**MINI PROJECT REPORT**

**Fake News Detection**

*by*

**Abhinav Maurya Entry No.: 19BCS003**

**Abheet Kumar Entry No.: 19BCS002**

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

*of*

**BACHELOR OF TECHNOLOGY**

*in*

COMPUTER SCIENCE & ENGINEERING



**SHRI MATA VAISHNO DEVI UNIVERSITY, KATRA**

**(School of Computer Science & Engineering)**

**JAMMU & KASHMIR – 182 320**

**Session 2021-22**

**Acknowledgments**

We two have made efforts towards this project. However, it would not have been possible without the guidance of Dr. Manoj Kumar Gupta at every moment of the project, he had been supporting and made valuable remarks over which we made changes and got better results than what we could have got.

We would also thank all of the faculty members who directly or indirectly inspired, supported, and gave their valuable suggestions. Lastly, we would like to thank all of the classmates who helped and supported us.

**Declaration**

We undersigned solemnly declare that the project report Fake News Detection is based on my work carried out during the course of our study under the supervision of Dr. Manoj Kumar Gupta**.**

We assert the statements made and conclusions are drawn are an outcome of my research work. I further certify that

1. The work contained in the report is original and has been done by us under the supervision of my supervisor.
2. The work has not been submitted to any other Institution for any other degree/diploma/certificate in this university or the any other University of India or abroad.
3. Whenever we have used materials (data, theoretical analysis, and text) from other sources, we have given due credit to them in the text of the report and given their details in the references.

**Abhinav Maurya**

**19bcs003, B. Tech, 6th Sem**

**Abheet Kumar**

**19bcs002, B. Tech, 6th Sem**

School of Computer Science & Engineering

Shri Mata Vaishno Devi University, Katra

I endorse the above declaration of the student.

**Dr Manoj Kumar Gupta**

(Name and Signature of the Supervisor)

# ABSTRACT

Fake News has become one of the major problem in the existing society. Fake News has high potential to change opinions, facts and can be the most dangerous weapon in influencing society.

The proposed project uses NLP techniques for detecting the 'fake news', that is, misleading news stories which come from the non-reputable sources. By building a model based on a K-Means clustering algorithm, the fake news can be detected . The data science community has responded by taking actions against the problem. It is impossible to determine a news as real or fake accurately. So the proposed project uses the datasets that are trained using count vectorizer method for the detection of fake news and its accuracy will be tested using machine learning algorithms.

# CONTENTS

Contents

[Acknowledgments 2](#_Toc106026441)

[Declaration 3](#_Toc106026442)

[ABSTRACT 4](#_Toc106026443)

[CONTENTS 5](#_Toc106026444)

[LIST OF FIGURES 8](#_Toc106026445)

[LIST OF TABLES Error! Bookmark not defined.](#_Toc106026446)

[CHAPTER 1 11](#_Toc106026447)

[INTRODUCTION 11](#_Toc106026448)

[1.1 MACHINE LEARNING AND NLP: 11](#_Toc106026449)

[1.1.2.1 STAGES IN NLP 13](#_Toc106026450)

[1.3 PROBLEM STATEMENT 16](#_Toc106026451)

[CHAPTER 2 17](#_Toc106026452)

[LITERATURE SURVEY 17](#_Toc106026453)

[2.1 INTRODUCTION 17](#_Toc106026454)

[2.2 REVIEW OF LITERATURE 17](#_Toc106026455)

[2.3 RELATED WORK 18](#_Toc106026456)

[CHAPTER 3 20](#_Toc106026457)

[METHODOLOGY 20](#_Toc106026458)

[3.1 PROPOSED SYSTEM 20](#_Toc106026459)

[3.2 SYSTEM ARCHITECTURE 20](#_Toc106026460)

[3.3 ALGORITHM FOR THE PROPOSED SYSTEM: 22](#_Toc106026461)

[CHAPTER 4 23](#_Toc106026462)

[DATASET 23](#_Toc106026463)

[4.1 EXISTING DATASETS FOR THIS SYSTEM: 23](#_Toc106026464)

[4.2 : PROPOSED DATASET USED: 23](#_Toc106026465)

