**Fake News Detection**

Team 3

**Team Members**

**Student 1**

**سليمان سنوسي سليمان مبروك**

**ID : 20191700891**

**Student 2**

**……..**

**ID : …….**

**Student 3**

**…….**

**ID : …….**

**Student 4**

**………**

**ID : ……..**

**Student 5**

**……..**

**ID : ……**

**Student 6**

**……..**

**ID : …….**

Introduction:

We consume news through several mediums throughout the day in our daily routine, but sometimes it becomes difficult to decide which one is fake and which one is authentic.

Do you trust all the news you consume from online media?

Every news that we consume is not real. If you listen to fake news it means you are collecting the wrong information from the world which can affect society because a person’s views or thoughts can change after consuming fake news which the user perceives to be true.

Since all the news we encounter in our day-to-day life is not authentic, how do we categorize if the news is fake or real?

In this article, we will focus on text-based news and try to build a model that will help us to identify if a piece of given news is fake or real.

**Description**

To get the accurately classified collection of news as real or fake we have to build a machine learning model.

To deals with the detection of fake or real news, we will develop the project in python with the help of ‘sklearn’, we will use ‘TfidfVectorizer’ in our news data which we will gather from online media.

After the first step is done, we will initialize the classifier, transform and fit the model. In the end, we will calculate the performance of the model using the appropriate performance matrix/matrices. Once will calculate the performance matrices we will be able to see how well our model performs.

**Data Analysis**

Here I will explain the dataset.

In this python project, we have used the CSV dataset. The dataset contains 7796 rows and 4 columns.

This dataset has four columns,

title: this represents the title of the news.

text: this column has the news itself.

label: this is a binary column representing if the news is fake or real .

**Data Preprocessing**

Techniques for Text Preprocessing:

- Expand Contractions. ...

- Lower Case. ...

- Remove punctuations. ...

- Remove words and digits containing digits. ...

- Remove Stopwords. ...

- Rephrase text. ...

- Stemming and Lemmatization. ...

- Remove Extra Spaces.

**TF-IDF Vectorizer**

TF (Term Frequency):  In the document, words are present so many times that is called term frequency. In this section, if you get the largest values it means that word is present so many times with respect to other words. when you get word is parts of speech word that means the document is a very nice match.

IDF (Inverse Document Frequency): in a single document, words are present so many times, but also available so many times in another document also which is not relevant. IDF is a proportion of how critical a term is in the whole corpus.

collection of word Documents will convert into the matrix which contains TF-IDF features using  TF-IDF Vectorizer.

**Count Vectorizer**

Machines cannot understand characters and words. So when dealing with text data we need to represent it in numbers to be understood by the machine. Countvectorizer is a method to convert text to numerical data.

Convert a collection of text documents to a matrix of token counts.

**Label Encoder**

Encode target labels with value between 0 and n\_classes-1.

This transformer should be used to encode target values, i.e. y, and not the input X.

In many Machine-learning or Data Science activities, the data set might contain text or categorical values (basically non-numerical values).

**Libraries**

The very basic data science libraries are sklearn, pandas, NumPy e.t.c and some specific libraries such as transformers.

**Classification Models**

Multinomial NB

Support Vector Machine Classifier

Logistic Regression Classifier

Stochastic Gradient Descent

**Models With Encoding:**

Logistic Regression Classifier

Multinomial NB

Data Set Summary:

**1-What is the data set used?**

Fake News Detection data set

**2- What is the summary of the dataset columns?**

Data Set Contains Columns Title , Text , and Label

Title Column that is not important in classification problem

Used Columns : Text and Label Columns.

Techniques Used for Text Preprocessing:

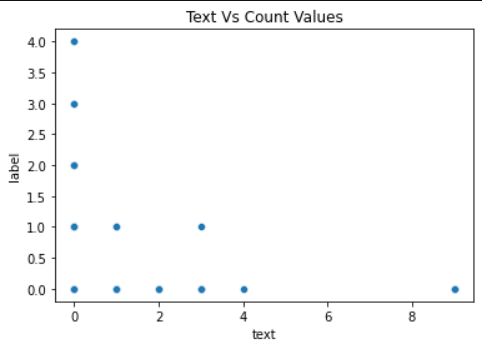
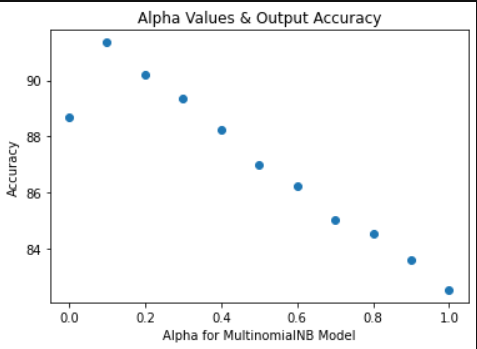
- Expand Contractions. - Lower Case.

- Remove punctuations. - Remove words and digits containing digits.

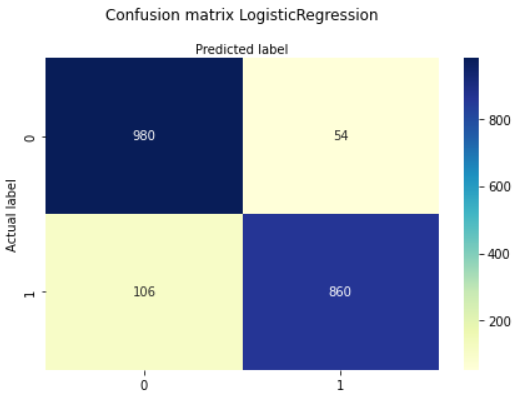
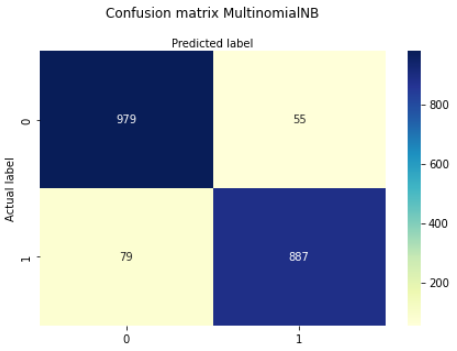
- Remove Stopwords. - Rephrase text.

