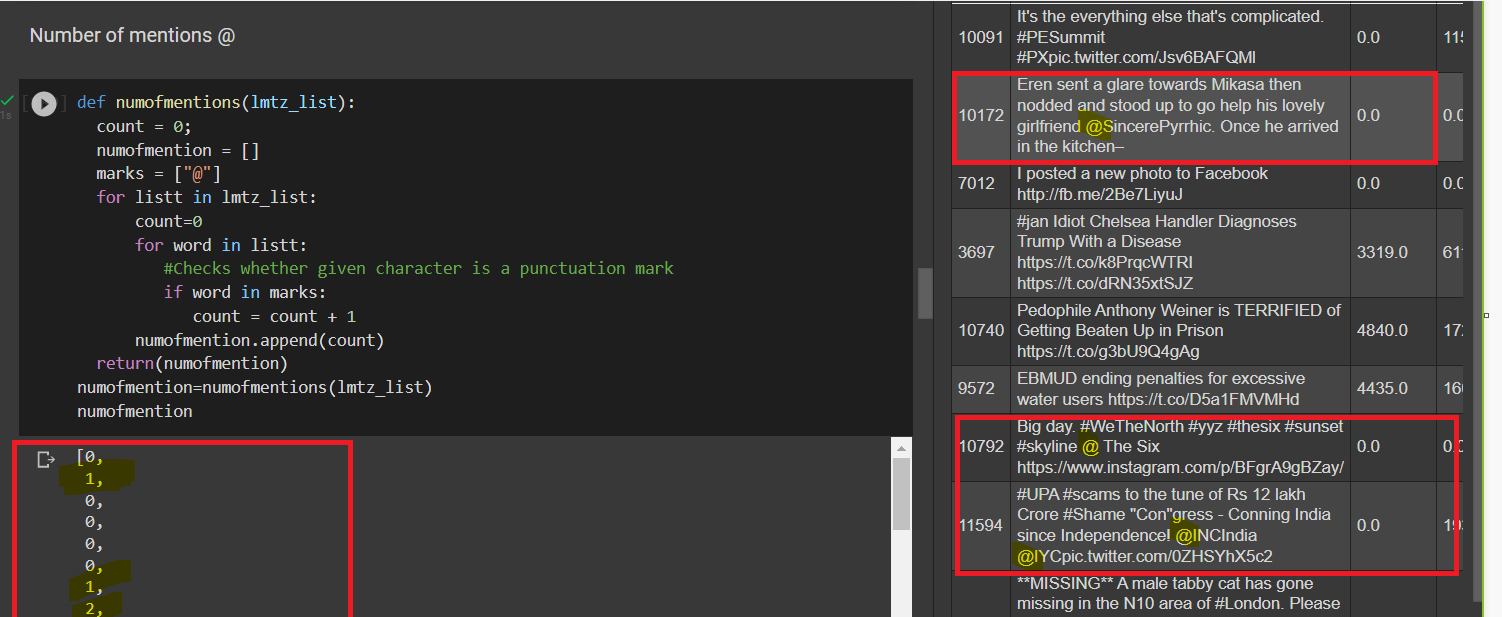


When building the vocabulary ignore terms that have a document frequency strictly higher than the given max threshold (corpus-specific stop words) and ignore terms that have a document frequency strictly lower than the given min threshold.

