# FAKE NEWS DETECTION USING NLP

A MINOR PROJECT REPORT
[INTERNSHIP REPORT]

***Submitted by***

**P.S.A.BHASKAR REDDY**
**[RA1911003010072]**
**M.GNANESWAR REDDY**
**[RA1911003010113]**

*Under the guidance of*
**Dr.R.S.PONMAGAL**
(ASSOCIATE PROFESSOR,CTECH)

***in partial fulfillment for the award of the degree of***

**BACHELOR OF TECHNOLOGY**

in

**COMPUTER SCIENCE & ENGINEERING**

of

**FACULTY OF ENGINEERING AND TECHNOLOGY**

SRM
INSTITUTE OF SCIENCE & TECHNOLOGY
Deemed to be University u/s 3 of UGC Act, 1956

S.R.M.Nagar, Kattankulathur, Chengalpattu District

**NOVEMBER 2022**

# SRM INSTITUTE OF SCIENCE ANDTECHNOLOGY

(Under Section3 of UGC Act,1956)

## BONAFIDE CERTIFICATE

Certified that 18CSP107L minor project report [18CSP108L internship report] titled **“FAKE NEWS DETECTION USING NLP”** is the bonafide work of **“P.S.A.BHASKAR REDDY[RA1911003010072], M.GNANESWAR REDDY [RA1911003010113]”** who carried out the minor project work[internship] under my supervision. Certified further, that to the best of my knowledge the workreported herein does not form any other project report or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

| | |
|---|---|
| **SIGNATURE** | **SIGNATURE** |
| Dr.R.S.Ponmagal | Dr.M.Pusphalatha |
| Guide | Head Of The Department |
| Associate Professor | Professor |
| CTech | Dept Of Computing Technologies |

**Signature Of The Panel Head**

Dr.R.S.Ponmagal

Associate Professor

CTech

# ABSTRACT

In this modern era, Everyone relies on different online news sources. News quickly disseminated among the thousands of users of social media platforms such as Twitter, etc. However, there may be some misleading content for damaging the reputation of people or firms. The fake news propagators intentionally spread fake news tinfluence public opinions on particular issues. Therefore,detection of bogus news in advance is crucial order to stopits growth of this false information and protect innocent people from those who promote it. There are numerous ways to identify bogus news. In them natural language processing is one of the techniques which works effectively and efficiently. In natural language processing,Regular expression, tokenization, and lemmatization are examples of ext pre-processing techniques is used before vectorization.Vectorization is vectorizing the data iemploying to N-gram vectors or sequence vectors, the terms frequency-inverse document frequency or one-hot encoding, respectively.N-grams concept was mainly used to enhance the proposed model. In order to observe the accuracy of the model, classification algorithms of machine learning can be used. False Article detection aims to give the user the option of identifying news as false or real.

# TABLE OF CONTENTS

# LIST OF FIGURES

# ABBREVIATIONS

| | |
|---|---|
| **FN** | Fake News |
| **CSS** | Cascading Style Sheet |
| **CV** | Computer Vision |
| **DB** | Data Base |
| **UI** | User Interface |
| **TN** | True News |
| **ML** | Machine Learing |

# CHAPTER 1
# INTRODUCTION

## 1.1 Introductory Part

The number of people using the Internet has drastically expanded because to the development of information and communication technologies. Everyone relies on different online news sources. Due to the abundance of news immediately spread through millions of users of social media sites like Twitter, and others because they made the process of sharing information easier. It eases the process of accessing with data sharing in light of the technological revolution. Since On these online platforms, news can be easily produced.Platforms there may exist some fake news. This fake news is generated in order to damage or ruin someone's reputation or firm's reputation. The fake news propagators may do this to seek Ransome. It's not at all good for society if it continues. The spammers take this as an opportunity for making money and spam the news continuously. So, It is necessary to identify false information to save innocent individuals who spread bogus news and spam. Thus, in order to curb the spread of this, early hoax detection is essential.False information and protect innocent people from those who promote it. There are numerous ways to identify fake news, and natural language processing is one of them. Because it involves communication between people and machines, it is effective and efficient.We are using Natural Language Processing in our project since it isconcerned with building machines that can easily understand and respond to the text or voice data the same way humans do. With NLP, machines can even perform tasks on spoken or written text. The data pre-processing techniques that we are using in our project are Lemmatization, Tokenization, Stemming, Stop words and Vectorization. This is how fake news detection can be done using artificial intelligence to save innocent people from the fake news propagators and spammers. Fake news detection aimsto provide users with the options for categorizing news as false or accurate.

## Challenges:

There are various platforms for social media like Facebook, Twitter, etc., where people rely on for the news updates. In these platforms Any user can publish content or disseminate news. However,The posts on these platforms are not verified. So, some users intentionally spread fake news in these platforms in order to ruin reputation of a firm or a person.The Internet system leads to generate a lot of fake news content. The misleading content is generated by someone in order to damage the reputation of people or firms.

The fake news propagators may do this to seek Ransomer. It's not at all good for society if it continues. The spammers take this as an opportunity for making money and spam the news continuously.Spotting the fake news in the social media platforms is challenging because ofits dynamics. Since the spammers continuously spam the news to make some money, the news appears everywhere on the internet.Fake news is being widely disseminated, which might have very negative impacts on society and people. Persons opinions may also change on that particular firm or person.These are the challenges that world is facing due to these online platforms.It is creating negative impacts on the innocent people and the fake news propagators and spammers who are the actual spoilers are generating revenue by spamming the fake news.

## Solutions to those challenges:

There are so many that people are facing in this modern era due to the propagation of fake news. Any user can easily mislead the society by posting fake content in the social media. So, to stop spreading this fake news and to rescue innocent people from fake news propagators and spammers Early fake news detection is crucial to preventing further damage. There are numerous methods available to identify bogus news, among them natural language processing is one of the techniques which works effectively and efficiently. The required step that is being followed to implement an application are as follows:

(1) Establish a process for detecting misleading news online.
(2) The project's natural language analysis also produced a feature selection method.
(3) Collecting a dataset, we have used IFND dataset.
(4) We develop a fake news detection application.

Term frequency weighting, term frequency, word halting, word stemming, term frequency, inverse document frequency weighting, and word segmentation are all methods for analysing words.Components of the basic preprocessing phase for NLP. To adequately cover the news domain, we require as much data as we can. In order to create the model, data must be gathered. The feature data is then sent to machine learning, where it is divided into three sets with relative weights of 50%, 20%, and 30%: sets for teaching, validating, and testing.And divides news stories into two categories: true and false.When the data is provided in any language, we have employed translater to determine whether the news is authentic or fictitious. Naive Baye's Classifier machine learning models are the ones utilised in this research. In the end, this approach aids in determining if the news is legitimate or phoney. The goal of false information identification is to give the user the option of categorising the news as genuine or spurious.

## 1.2 Overview:

The goal of the fake information detection project is to give the user the option of classifying the news as true or phoney. Using a variety of methodologies, a model must be built to determine whether the news is legitimate or phoney. Since natural language processing is concerned with creating machines that can readily comprehend and react to text or speech data in the same way people do, we are employing it in our project. Machines can now carry out tasks on both written and spoken material thanks to natural language processing. For many different NLP tasks, the Python programming language offers a large variety of libraries and tools. The open-source set of tools, programmes, and libraries known as Natural Language Toolkit is used to create NLP programmes.The use of bots, ques solutions, text analytics, and language understanding are some examples of applications for natural language processing. In our project, we're using the data pre-processing techniques lemmatization, tokenization, stemming, stop words (methods for breaking sentences into tokens and trimming words), and vectorization. Vectorization can be applied to the pre-processed data after data pre-processing to turn the text into a numerical representation. When the data is provided in any language, we have employed translater to determine whether the news is authentic or phoney. This is how artificial intelligence can be used to detect bogus news in order to protect innocent people from spammers and those who spread false information. The goal of fake news identification is to give the user the option of categorising the news as true or untrue..

## 1.3 Problem Statement

Making a machine learning model that can identify if the provided news is legitimate or fraudulent is the goal of this project.

## 1.4 System Requirements

- Google Colab IDE
- Python
- OS : Windows, Linux (environment)
- Processor : Intel Dual Core(Minimum)
- RAM : 8GB(Minimum)
- Disk Space: 20GB(minimum)

# CHAPTER - 2

# LITERATURE SURVEY

## 2.1 Review of Literature Survey

| SL. no | Title | Year | Description | Limitations | Advantages |
|---|---|---|---|---|---|
| 1 | Ahmad, T.; Faisal, M.S.; Rizwan, A.; Alkanhel, R.; Khan, P.W.; Muthanna, A.<br>Efficient Fake News Detection Mechanism Using Enhanced Deep Learning Model. | 2022 | Machine Learning for Rumor Detection using ML algorithms are SVM, Random Forest, Logistic Regression, Gaussian Naïve Bayes and also used Neural network and recurrent Neural network. | Machine Learning for Rumor Detection using ML algorithms are SVM, Random Forest, Logistic Regression, Gaussian Naïve Bayes and also used Neural network and recurrent Neural network. | This research uses a new set of content-based and social-based features for rumor detection. |
| 2 | Meesad, P. Thai Fake News Detection Based on Information Retrieval, Natural Language Processing and Machine Learning. | 2021 | In this paper, author proposed natural language processing for fake news detection. | We have a research question on how to make deep learning understand the new s more as humans do. Besides, if news comprise sound, and video, the machine must analyze &respond correctly. | Fake news detection aims to provide the user with the ability to classify the news as fake or real. |
| 3. | Jamal Abdul Nasir, Osama Subhani Khan, Iraklis Varlamis, Fake news detection: A hybrid CNN-RNN based deep learning approach. | 2021 | The TI-CNN (Text and Image information based Convolutional Neural Network) model has been proposed. | If news comprise sound, and video, the machine must analyze and respond correctly. | It is difficult to evaluate the fake news content in the online resources. By using an algorithm for detecting the fake news, the innocent |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | people can besaved. |
| 4 | "Fake Media Detection Based on Natural Language Processing by Z. Shahbazi and Y. -C. Byun. | 2021 | Applied the reinforcement learning technique, a learning based algorithm, to make a strong decision-making architecture andcombine it with block chain framework, smart contract, and customized consensus algorithm. | There are some existing softwaretools for micro blogging sites which are mainlybuild to combat fake news problem. | Fakenews sharing is one of the popular research problems in recent technology based on lack of security andtrust in terms of the truth of shared news insocial media. |
| 5 | Fake News Detection Using Machine Learning Approaches, BN Alwasel1, H Sirafi1 and M Rashid, ZKhanam. | 2021 | This paper makes analysis of the research related to fake news detection and explores the traditional machine learning models to create a model of a product with supervisedmachine learning algorithm. | There are some existing softwaretools for micro blogging sites which are mainlybuild to combat fake news problem. | Theauthor ex plored the traditional machine learning models to choose the best, in order to create a bestmodel with supervised machine learning algorithm. |
| 6 | Kaliyar, R.K., Goswami, A. & Narang, P. DeepFakE: improvingfake news detection using tensor decomposition-based deep neural network. | 2021 | In this paper the author proposed natural language processing forsocial media fake news detection using echo-chamber technique. Theproposed method has been tested on a real- world dataset | If news contents arein the form of text, sound, video the predicting whether the news is fake or real is difficult. | The use of ourproposed DNN further improves the performance as compared to both traditional machine learning as well as deep learning algorithms. |

| | | | | | |
|---|---|---|---|---|---|
| 7. | Uma Sharma, Sidarth Saran, Shankar M. Patil,2021, Fake News Detection using Machine Learning Algorithms. | 2020 | In this paper, four different machine learning algoritms suchas Naïve Bayes, Random forest and Logistic regression algorithms are used forclassification. | If news contents arein the form of text, sound, video the predicting whether the news is fake or real is difficult | Detecting the fake news at the early stageis very helpfulto save innocent people from fake news propagator |
| 8 | Zervopoulos A., Alvanou A.G., Kermanidis K. (2020) Hong Kong Protests: Using Natural LanguageProcessing for Fake News Detection on Twitter. In: Maglogiannis I., Iliadis L., Pimenidis E. (eds) Artificial Intelligence Applications and Innovations. springer | 2020 | In this paper the author proposed natural language processing forsocial media fake news detection on twitter. In this paper ML algorithms are used fot feature preprocessing and selection methods are considered. | The data content inthe dataset is very less. So, the accuracy may be less since the dataset does not contain large volumes of labelleddata | Our proposed model outperforms with the existing fake news detection methods by applying deep learning on combined news content and social context-based feauters. |
| 9 | J. C. S. Reis, A. Correia,<br>F. Murai, A. Veloso and<br>F. Benevenuto, "Supervised Learningfor Fake News Detection,"<br>in IEEE Intelligent Systems. | 2019 | In this paper, author discussed how supervisory algorithmscan be used for detecting the fake news. | The data content inthe dataset is very less. So, the accuracy may be less since the dataset does not contain large volumes of labelleddata | This paper proposes the main features for fake news detection. Thispaper present a new set of features and measure the performance of current approaches and features for automatic detection of fake news |

| | | | | | |
|---|---|---|---|---|---|
| 10 | “Fake News Detection using Machine Learningand Natural Language Processing” Kushal Agarwalla, Shubham Nandan, Varun Anil Nair, D. Deva Hema, IJRTE. | 2019 | In this paper NPL(natural preprocessing languages) algorithmsNLTK is used NLTKin python was used to tokenize the body andheadline. | There are some existing softwaretools for micro blogging sites which are mainlybuild to combat fake news problem. | The maximum accuracy of 83% was attained by using Naïve Bayes classifier with lidstone smoothing. Whereas in the model which consisted of only Naïve Bayes attained an accuracyof 74%. |

# CHAPTER 3

## SYSTEM ARCHITECTURE AND DESIGN

## 3.1 UML DIAGRAMS

### 3.1.1 Data Flow Diagram

A data flow diagram shows the way information flows through a process or system. It includes data inputs and outputs, data stores, and the various subprocesses the data moves through. DFDs are built using standardized symbols and notation to describe various entities and their relationships.