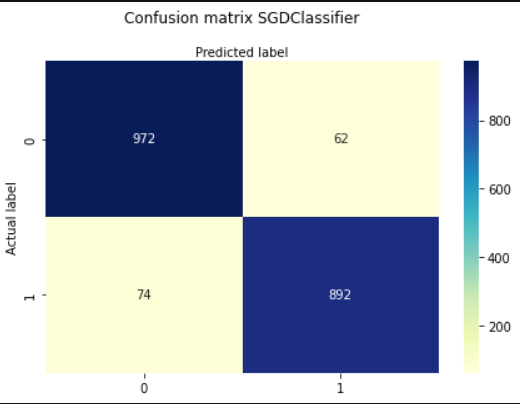
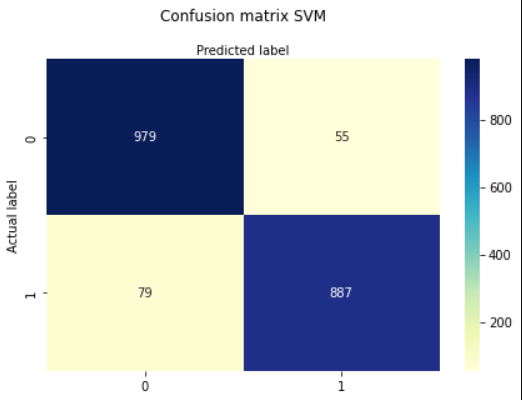
- Stemming and Lemmatization. - Remove Extra Spaces.

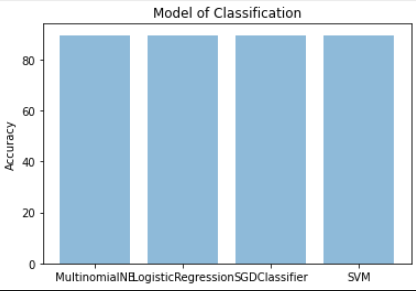
**3- Visualize the dataset statistics**

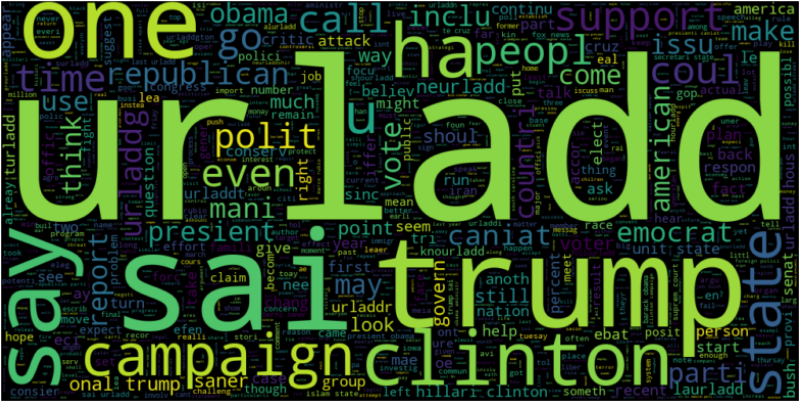
Results:

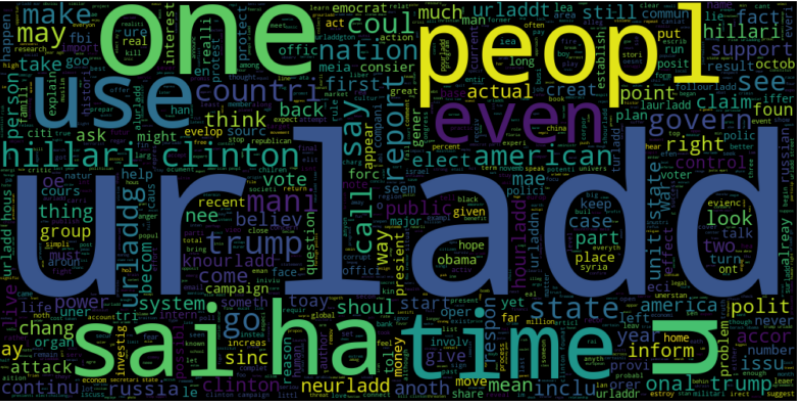
 

****

**Real News Words:**



**Feck News Words:**



**Conclusion**

The passive-aggressive classifier performed the best here and gave an accuracy of 92.59% without encoding by using TF-IDF Vectorizer , and 93.9% with encoding by using Count Vectorizer and TF-IDF Vectorizer as same result .

We can print a confusion matrix to gain insight into the number of false and true negatives and positives

Fake news detection techniques can be divided into those based on style and those based on content, or fact-checking. Too often it is assumed that bad style (bad spelling, bad punctuation, limited vocabulary, using terms of abuse, ungrammaticality, etc.) is a safe indicator of fake news.

More than ever, this is a case where the machine’s opinion must be backed up by clear and fully verifiable indications for the basis of its decision, in terms of the facts checked and the authority by which the truth of each fact was determined.

Collecting the data once isn’t going to cut it given how quickly information spreads in today’s connected world and the number of articles being churned out.