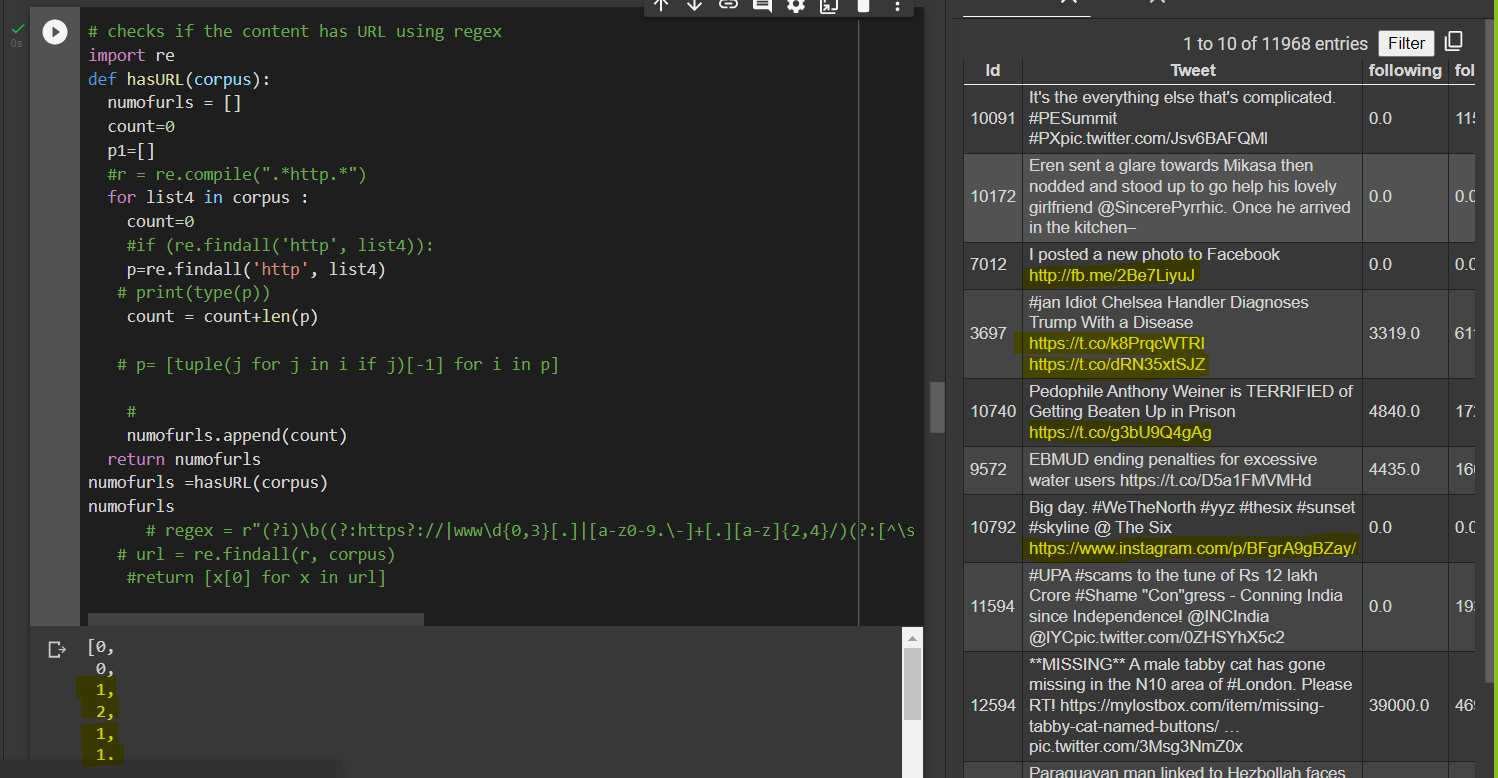
1. Number of Hashtags: This feature is very important , as it indicates whether or not the tweet is spam, because a lot of hashtags in the tweet , Spammers tweets mostly contain hashtags to attract the attention of legitimate users.



1. number of mentions : This is another important feature , it also gives a hint if this tweet is a spam or not. A lot of mentions in a tweet indicates that the tweet is a spam .



1. if the tweet has a URL or not : Spammers’ tweets mostly contain links to attract the attention of legitimate users and lead them to their malicious purposes