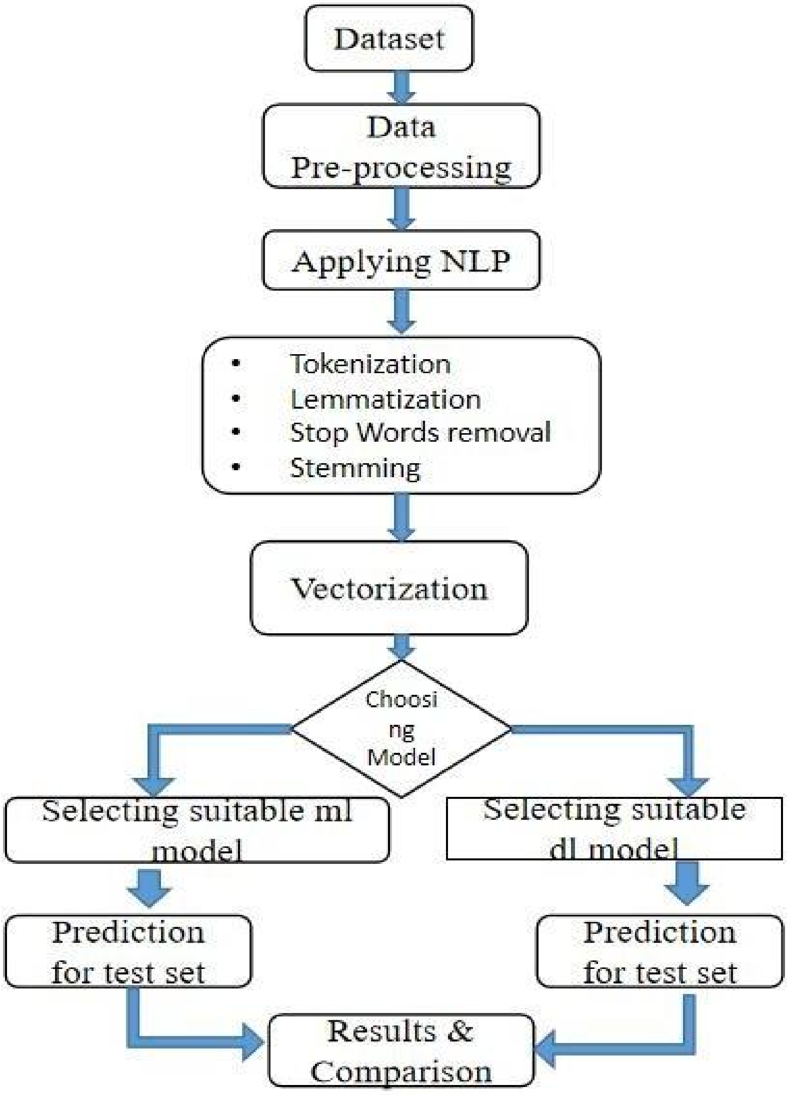


**FIG - 3.1.1(Data Flow Diagram)**

## 3.1.2 Architecture Diagram

An architecture diagram is a visual representation of all the elements that make up part, or all, of a system. Below all, it helps the engineers, designers, stakeholders — and anyone else involved in the project understand a system or app's layout.

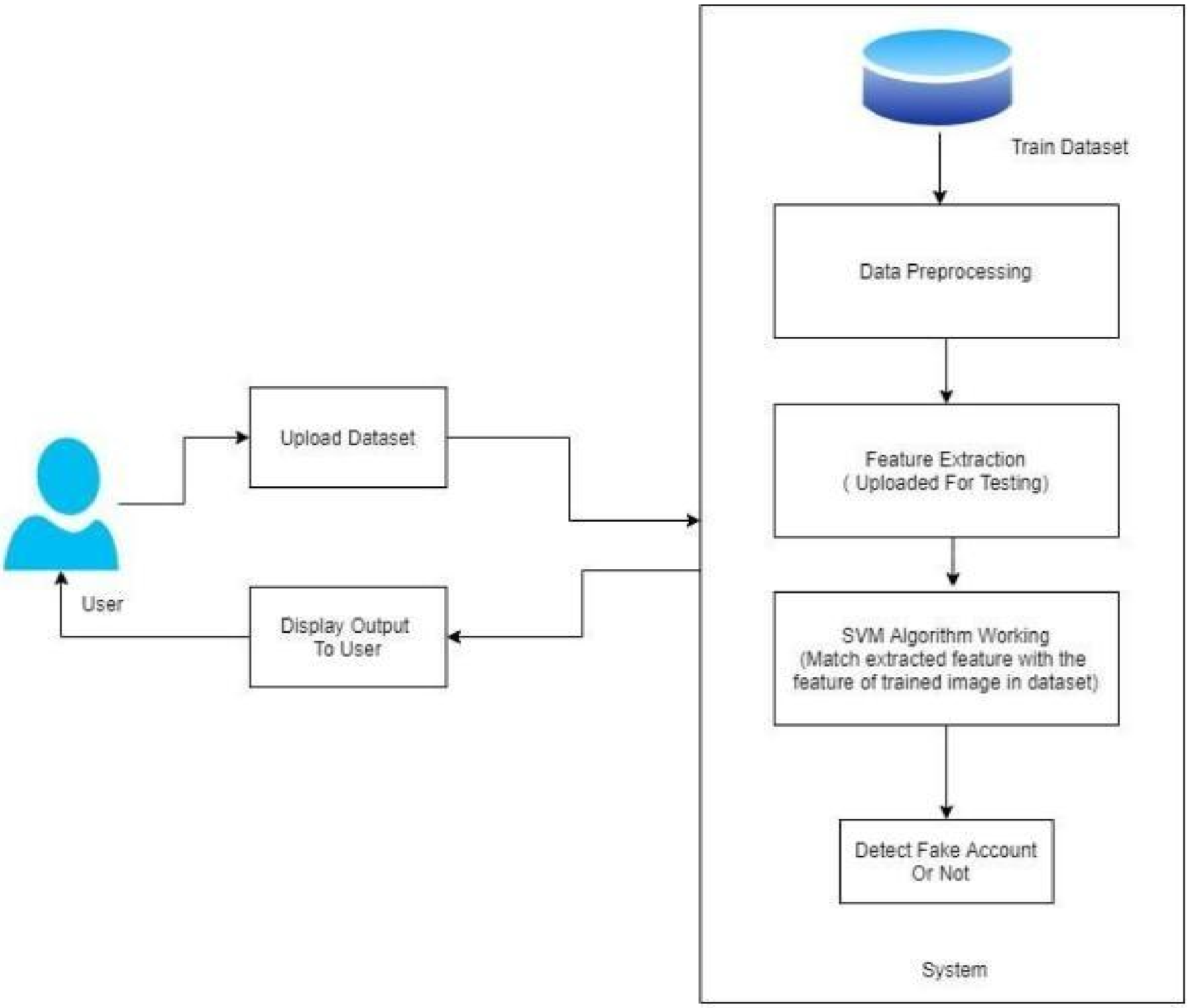


**FIG - 3.1.2(Architecture Diagram)**

# CHAPTER 4

# METHODOLOGY

## 4.1 Proposed Model

The proposed model for Fake News Detection is Natural Language Processing. Natural language processing is a form of Artificial Intelligence which is concerned with building machines that can easily understand and respond to the text or voice data the same way humans do. With NLP, machines can even perform task on spoken or written text.The Python programming provides a wide range of libraries and tools for various NLP tasks. Natural Language Toolkit is the open-source collection oflibraries, programs, and resources for building NLP programs.The applications of natural language processing are speech recognition, sentiment analysis, question/answer systems, chatbots, automatic text summarization, market intelligence, automatic text classification, and automatic grammar checking. etc.

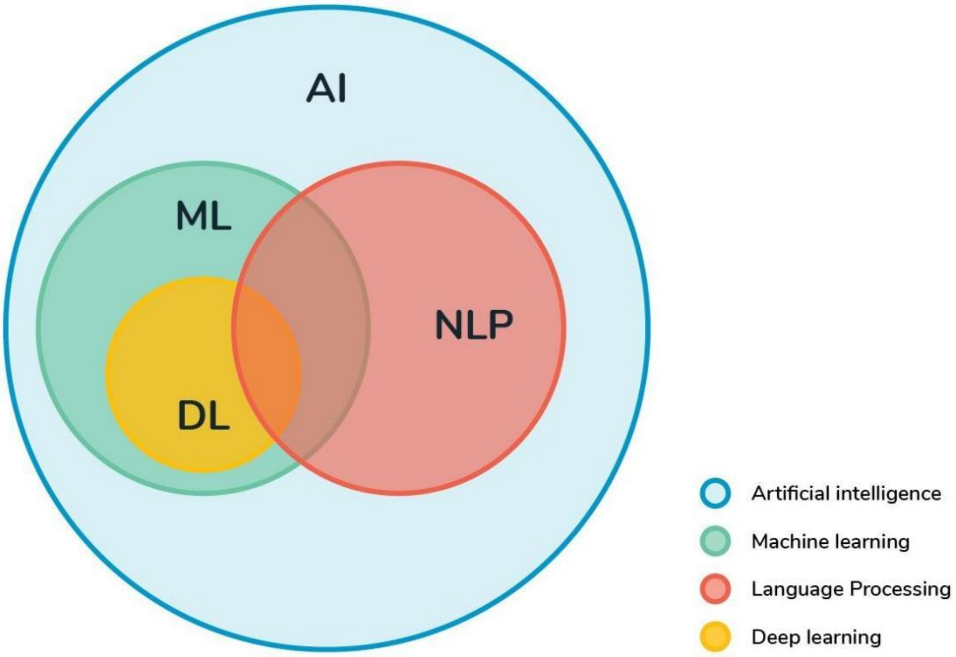


FIG - 4.1(Proposed Model Diagram)

### 4.1.1 Data Pre-Processing

Data pre-processing is the first and crucial step while building the machine learning models since it concerned with preparing the raw data and making it suitable for machine learning model.The Natural Language Toolkit includes libraries for NLP tasks such as stemming, lemmatization, tokenization(methods that are usedfor breaking sentences into tokens and trimming words).After data pre-processing vectorization can be done to the pre-processed data for converting the text into numerical representation.The three methods that are involved in data pre-processing are:

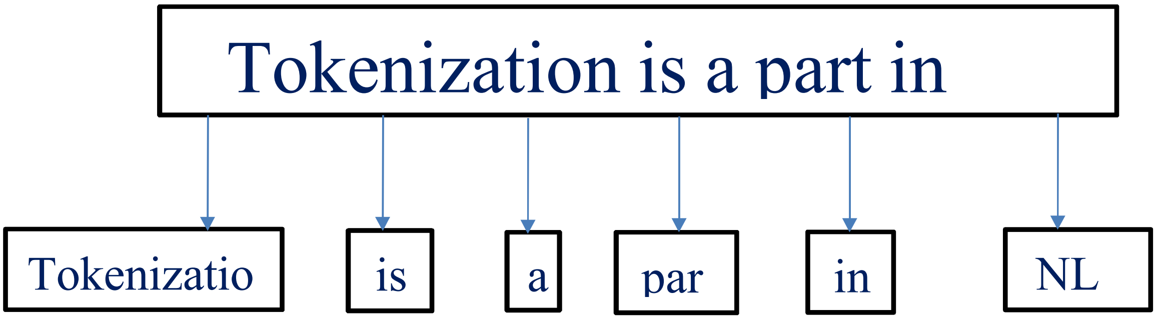
1. Tokenisation
2. Lemmatization
3. Stemming
4. Stop words

### Tokenization:

Tokenization is the process of breaking down the natural language text data into chunks of information I.e., smaller units called as tokens. These tokens helps in understanding and developing the model for the Natural Language Processing.It is the basic and crucial step in natural language processing. It further helps in interpreting the meaning of the given text by analyzing the sequence of the words in the given data.Tokenization is classified into three types- word, character and n-gram characters tokenization. We are using in our fake news detection project.Each sentence from the dataset will undergoes the primary preprocessing i.e., tokenization using the in built split function as shown below.

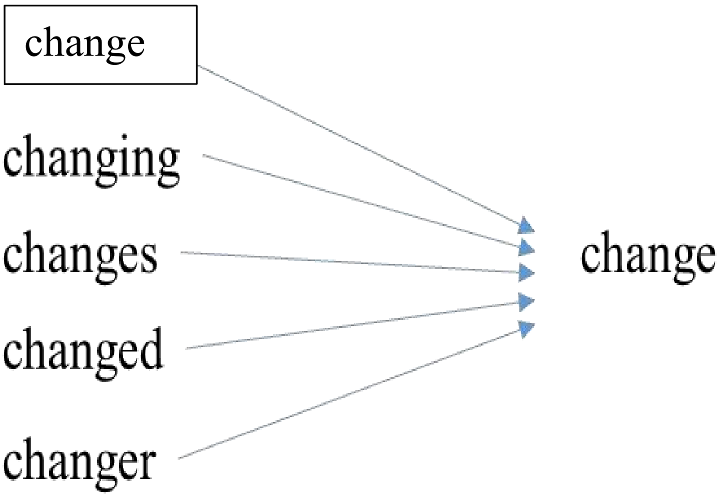
```
sentence='tokenization is part of NLP'
```

Tokenizer_list =sentence.split()



## Lemmatization:

Lemmatization is the technique which is used to reduce the tokens into normalized form i.e., root dictionary form.This technique takes into consideration of the morphological analysis of the words to convert the words into normalized form.Lemmatization mainly focus on the context in which the word is being used.Lemmatization techniques are used by the search engines and chatbots toanalyze the meaning behind the words.

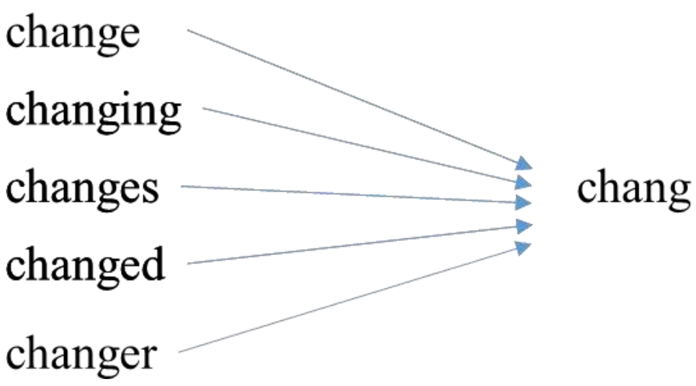


## Stemming:

Stemming is the technique of reducing the word to its word stem i.e., word base. Stemming is basically removing the suffix from a word and reduce it to its root word. This technique uses the stem of the word.Stemming techniques are used by the search engines and chatbots to analyze the meaning behind the words to produce the better results.The stemming will be applied to the dataset by creating object to the porter stemmer class.