[4.3 : FAKE NEWS SAMPLES: 25](#_Toc106026466)

[CONCEPTS 27](#_Toc106026467)

[5.1 PREPROCESSING: 27](#_Toc106026468)

[5.2 STEPS IN TEXT PRE-PROCESSING: 27](#_Toc106026469)

[5.3 COUNT VECTORIZER: 28](#_Toc106026470)

[5.3.1 Input to count vectorizer: 29](#_Toc106026471)

[5.4 WORD2VEC MODEL: 30](#_Toc106026472)

[5.4.1 WORD2VEC ALGORITHM : 31](#_Toc106026473)

[5.6.2 EVALUATION MEASURES: 31](#_Toc106026474)

[5.6.2 DIFFERENT TYPES OF EVALUATION METRICS 32](#_Toc106026475)

[5.6.3 DEFINING THE METRICS 33](#_Toc106026476)

[5.6.4 ACCURACY 33](#_Toc106026477)

[5.6.5 PRECISION 33](#_Toc106026478)

[5.6.6 RECALL 33](#_Toc106026479)

[CHAPTER 6 34](#_Toc106026480)

[Model Used 34](#_Toc106026481)

[6.1 Logistic Regression 34](#_Toc106026482)

[6.2 Binary Logistic Regression 34](#_Toc106026483)

[6.3 Confusion Matrix for Logistic Regression 34](#_Toc106026484)

[6.4 Decision Tree 34](#_Toc106026485)

[6.5 Information Gain: 35](#_Toc106026486)

[6.6 Gini Index: 35](#_Toc106026487)

[6.7 Results for Decision Tree Classification 35](#_Toc106026488)

[6.8 Gradient boosting Classifier 36](#_Toc106026489)

[6.9 Random Decision Forest 38](#_Toc106026490)

[6.10 Results for Random Forest Classifier 38](#_Toc106026491)

[CHAPTER 7 39](#_Toc106026492)

[EXPERIMENT ANALYSIS 39](#_Toc106026493)

[True.csv 41](#_Toc106026494)

[CHAPTER 8 43](#_Toc106026495)

[USER INPUT INTERFACE 43](#_Toc106026496)

[OUTPUT 44](#_Toc106026497)

[9.1 CONCLUSION: 45](#_Toc106026500)

[9.2 FUTURE WORK: 45](#_Toc106026508)

[REFERENCES 46](#_Toc106026509)

# LIST OF FIGURES

|  |  |
| --- | --- |
| **FIGURE NO.** | **TITLE** |
| 1 | Relationship of different fields in AI |
| 2 | Architecture of the proposed System. |
| 3 | Diagram showing the Work Flow |
| 4 | Description about the data set |
| 5 | Samples of fake news |
| 6 | number of fake words in the data using count vectorizer |
| 7 | number of true words in the data using count vectorizer |
| 8 | Logistic Regression Confusion Matrix |
| 9 | Diagram of Decision tree |
| 10 | Results for Decision Tree Classificatio**n** |
| 11 | Results for Gradient Boosting Classifier |
| 13 | Gradient Boosting Classifier Confusion Matrix |
| 14 | Results for Random Forest Classifier |
| 15 | Random Forest Classifier Confusion Matrix |
| 16 | True.csv |
| 17 | Fake.csv |
| 18 | User Interface |
| 19 | Output |

## SYMBOLS AND ABBREVIATIONS

**LIST OF ABBREVATIONS**

|  |  |
| --- | --- |
| **SHORT FORM** | **FULL FORM** |
| CV | Count Vectorizer |
| W2V | Word 2 Vector |
| AI | Artificial Intelligence |
| DL | Deep Learning |
| ML | Machine Learning |