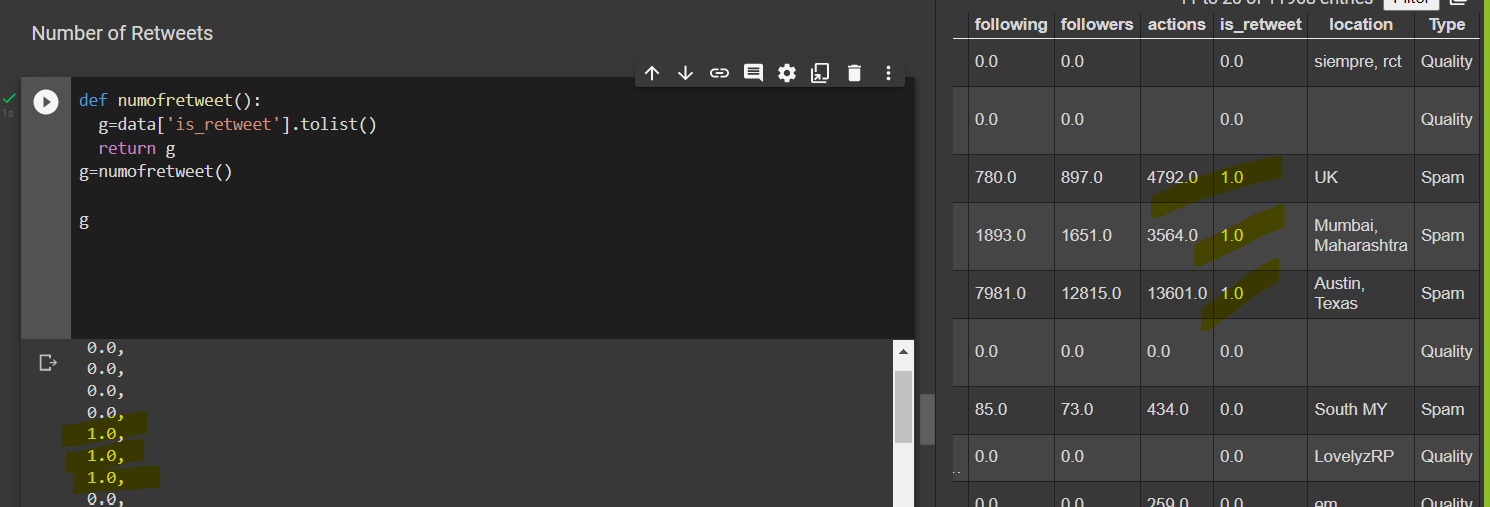


1. number of retweets :

Retweets- Retweets are the replies to any tweet using

@RT symbol and spammers use maximum @RT in

their tweets.



# the ratio of the number of following and number of followers:

This feature is very important , as it indicates whether or not the tweet is spam, because Compared to legitimate users, spammers are more likely to follow too many legitimate accounts to attract attention.

Since spammers are not followed by legitimate users, spammers have fewer followers than legitimate users. Also, their tweets are less likely to receive likes and retweets than those from legitimate users.