Word=PorterStemmer()

Word.stem('changing')



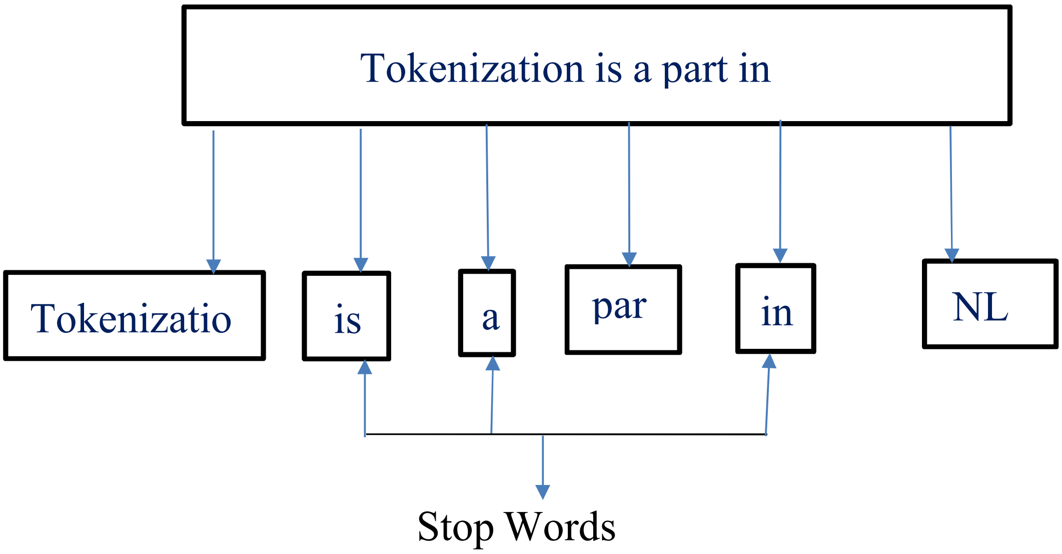
## Stop words:

Stop words are used to eliminate the unimportant words, allowing the applications to focus on the important words instead. The stop words are not necessary in our project because they has no scope in training or testing the data.This method can be done by maintaining a list of stop words and preventing all the stop words from analyzed.The necessary module for importing the stopwords is nltk.corpus. From nltk.corpus stopwords are imported.

Some of the known stop words are given below:

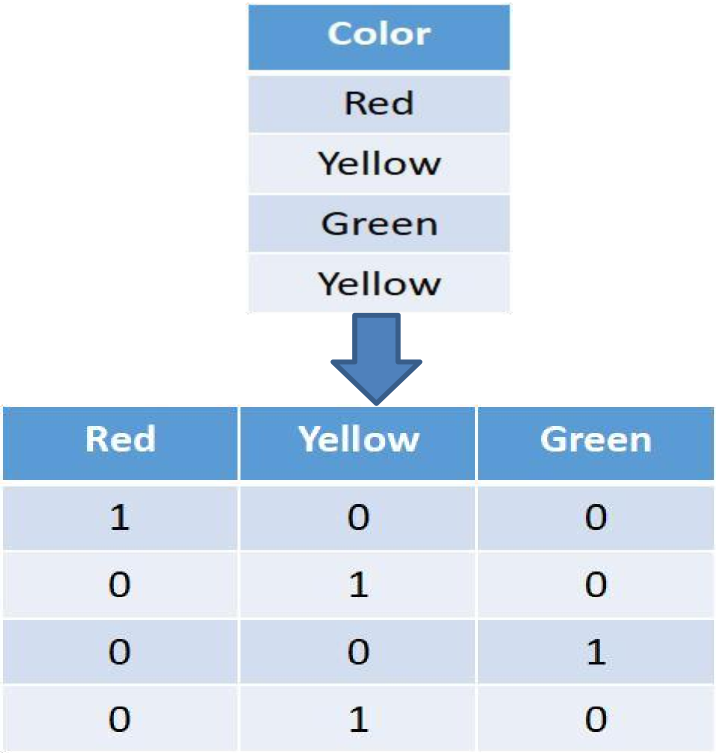
ourselves, hers, between, yourself, but, again, there, about, once, during, out, very, having, with, they, own, an, be, some, for, do.

most, itself, other, off, is, s, am, or, who, as, from, him, each, the, themselves, until, below, are, we, ‘these, your, his, through, don, nor, me, were, her, more,himself, this, down, should, our, their, while, above, both, up, to, ours, had, she, all, no, when, at, any, before, them, same, and, been, have, in, will, on, does, yourselves, then, that, because, what.



### 4.1.2 Vectorization

After data pre-processing vectorization can be done on the pre- processed data. Vectorization is the process of converting the textdata to numerical variables.We will be creating vectors that have a dimensionality equal to the size of our vocabulary, and if the text data features that vocabword, we will put a one in that dimension. Every time we encounter that word again, we will increase the count, leaving 0severywhere we did not find the word even once.The result of this will be very large vectors, if we use them on realtext data, we will get very accurate counts of the word content of our text data.



| Color |
|---|
| Red |
| Yellow |
| Green |
| Yellow |

| Red | Yellow | Green |
|---|---|---|
| 1 | 0 | 0 |
| 0 | 1 | 0 |
| 0 | 0 | 1 |
| 0 | 1 | 0 |

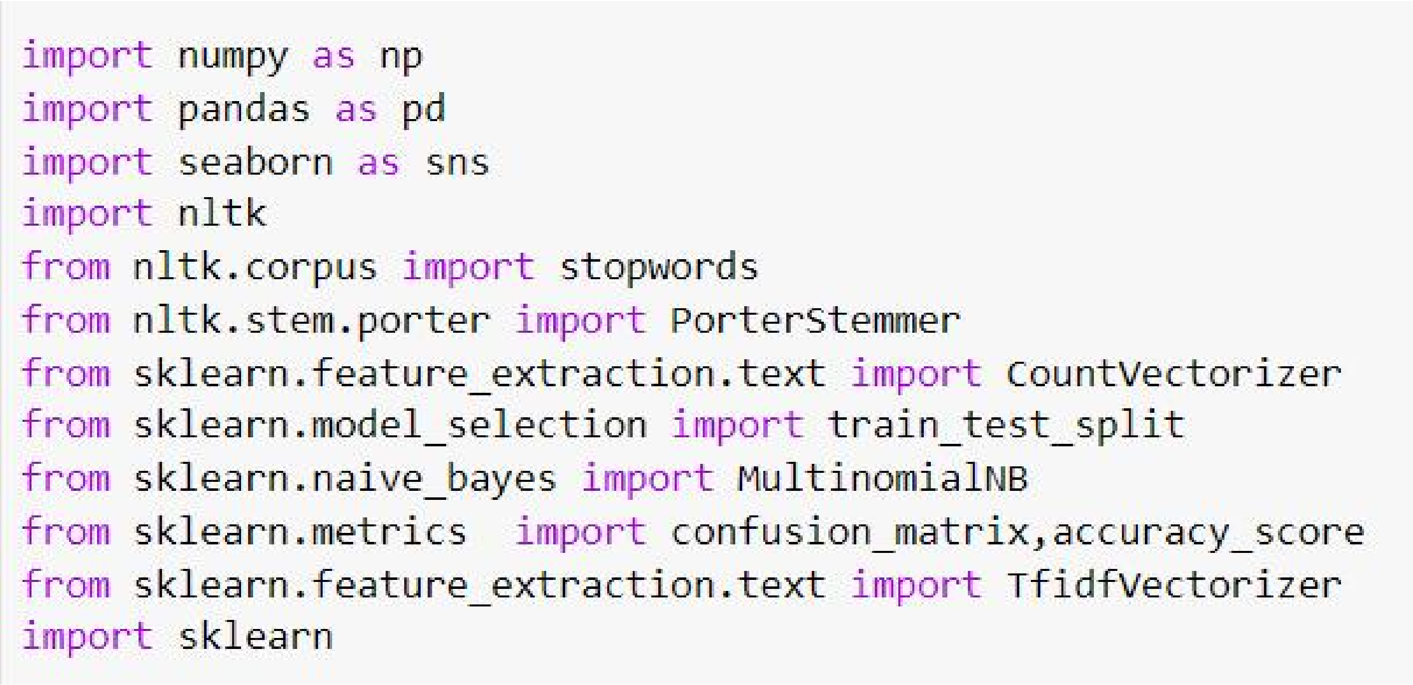
**FIG - 4.1.2(Vectorization Diagram)**

## 4.1 Naive Baye’s Classifier

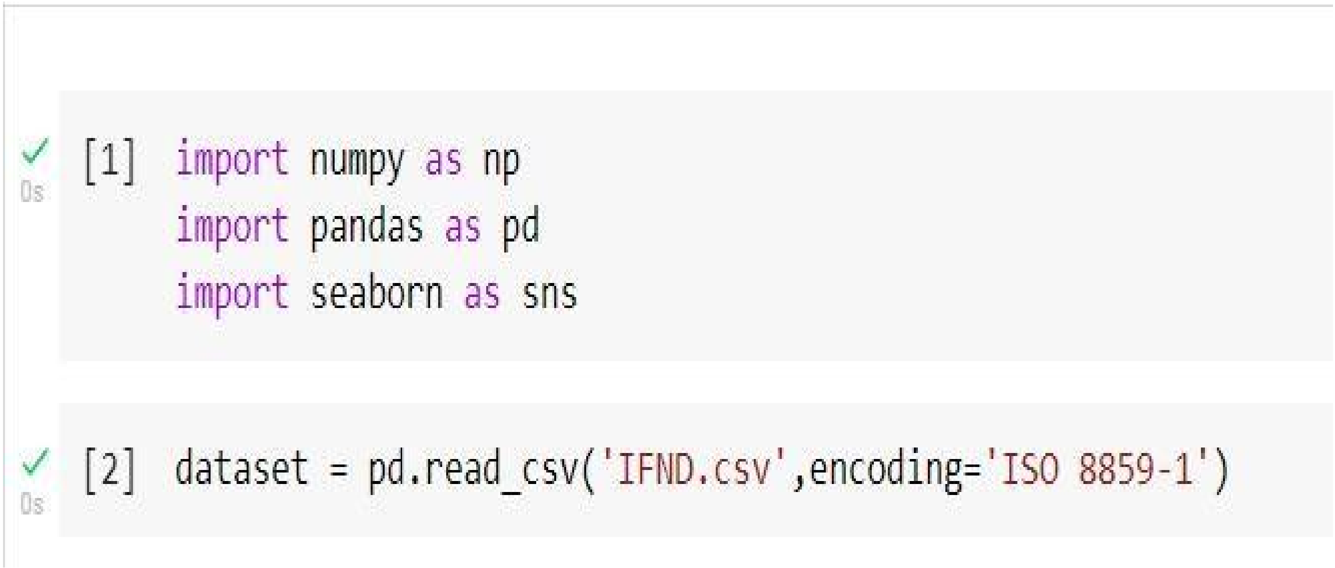
Naive bayes classifiers are a collection of classification algorithms based onBayes Theorem. These classifiers mainly used for text classification and text analysis machine learning problems.The principle of naïve bayes algorithm is every pair of features being classified is independent of each other. Naive bayes algorithmis family of algorithms where all of them share a common principle.There are two main assumptions of naive bayes algorithm i.e., it assumes that each variable or a feature of the same class makes an independent and equal contribution to the outcome.However, the assumptions made by Naive Bayes are not generally correct inthe real-world situations. Due to its independence assumption, it is called asnaive i.e., because it assumes something that might not be true.The naive bayes classifier converts the collection of text documents into a matrix of token counts during implementation.

# CHAPTER - 5
# 5 .CODING AND TESTING

## 5.1 Import Libraries



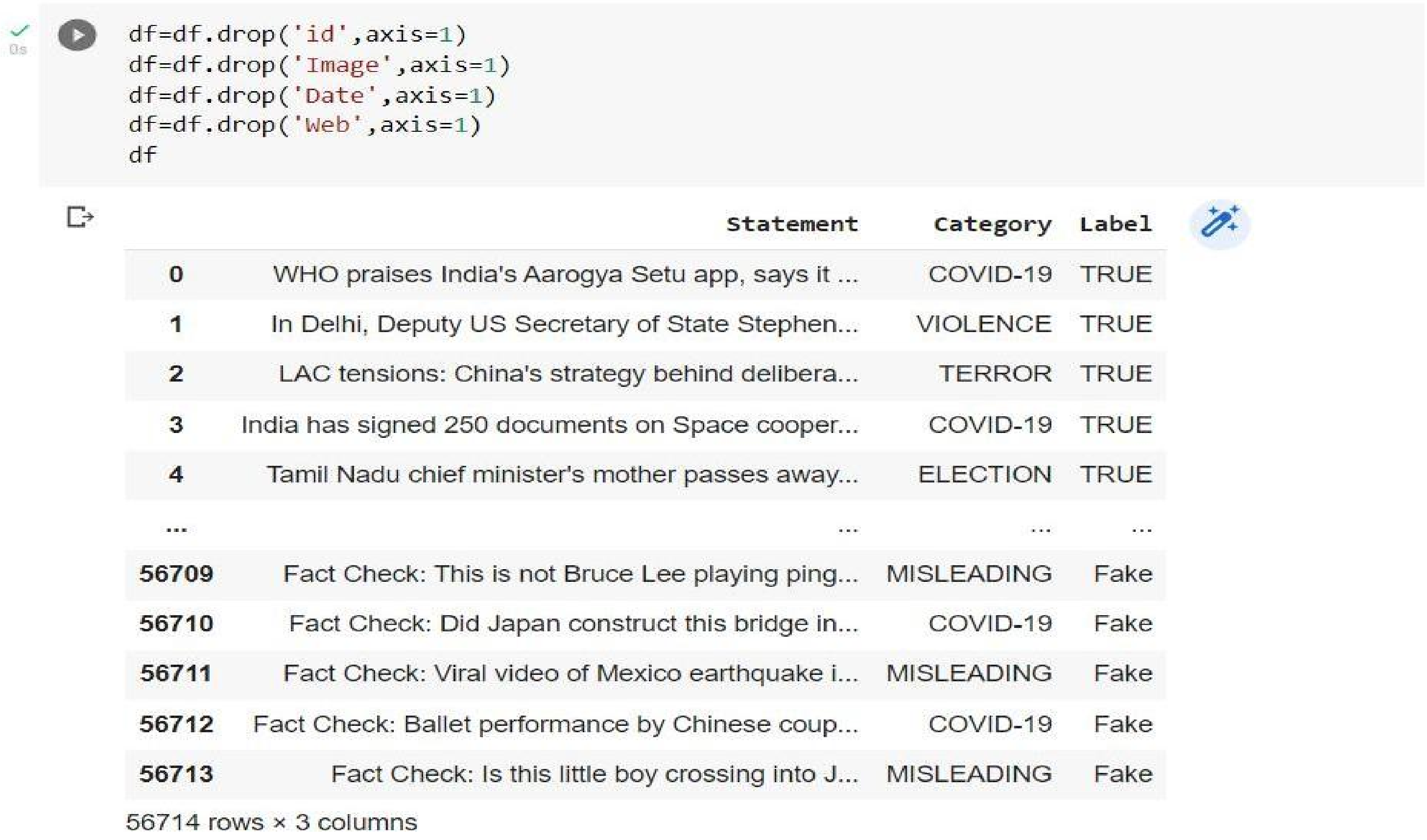
```
import numpy as np
import pandas as pd
import seaborn as sns
import nltk
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.model_selection import train_test_split
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics  import confusion_matrix,accuracy_score
from sklearn.feature_extraction.text import TfidfVectorizer
import sklearn
```



```
import numpy as np
import pandas as pd
import seaborn as sns
```

```
dataset = pd.read_csv('IFND.csv',encoding='ISO 8859-1')
```