*The ratio (Followers/Following+Followers) was considered. since:*

*1. Numbers of followers-spammers have less number of*

*followers.*

*2. Numbers of followings-Spammers tend to follow a*

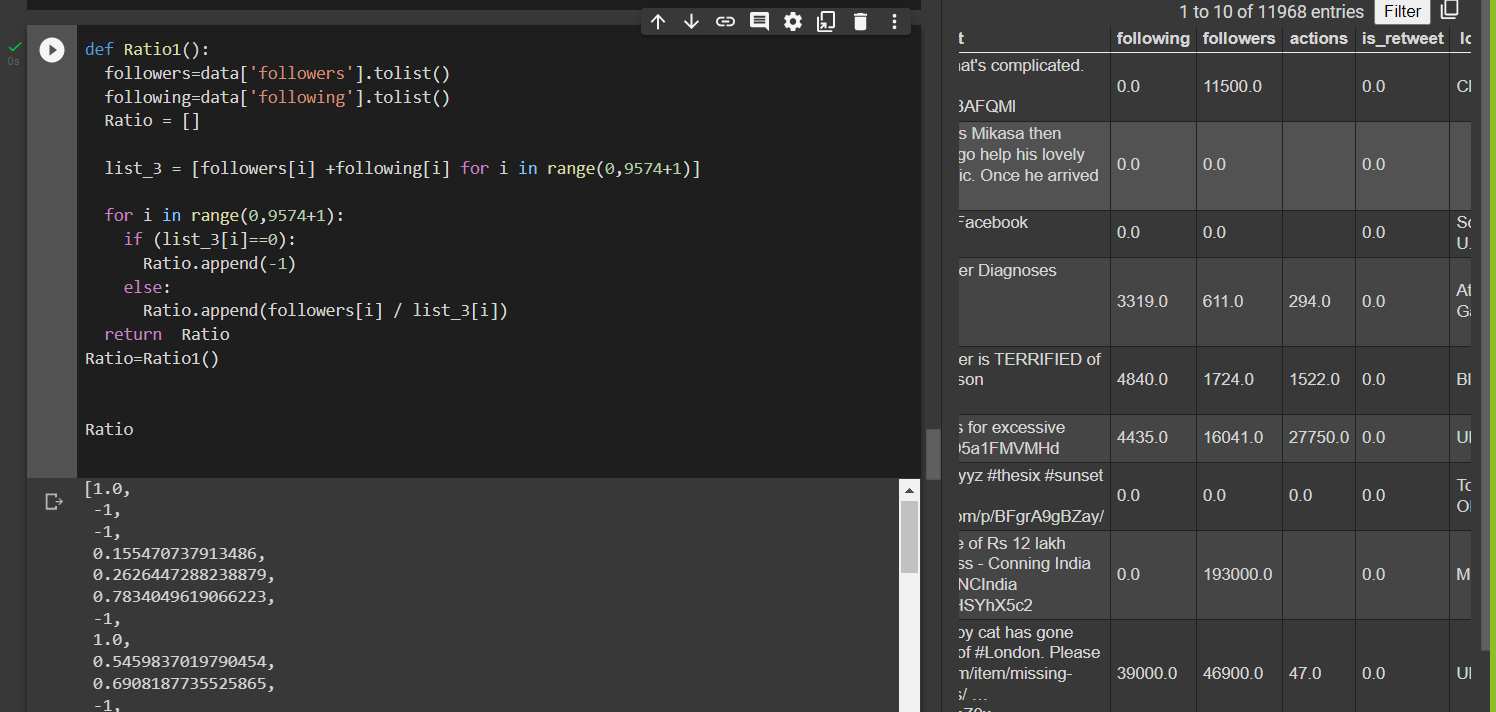
*large number of users.*

*3. Followers/Following Ratio- this ratio is less than 1*

*for spammers.*

*4. Reputation is defined as the ratio of followers to the*

*sum of followers*



The first account for example has 11500 followers while they don’t follow anyone.

so their following Ratio= 11500/(11500+0) =1.

4. Classification :

These steps were applied before applying classifiers:



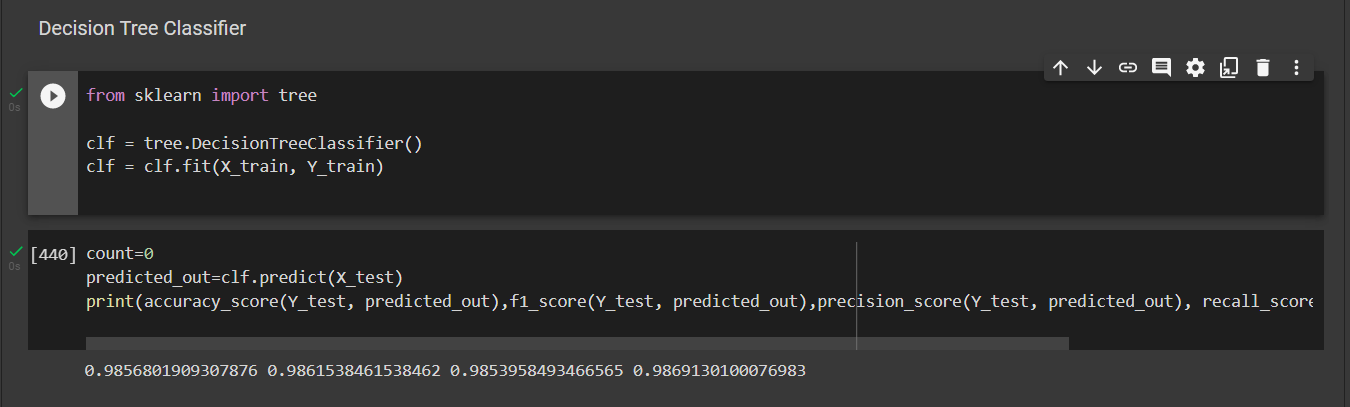
np.c Translates slice objects to concatenation along the second axis.