## 5.2 Checking Dataset Labels

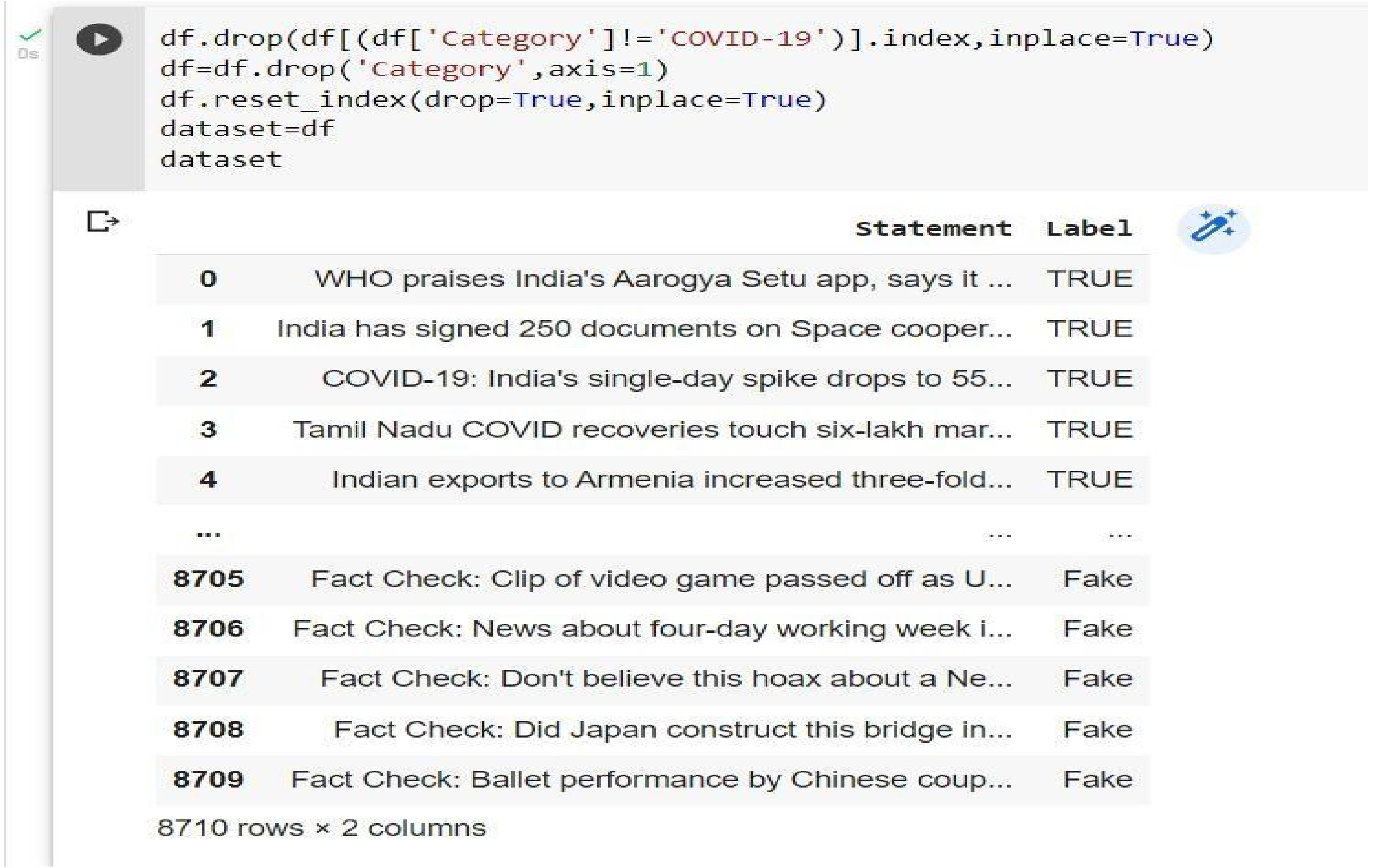


```
df=df.drop('id',axis=1)
df=df.drop('Image',axis=1)
df=df.drop('Date',axis=1)
df=df.drop('Web',axis=1)
df
```

|  | Statement | Category | Label |
|---|---|---|---|
| 0 | WHO praises India's Aarogya Setu app, says it ... | COVID-19 | TRUE |
| 1 | In Delhi, Deputy US Secretary of State Stephen... | VIOLENCE | TRUE |
| 2 | LAC tensions: China's strategy behind delibera... | TERROR | TRUE |
| 3 | India has signed 250 documents on Space cooper... | COVID-19 | TRUE |
| 4 | Tamil Nadu chief minister's mother passes away... | ELECTION | TRUE |
| ... | ... | ... | ... |
| 56709 | Fact Check: This is not Bruce Lee playing ping... | MISLEADING | Fake |
| 56710 | Fact Check: Did Japan construct this bridge in... | COVID-19 | Fake |
| 56711 | Fact Check: Viral video of Mexico earthquake i... | MISLEADING | Fake |
| 56712 | Fact Check: Ballet performance by Chinese coup... | COVID-19 | Fake |
| 56713 | Fact Check: Is this little boy crossing into J... | MISLEADING | Fake |

56714 rows × 3 columns

## 5.3 Checking Empty Values

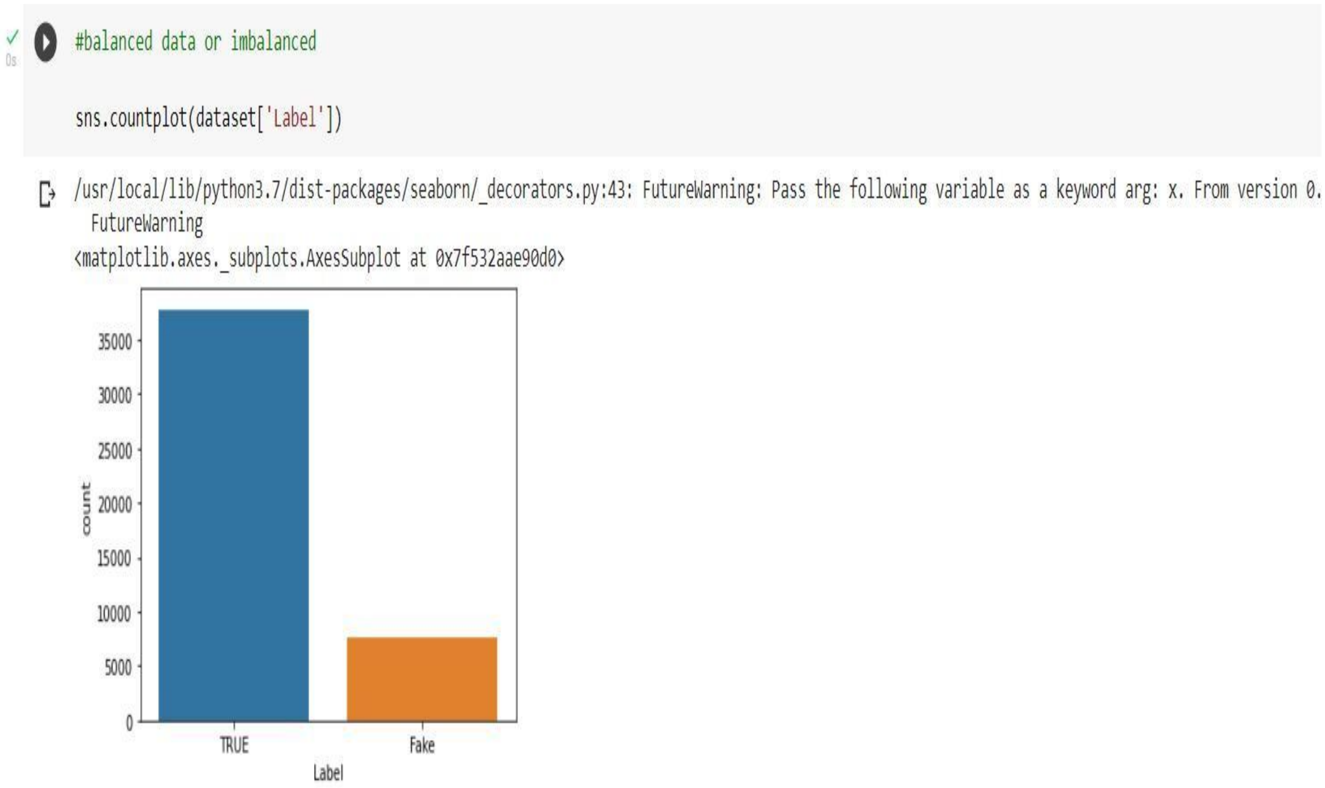


```
df.drop(df[(df['Category']!='COVID-19')].index,inplace=True)
df=df.drop('Category',axis=1)
df.reset_index(drop=True,inplace=True)
dataset=df
dataset
```

|  | Statement | Label |
|---|---|---|
| 0 | WHO praises India's Aarogya Setu app, says it ... | TRUE |
| 1 | India has signed 250 documents on Space cooper... | TRUE |
| 2 | COVID-19: India's single-day spike drops to 55... | TRUE |
| 3 | Tamil Nadu COVID recoveries touch six-lakh mar... | TRUE |
| 4 | Indian exports to Armenia increased three-fold... | TRUE |
| ... | ... | ... |
| 8705 | Fact Check: Clip of video game passed off as U... | Fake |
| 8706 | Fact Check: News about four-day working week i... | Fake |
| 8707 | Fact Check: Don't believe this hoax about a Ne... | Fake |
| 8708 | Fact Check: Did Japan construct this bridge in... | Fake |
| 8709 | Fact Check: Ballet performance by Chinese coup... | Fake |

8710 rows × 2 columns

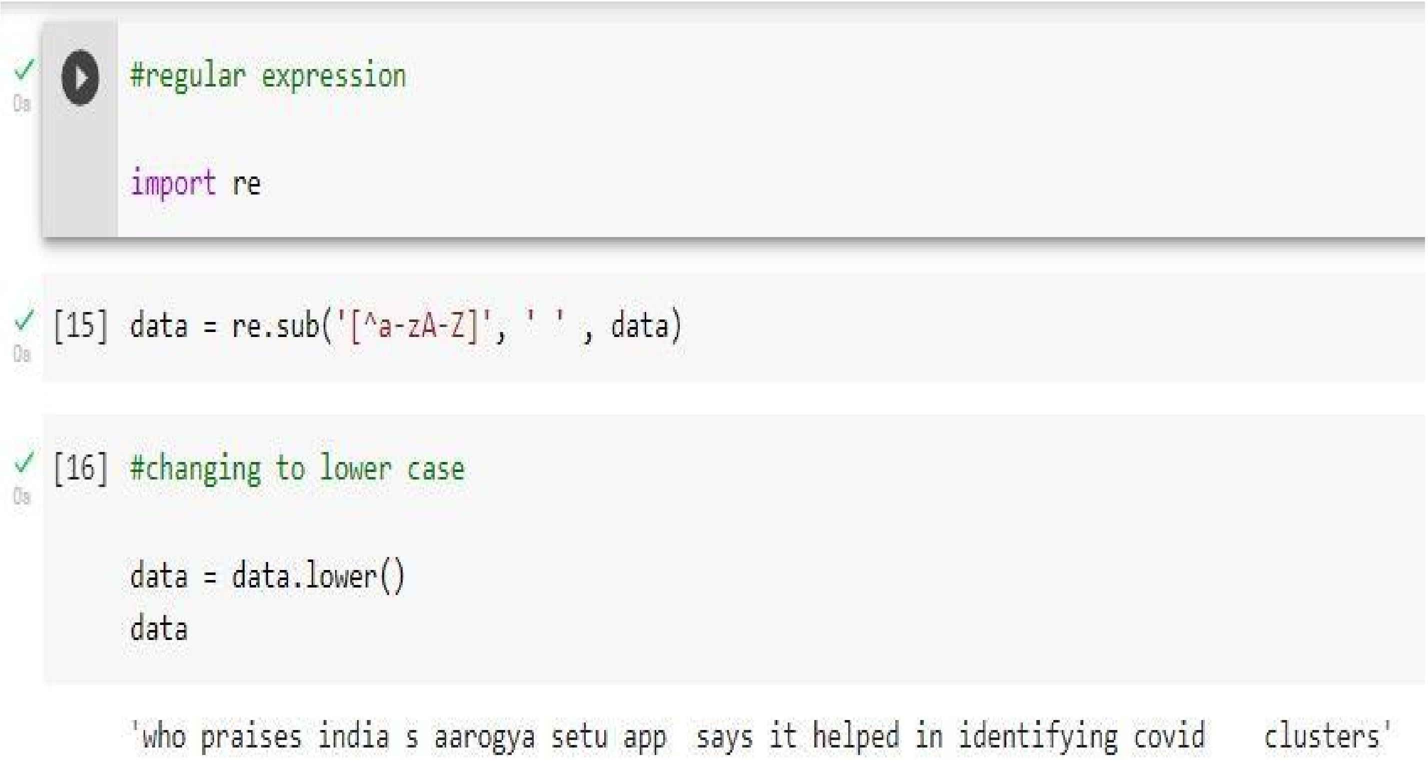
## 5.4 Checking The Data



```
#balanced data or imbalanced

sns.countplot(dataset['Label'])
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning: Pass the following variable as a keyword arg: x. From version 0.
  FutureWarning
<matplotlib.axes._subplots.AxesSubplot at 0x7f532aae90d0>
```

## 5.5 Importing Regular Expression



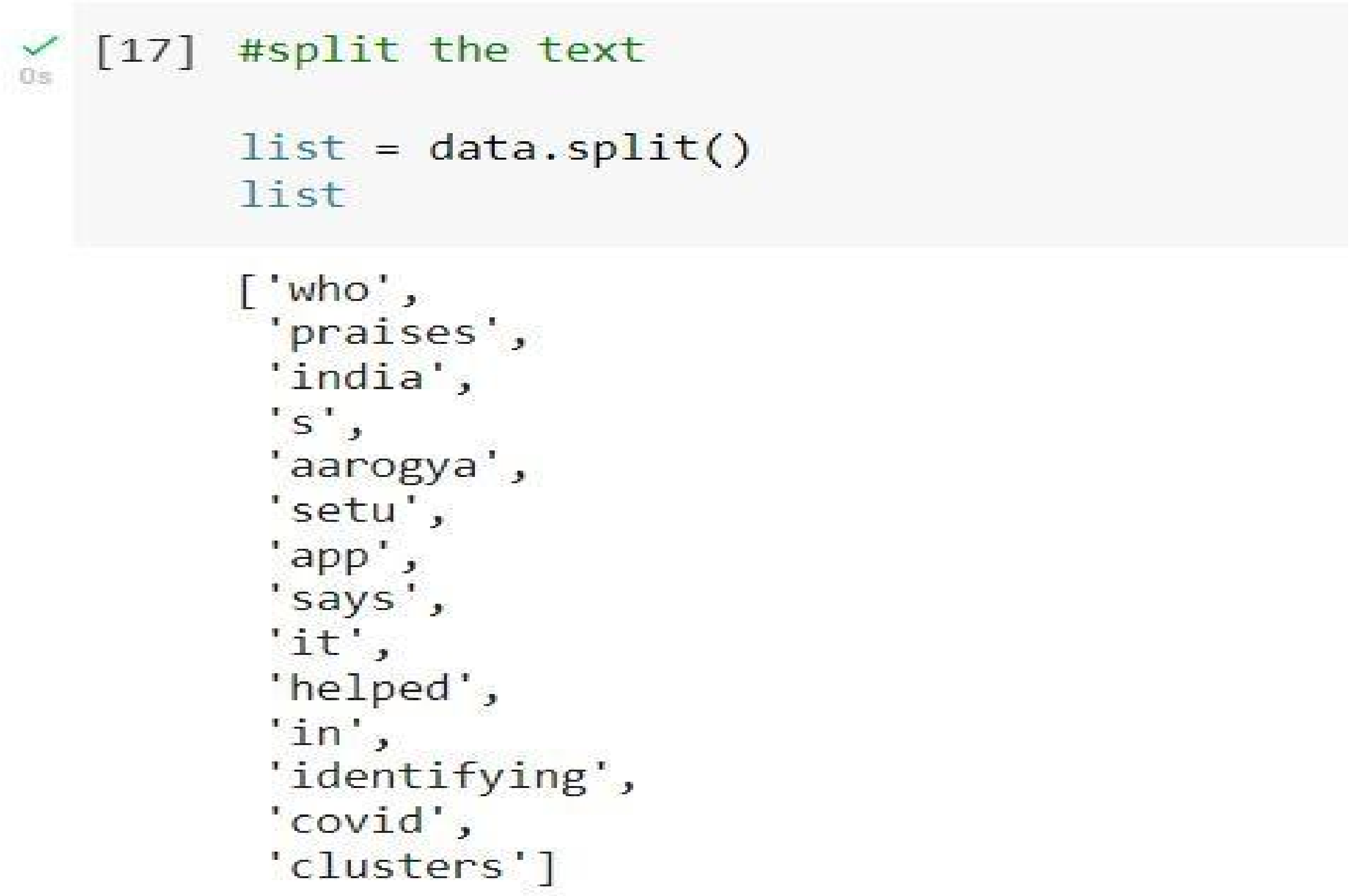
```
#regular expression

import re
```

```
[15] data = re.sub('[^a-zA-Z]', ' ' , data)
```

```
[16] #changing to lower case

     data = data.lower()
     data
```

```
'who praises india s aarogya setu app  says it helped in identifying covid    clusters'
```

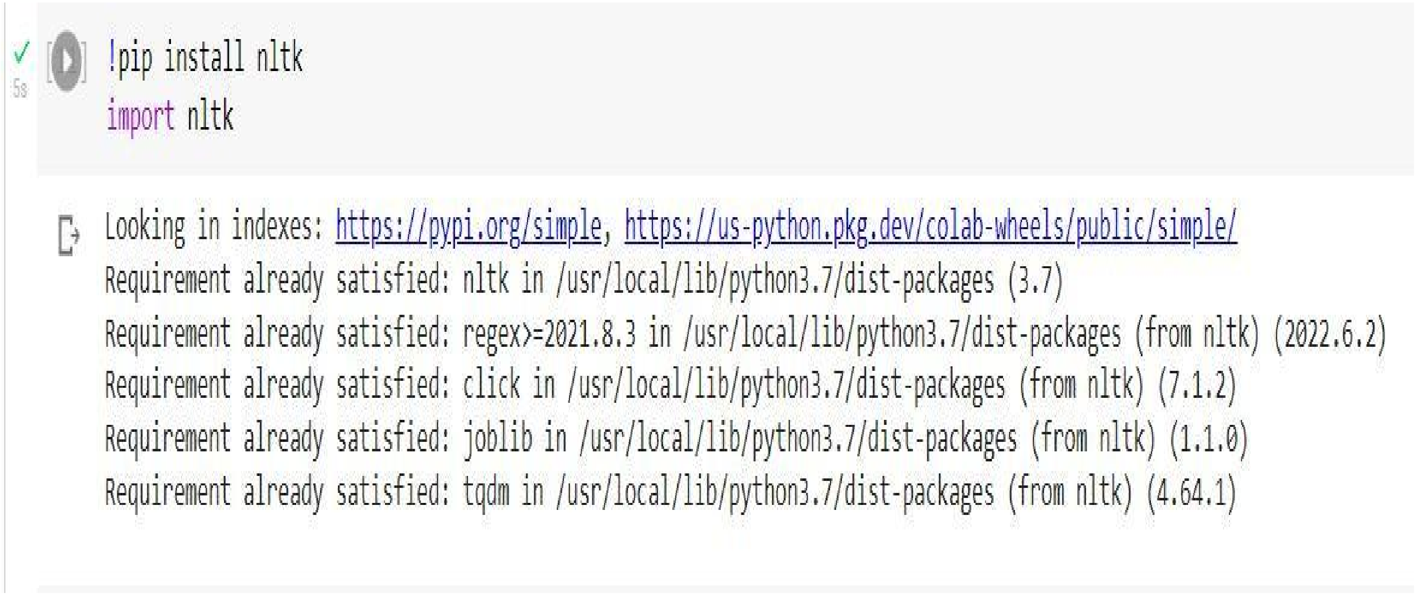
## 5.6 Splitting The Data



```
[17] #split the text

     list = data.split()
     list
```

```
['who',
 'praises',
 'india',
 's',
 'aarogya',
 'setu',
 'app',
 'says',
 'it',
 'helped',
 'in',
 'identifying',
 'covid',
 'clusters']
```

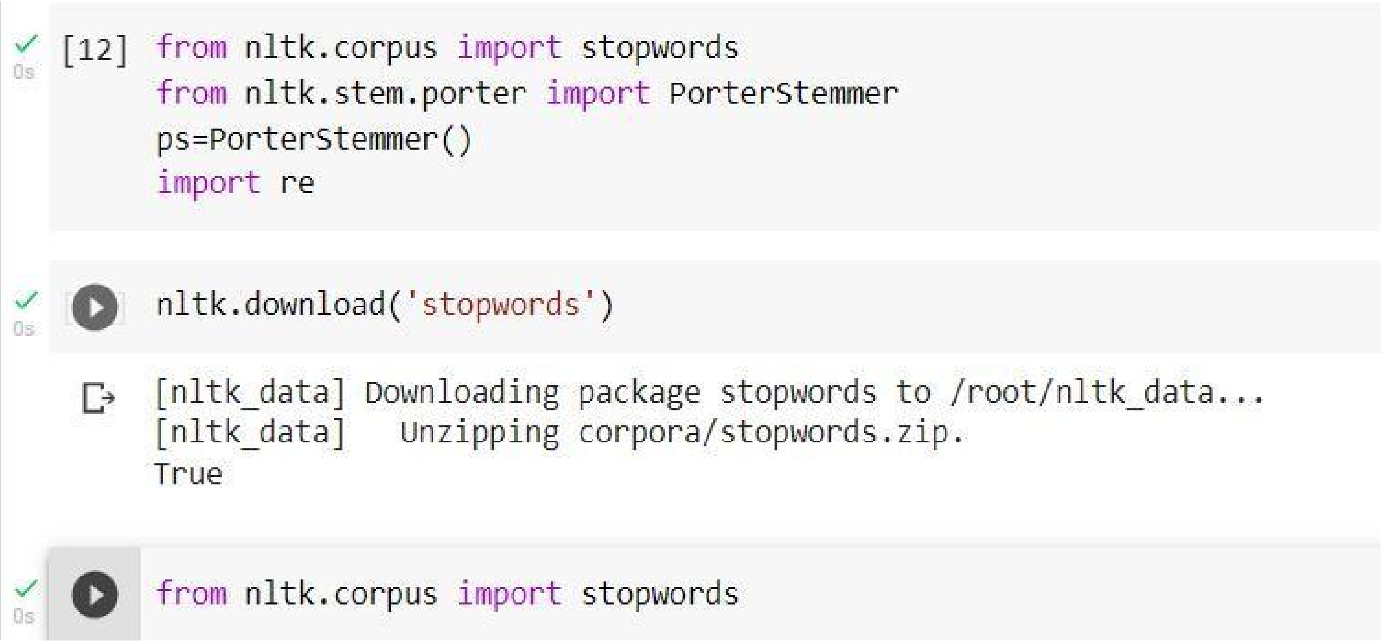
## 5.7 Installing NLP Toolkit



```
!pip install nltk
import nltk
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: nltk in /usr/local/lib/python3.7/dist-packages (3.7)
Requirement already satisfied: regex>=2021.8.3 in /usr/local/lib/python3.7/dist-packages (from nltk) (2022.6.2)
Requirement already satisfied: click in /usr/local/lib/python3.7/dist-packages (from nltk) (7.1.2)
Requirement already satisfied: joblib in /usr/local/lib/python3.7/dist-packages (from nltk) (1.1.0)
Requirement already satisfied: tqdm in /usr/local/lib/python3.7/dist-packages (from nltk) (4.64.1)
```

## 5.8 Importing Stopword Removal



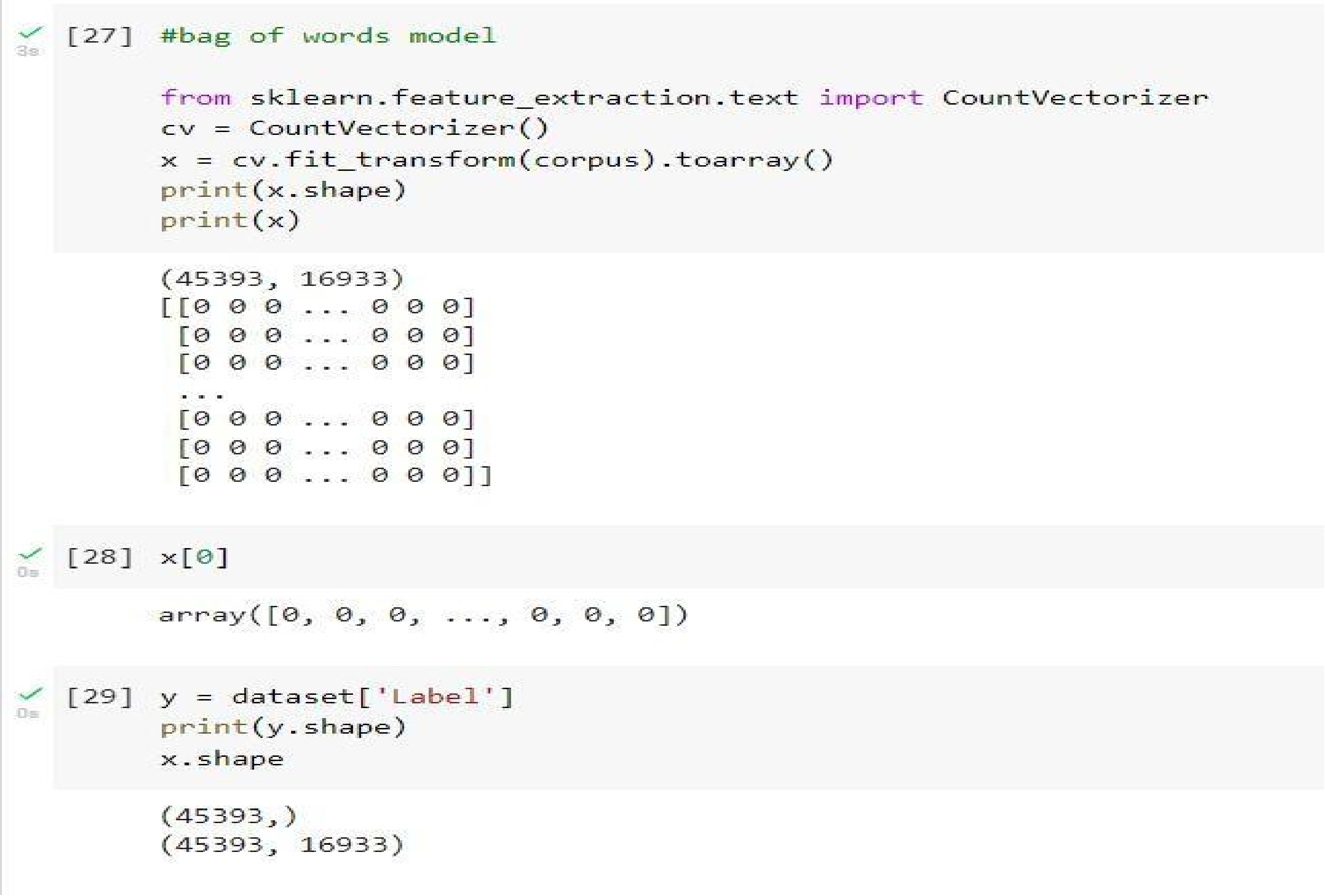
```
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer
ps=PorterStemmer()
import re
```

```
nltk.download('stopwords')
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
True
```

```
from nltk.corpus import stopwords
```