Since the Classifier’s output classes are either 1 or 0 so

Quality was given a value of 1 and Spam was given a value of 0. Then, The dataset was divided dataset into 70% training 30% testing finally, NaN data was handled.now the data is ready to get passed to classifiers.

Result and Testing Classifiers:

1. Decision tree Before Balancing



# After Balancing

# 

# 

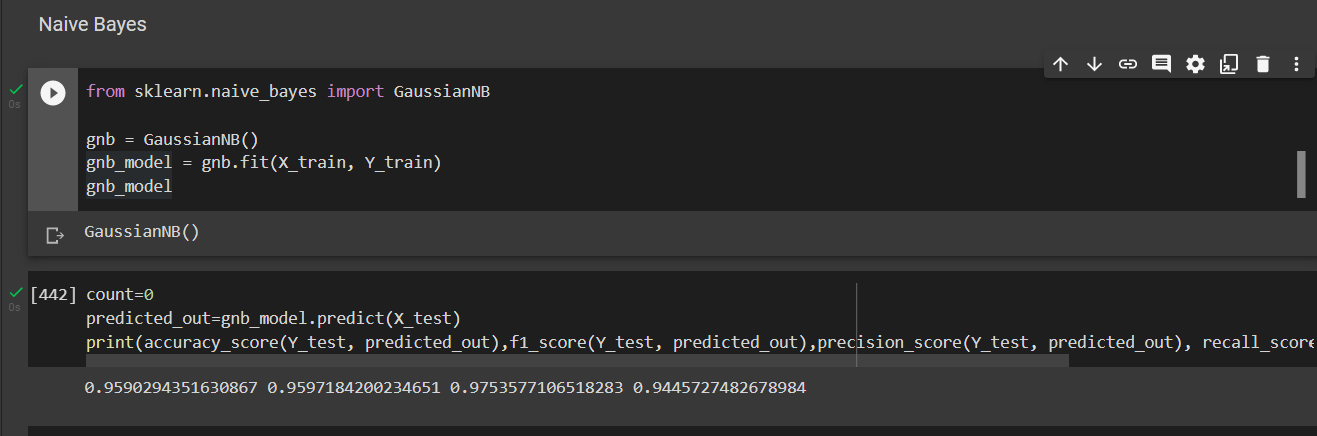
# 

# 

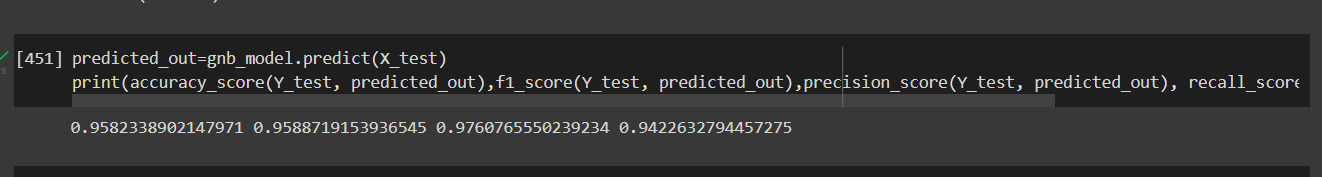
# 

# 

# Naive Bayes before balancing :



Naive Bayes after Balancing:



# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# Conclusion

# Reverences

[1] [Additive White Gaussian Noise (AWGN) - Wireless Pi](https://wirelesspi.com/additive-white-gaussian-noise-awgn/) accessed on 19-12-2021

[2] [What is AWGN - Noisecom](https://www.noisecom.com/resource-library/articles/artmid/1867/articleid/1233/what-is-awgn) accessed on 19-12-2021

[3] [What is signal-to-noise ratio and how is it measured? (techtarget.com)](https://www.techtarget.com/searchnetworking/definition/signal-to-noise-ratio) accessed on 19-12-2021

# Appendix