## 5.9 Bag Of Words Model



```
#bag of words model

from sklearn.feature_extraction.text import CountVectorizer
cv = CountVectorizer()
x = cv.fit_transform(corpus).toarray()
print(x.shape)
print(x)
```

```
(45393, 16933)
[[0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 ...
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]]
```

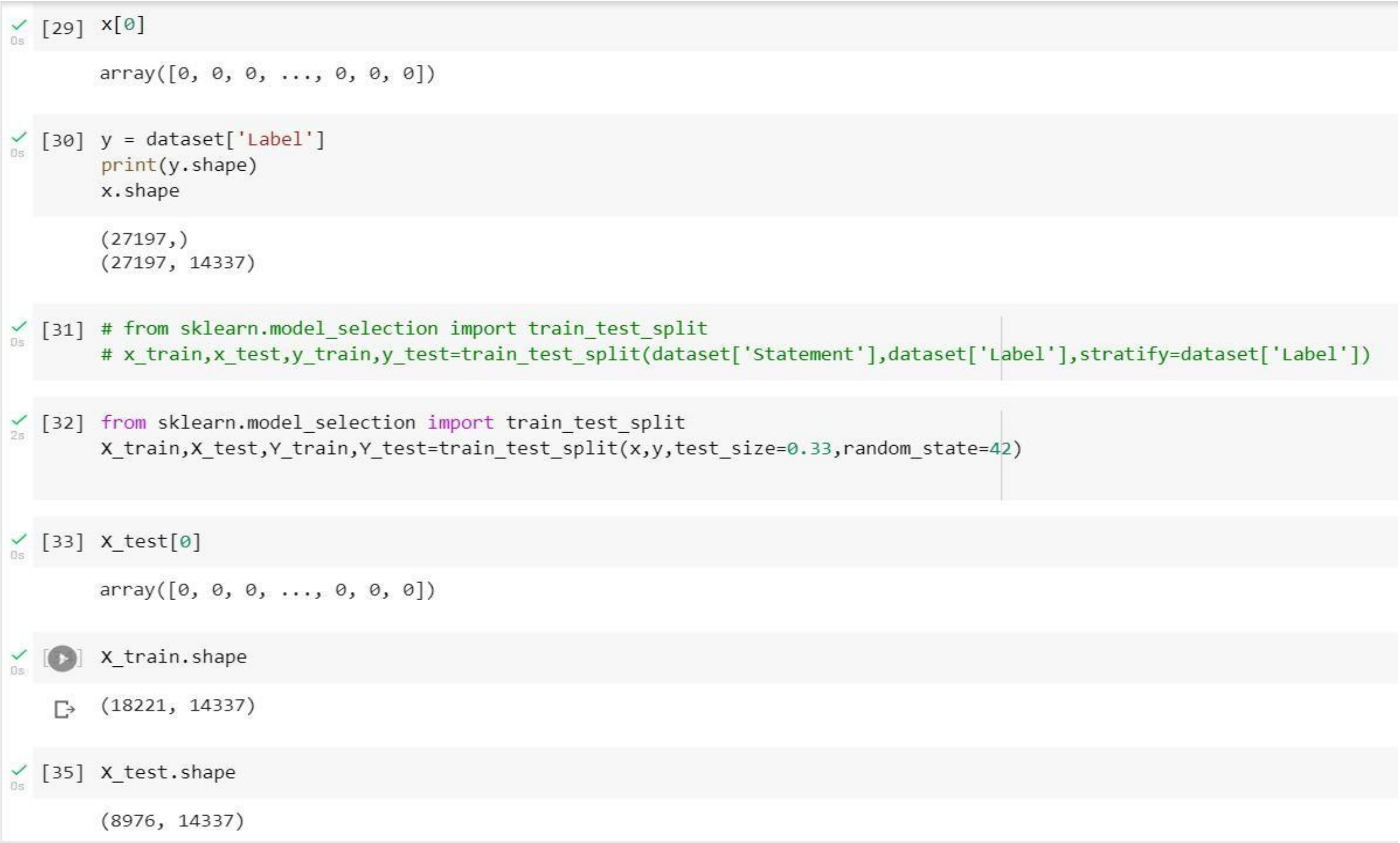
```
x[0]
```

```
array([0, 0, 0, ..., 0, 0, 0])
```

```
y = dataset['Label']
print(y.shape)
x.shape
```

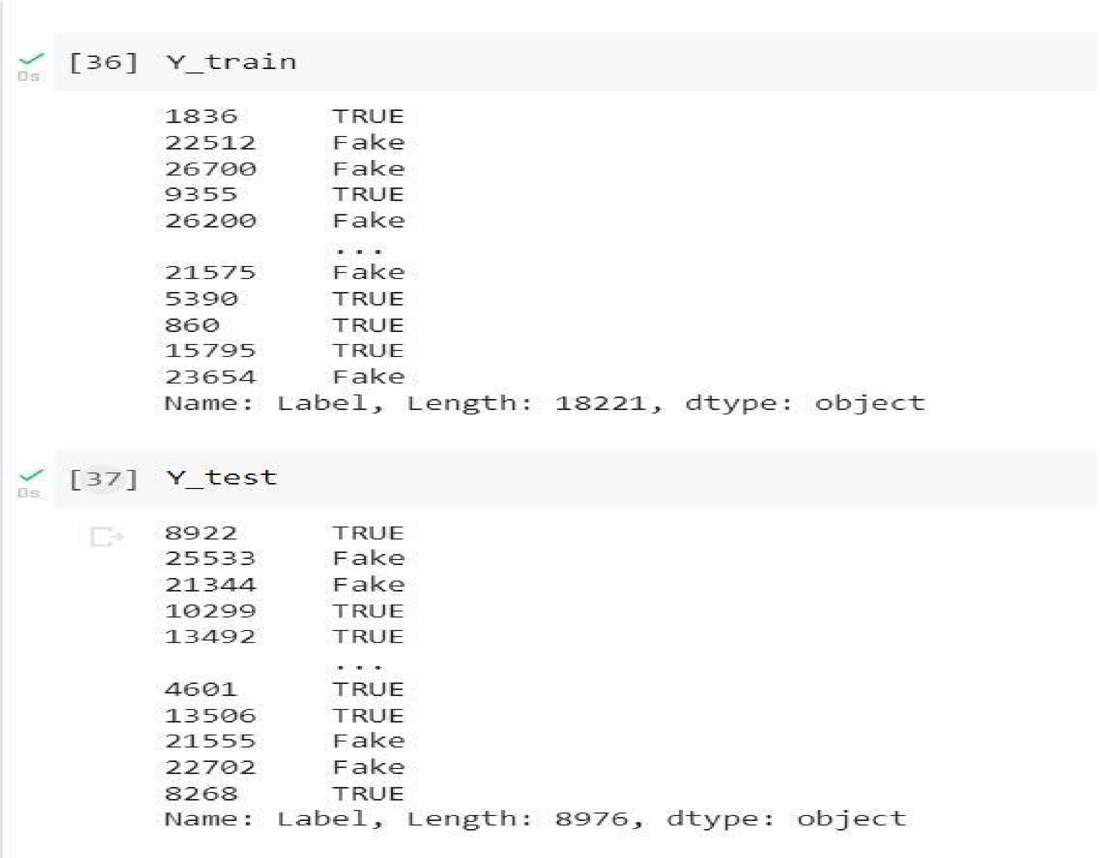
```
(45393,)
(45393, 16933)
```

## 5.10 Training The Dataset



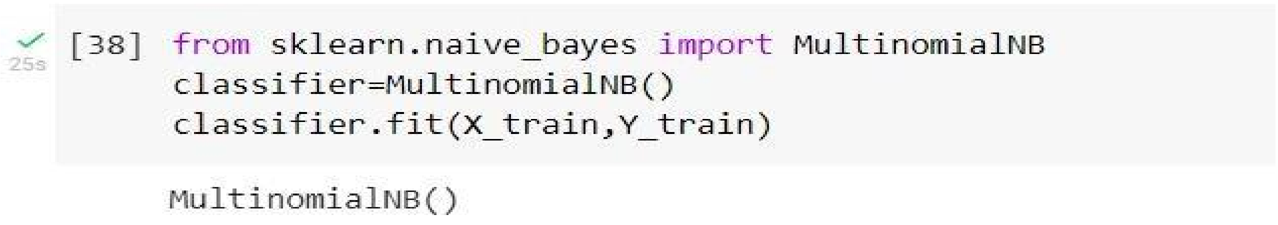
```
[29] x[0]

     array([0, 0, 0, ..., 0, 0, 0])

[30] y = dataset['Label']
     print(y.shape)
     x.shape

     (27197,)
     (27197, 14337)

[31] # from sklearn.model_selection import train_test_split
     # x_train,x_test,y_train,y_test=train_test_split(dataset['Statement'],dataset['Label'],stratify=dataset['Label'])

[32] from sklearn.model_selection import train_test_split
     X_train,X_test,Y_train,Y_test=train_test_split(x,y,test_size=0.33,random_state=42)

[33] X_test[0]

     array([0, 0, 0, ..., 0, 0, 0])

     X_train.shape

     (18221, 14337)

[35] X_test.shape

     (8976, 14337)
```

## 5.11 Testing The Dataset

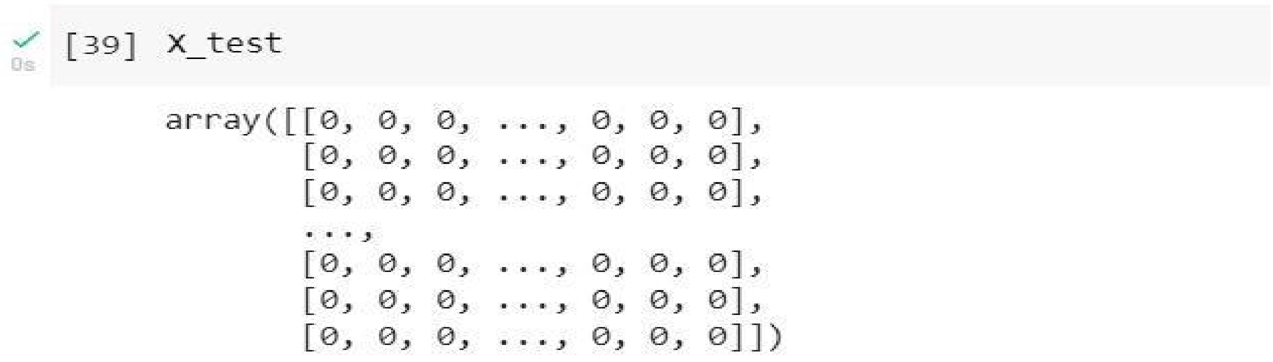


```
[36] Y_train

     1836      TRUE
     22512     Fake
     26700     Fake
     9355      TRUE
     26200     Fake
               ...
     21575     Fake
     5390      TRUE
     860       TRUE
     15795     TRUE
     23654     Fake
     Name: Label, Length: 18221, dtype: object

[37] Y_test

     8922      TRUE
     25533     Fake
     21344     Fake
     10299     TRUE
     13492     TRUE
               ...
     4601      TRUE
     13506     TRUE
     21555     Fake
     22702     Fake
     8268      TRUE
     Name: Label, Length: 8976, dtype: object
```

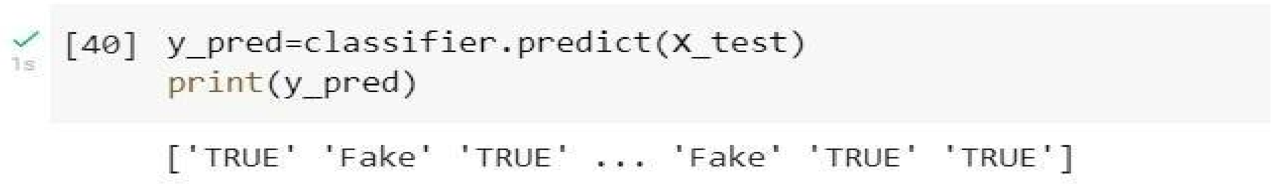
## 5.12 Naive Baye’s Classifier



```
[38] from sklearn.naive_bayes import MultinomialNB
     classifier=MultinomialNB()
     classifier.fit(X_train,Y_train)

     MultinomialNB()
```

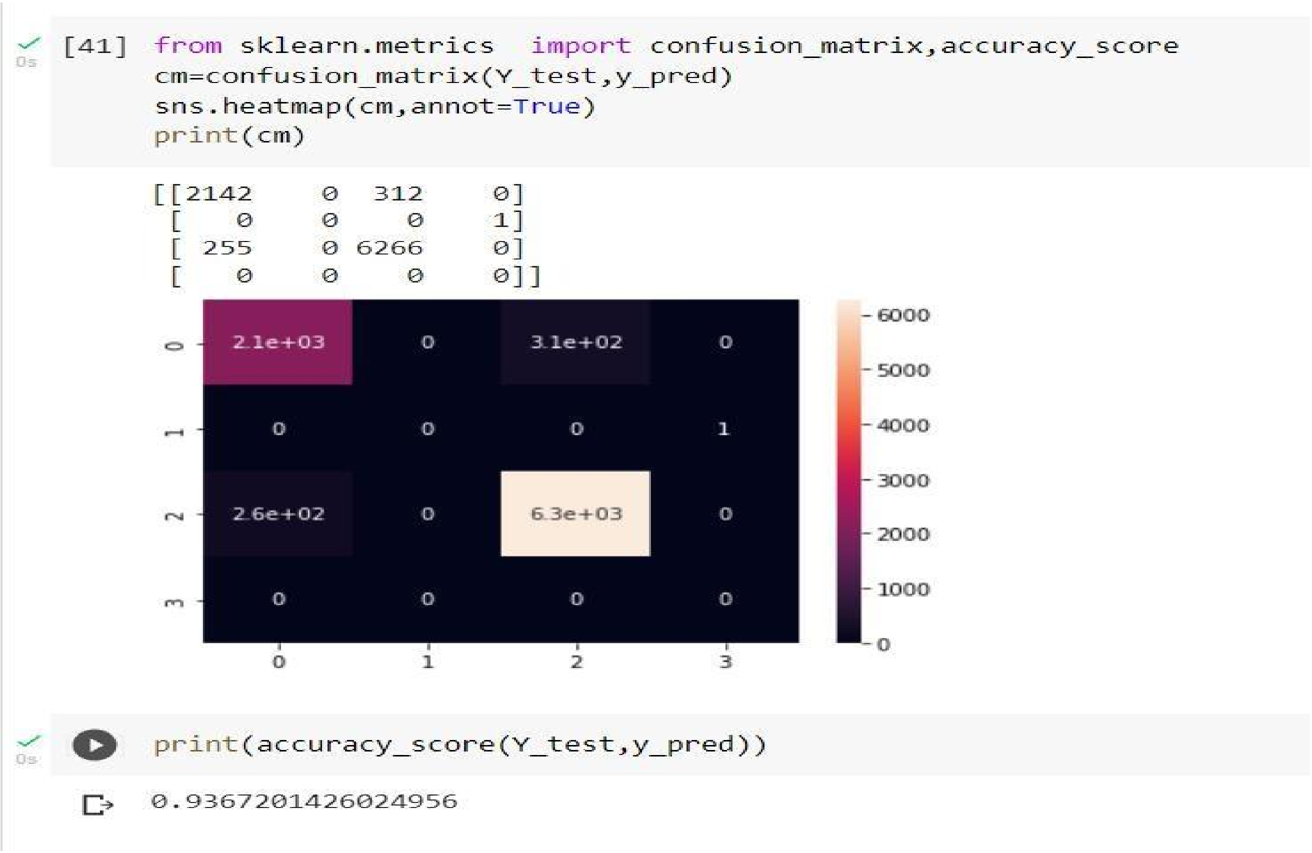


```
[39] X_test

     array([[0, 0, 0, ..., 0, 0, 0],
            [0, 0, 0, ..., 0, 0, 0],
            [0, 0, 0, ..., 0, 0, 0],
            ...,
            [0, 0, 0, ..., 0, 0, 0],
            [0, 0, 0, ..., 0, 0, 0],
            [0, 0, 0, ..., 0, 0, 0]])
```



```
[40] y_pred=classifier.predict(X_test)
     print(y_pred)

     ['TRUE' 'Fake' 'TRUE' ... 'Fake' 'TRUE' 'TRUE']
```

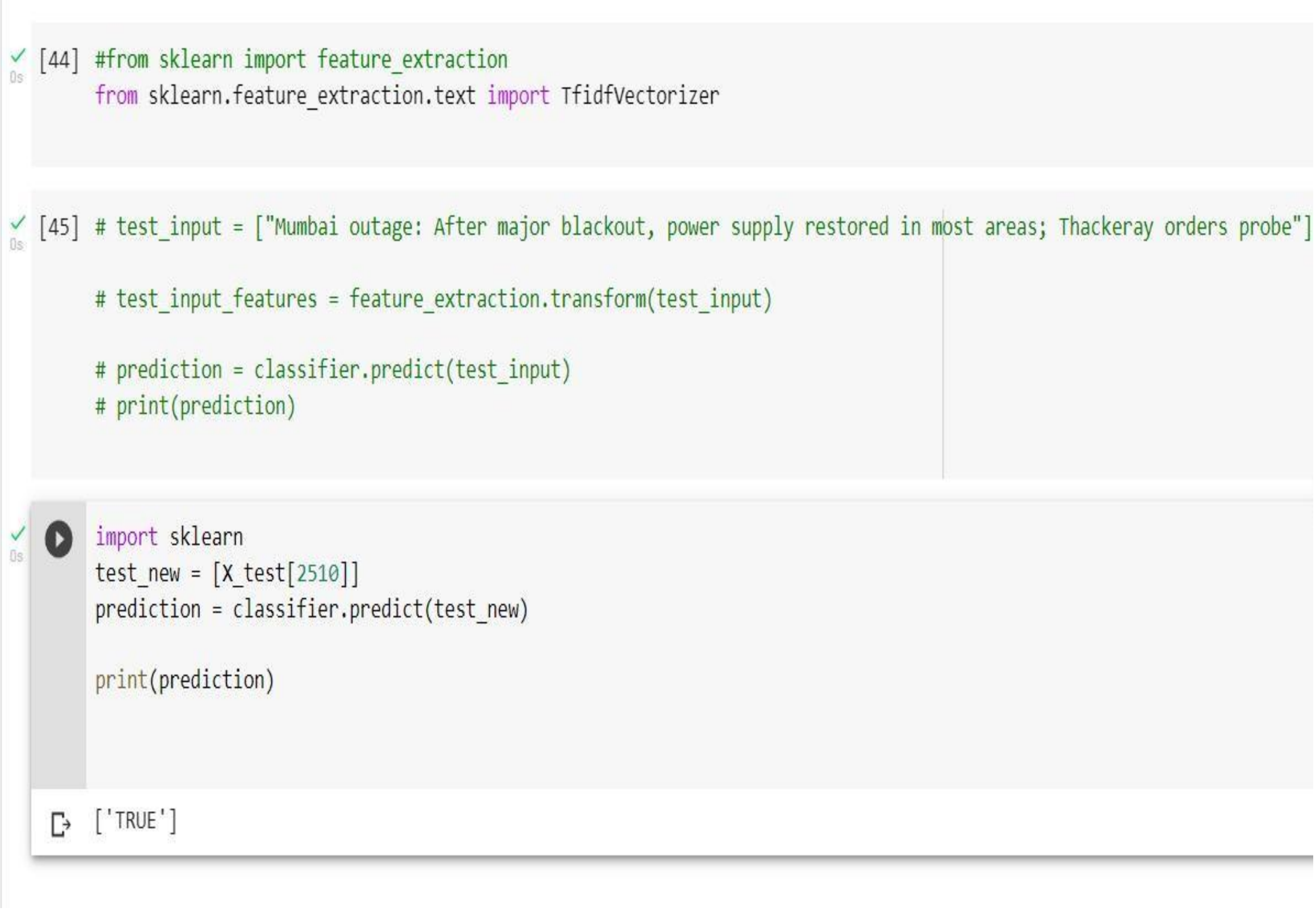
## 5.13 Accuracy Score



```
[41] from sklearn.metrics  import confusion_matrix,accuracy_score
     cm=confusion_matrix(Y_test,y_pred)
     sns.heatmap(cm,annot=True)
     print(cm)

     [[2142    0  312    0]
      [   0    0    0    1]
      [ 255    0 6266    0]
      [   0    0    0    0]]
```

```
print(accuracy_score(Y_test,y_pred))

0.9367201426024956
```

## 5.14 Testing For Examples



```
[44] #from sklearn import feature_extraction
     from sklearn.feature_extraction.text import TfidfVectorizer
```

```
[45] # test_input = ["Mumbai outage: After major blackout, power supply restored in most areas; Thackeray orders probe"]

     # test_input_features = feature_extraction.transform(test_input)

     # prediction = classifier.predict(test_input)
     # print(prediction)
```

```
import sklearn
test_new = [X_test[2510]]
prediction = classifier.predict(test_new)

print(prediction)
```

```
['TRUE']
```

# CHAPTER - 6

# RESULTS AND DISCUSSIONS

## 6.1 Results And Discussions

Splitting the dataset and appliying the models.

```
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.33,
random_state = 42)
```

**Machine learning model: Naive bayes:**

```
classifier=MultinomialNB()
classifier.fit(x_train,y_train)
```

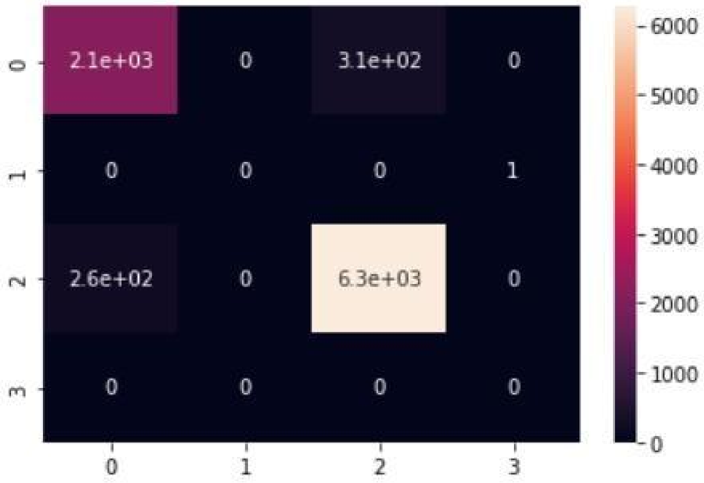
Predicting the x_test data:

```
y_pred=classifier.predict(x_te
st)
```

The naive bayes machine learning model are applied to the vectorized list(corpus). The algorithm gives the good accuracy when compared with other machine learning model. Naive bayes is good in classifying the text data.

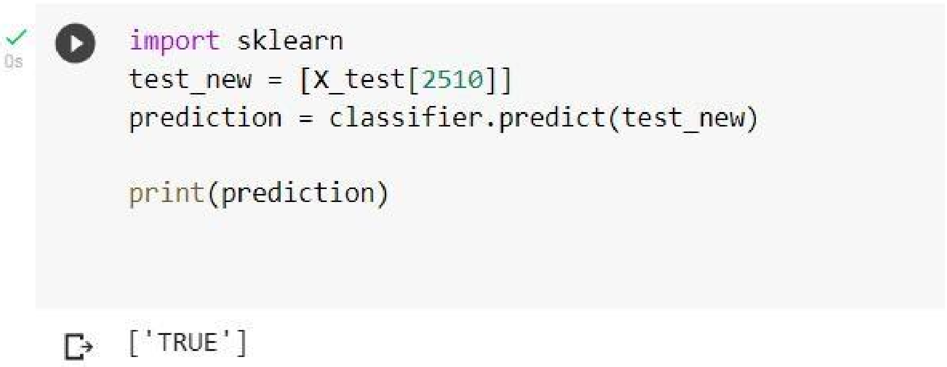
**confusion matrix to visualize the accuray**

```
cm=confusion_matrix(y_test,y_pred)
sns.heatmap(cm,annot=True)
Cm
print(accuracy_score(y_test,y_pred))
    0.9367201426024956
```



## Confusion Matrix

## Checking With Dataset Labels



```
import sklearn
test_new = [X_test[2510]]
prediction = classifier.predict(test_new)

print(prediction)
```

```
['TRUE']
```

# CHAPTER - 7

## CONCLUSION AND FUTURE ENHANCEMENT

Spotting fake news in the social media platforms is a difficult task asthe news stories are dynamic and any user in the social media platforms can post anything since they won't verify the user post. Therefore, the user may post fake news in the social media platforms to ruin reputation of a person or a firm.We proposed a Natural Language Processing to tackle fake news or misinformation. We employ Natural Language Processing to build automatic online fake news detection since it concerned with the understand and respond to the text data or spoken data. Itrespond to the given data the same way humans do. In our methodology, first, we retrieved data from an online news website and social media. We are working with IFND dataset that contains national wide news statements. Next, the natural language processing analyzes the retrieved news. The Python programming provides a wide range of libraries and tools for various NLP tasks. Natural Language Toolkit is the open-source collection of libraries, programs, and resources for building NLP programs. The Natural Language Toolkit includes libraries for NPL tasks such as stemming, lemmatization, tokenization(methods that are used for breaking sentences into tokens and trimming words) etc. Next, We have used translator to detect whether the news is real or fake when the data is given in any language. Lastly, machine learning receives the feature data and classifies the news articles into two classes: Real and Fake.

# REFERENCES

[1].Ahmad, T.; Faisal, M.S.; Rizwan, A.; Alkanhel, R.; Khan, P.W.; Muthanna, A. Efficient Fake News Detection Mechanism Using Enhanced Deep Learning Model. Appl. Sci. 2022.

[2].Meesad, P. Thai Fake News Detection Based on Information Retrieval, Natural Language Processing and Machine Learning. SN COMPUT. SCI. 2, 425 (2021)

[3].Jamal Abdul Nasir, Osama Subhani Khan, Iraklis Varlamis, Fake news detection: A hybrid CNN-RNN based deep learning approach, International Journal of Information Management Data Insights, Volume 1, Issue 1, 2021.

[4].Z. Shahbazi and Y. -C. Byun, "Fake Media Detection Based on Natural Language Processing,2021.

[5].Fake News Detection Using Machine Learning Approaches, B N Alwasel1, H Sirafi1 and M Rashid, Z Khanam et al 2021.

[6]. Kaliyar, R.K., Goswami, A. & Narang, P. DeepFakE: improving fake news detection using tensor decomposition-based deep neural network. JSupercomput 77, 1015–1037 (2021).

[7].Uma Sharma, Sidarth Saran, Shankar M. Patil, 2021, Fake News Detection using Machine Learning Algorithms, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) NTASU – 2020.

[8].Zervopoulos A., Alvanou A.G., Bezas K., Papamichail A., Maragoudakis M., Kermanidis K. Hong Kong Protests: Using Natural Language Processing for Fake News Detection on Twitter (2020).

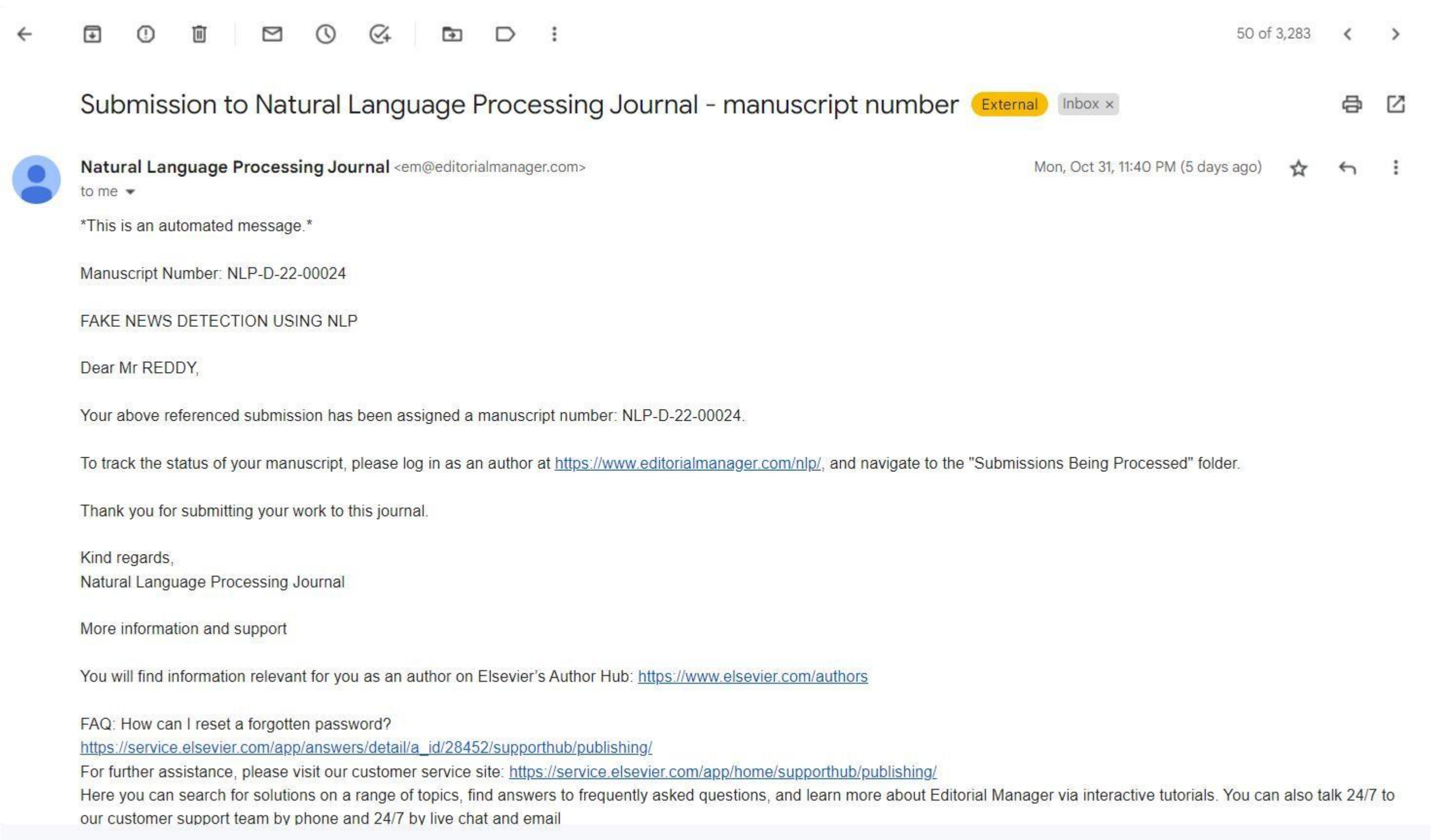
[9].J.C. S. Reis, A. Correia, F. Murai, A. Veloso and F. Benevenuto, "Supervised Learning for Fake News Detection," in IEEE Intelligent Systems, vol. 34, no. 2, pp. 76-81, March-April 2019.

[10].“Fake News Detection using Machine Learning and Natural Language Processing” Kushal Agarwalla, Shubham Nandan, Varun Anil Nair, D. Deva Hema,IJRTE, Vol-6, Issue-6, March 2019.

# APPENDIX A

## PUBLICATION DETAILS

We Submitted Our Research Paper For Publication At Editorial Manager[Natural Language Processing Journal].We Got The Mail From Natural Language Processing Journal On Oct 31st,2022 With Manuscript Number : NLP-D-22-00024.



Submission to Natural Language Processing Journal - manuscript number

**Natural Language Processing Journal** <em@editorialmanager.com>
to me

Mon, Oct 31, 11:40 PM (5 days ago)

*This is an automated message.*

Manuscript Number: NLP-D-22-00024

FAKE NEWS DETECTION USING NLP

Dear Mr REDDY,

Your above referenced submission has been assigned a manuscript number: NLP-D-22-00024.

To track the status of your manuscript, please log in as an author at https://www.editorialmanager.com/nlp/, and navigate to the "Submissions Being Processed" folder.

Thank you for submitting your work to this journal.

Kind regards,
Natural Language Processing Journal

More information and support

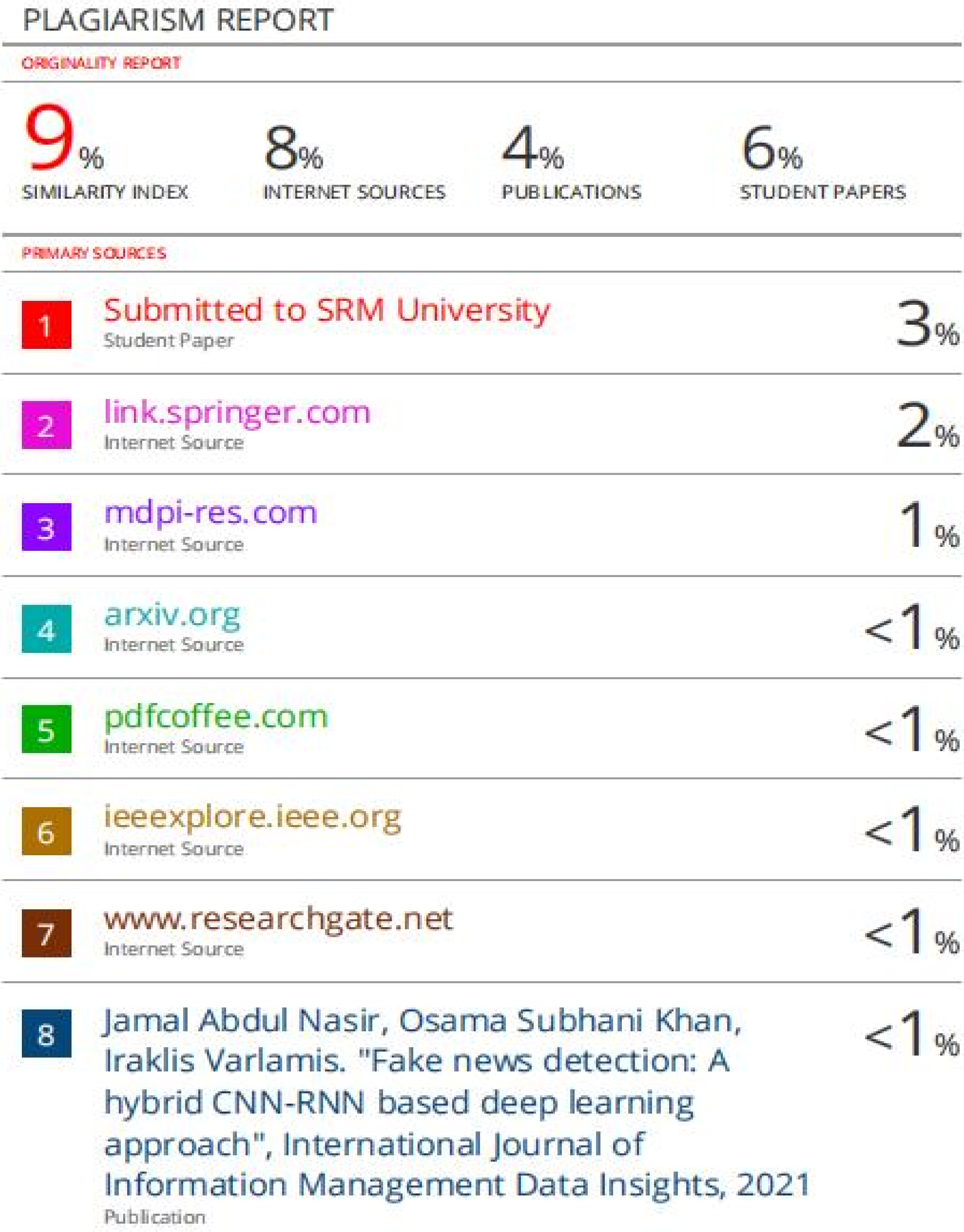
You will find information relevant for you as an author on Elsevier's Author Hub: https://www.elsevier.com/authors

FAQ: How can I reset a forgotten password?
https://service.elsevier.com/app/answers/detail/a_id/28452/supporthub/publishing/
For further assistance, please visit our customer service site: https://service.elsevier.com/app/home/supporthub/publishing/
Here you can search for solutions on a range of topics, find answers to frequently asked questions, and learn more about Editorial Manager via interactive tutorials. You can also talk 24/7 to our customer support team by phone and 24/7 by live chat and email

**Fig.1:Publication Notification**

# APPENDIX B

## PLAGIARISM REPORT



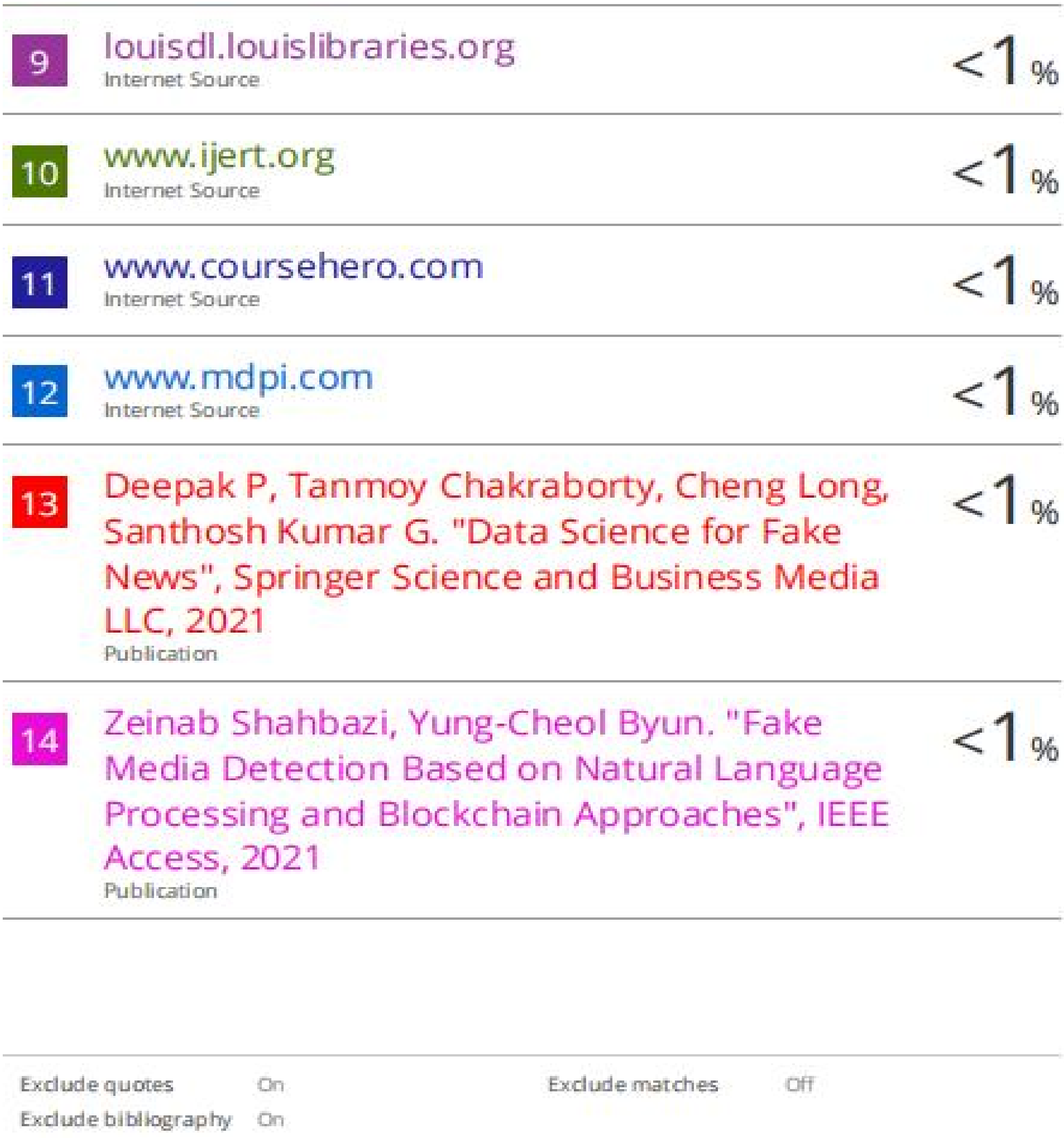
PLAGIARISM REPORT

ORIGINALITY REPORT

| 9% | 8% | 4% | 6% |
|---|---|---|---|
| SIMILARITY INDEX | INTERNET SOURCES | PUBLICATIONS | STUDENT PAPERS |

PRIMARY SOURCES

| | | |
|---|---|---|
| 1 | Submitted to SRM University<br>Student Paper | 3% |
| 2 | link.springer.com<br>Internet Source | 2% |
| 3 | mdpi-res.com<br>Internet Source | 1% |
| 4 | arxiv.org<br>Internet Source | <1% |
| 5 | pdfcoffee.com<br>Internet Source | <1% |
| 6 | ieeexplore.ieee.org<br>Internet Source | <1% |
| 7 | www.researchgate.net<br>Internet Source | <1% |
| 8 | Jamal Abdul Nasir, Osama Subhani Khan, Iraklis Varlamis. "Fake news detection: A hybrid CNN-RNN based deep learning approach", International Journal of Information Management Data Insights, 2021<br>Publication | <1% |



| | | |
|---|---|---|
| 9 | louisdl.louislibraries.org<br>Internet Source | <1 % |
| 10 | www.ijert.org<br>Internet Source | <1 % |
| 11 | www.coursehero.com<br>Internet Source | <1 % |
| 12 | www.mdpi.com<br>Internet Source | <1 % |
| 13 | Deepak P, Tanmoy Chakraborty, Cheng Long, Santhosh Kumar G. "Data Science for Fake News", Springer Science and Business Media LLC, 2021<br>Publication | <1 % |
| 14 | Zeinab Shahbazi, Yung-Cheol Byun. "Fake Media Detection Based on Natural Language Processing and Blockchain Approaches", IEEE Access, 2021<br>Publication | <1 % |

| | | | |
|---|---|---|---|
| Exclude quotes | On | Exclude matches | Off |
| Exclude bibliography | On | | |