# CHAPTER 1: INTRODUCTION

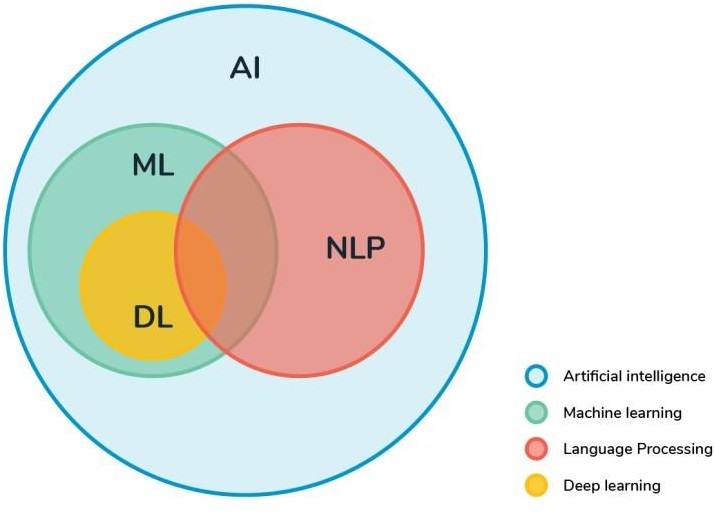
### MACHINE LEARNING AND NLP:

#### MACHINE LEARNING

Machine learning algorithms are the use of statistical models and algorithms by computers to automatically finish a task without the need for explicit instructions - they accomplish the pattern identification and inference on their own. Artificial intelligence is a subcategory of it. Machine learning algo make a modelibased on data, referred to as "train data," in command to predict or make judgments without being clearly programming to do so .

Standard definition of Machine Learning :

"A computer code is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, increases with experience E". The term "machine learning" was coined by Alan Turing.



**Fig. 1** : Relationship of different fields in AI

NLP is a field of computer science and practical wisdom related to the interaction of computers and human (natural) languages, especially how computers are programmed to manipulate the data of natural language to give meaningful result from that data. In NLP our main motive to make use of different machine learning Algorithms to make machine draw out the context or say the meaningful things from our natural language data.

### STAGES IN NLP

#### **Morphological Analysis**

Morphological Analysis includes the identifications of the structure of words. This analysis looks for the morphemes which means the smallest unit of word. Lexical analysis finds out the relationship between the morphemes and the root word.

#### SYNTAX ANALYSIS

Syntactic Analysis ensures that the structure of the given piece of the text is correct or not. In Simple words we can say it checks for the grammar at the sentence level.

#### SEMANTIC ANALYSIS

Semantic Analysis deals with drawing out the meaning from the given text.

#### DISCOURSE INTEGRATION

Deals with the result achieved in the previous phase to find out the relationship between the words in the sentence. Example “Ram is good student. He spend most time in library”. Here “he” refers to “Ram”.

#### PRAGMATIC ANALYSIS

This is the final step of NLP, we interpret the information achieved from the previous Step.

#### INSPIRATION OF WORK

During the Election there is an increase in fake news. The 2016 edition showed not only the risks of outcomes of the false newsibut also the tasks posed when attempts to distinguish false stories from real ones. Fake stories may be a new term, but they are by no means new. False stories have been around at least since the rise of the one-sided newspapers, which are part of the 20th century. However, new technology and the various media outlets are increasing the trends in fake news. Thus, to overcome the effect of the fake done in the past and make system that would help in future in identification of the fake news.

### PROBLEM STATEMENT

The use of news is like a sword which have double edges. On the other hand,its spreads very fast like a fire in the forest and misleads the people. It allows for a wide-ranging of “false stories”, that is, intentionally low-level stories with deliberate false information. The extensive spread of false news has the potential to have far-reaching effects on individuals and communities.

Thus the system should be designed so that the algorithm.must be able to spot false news in a particular situation.

# CHAPTER 2: LITERATURE SURVEY

### INTRODUCTION

In 20th generation the technologies are increasing at a high rate. And now sharing of the information have become very easy task. Internet is making our life easier and is helping to access a lot of information. We can see many news channel website some authorized some unauthorized sharing wide variety of news whenever we check that news or reads it appeals to be more realistic. The way they write them makes t look that this is true news. Today anyone can publish information on the internet that will be consumed by the WWW. People are misled by these fake news. Social media plays a very important role in supplying these fake news.

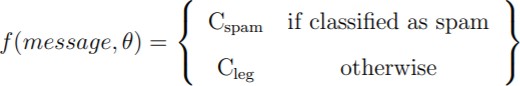
### REVIEW OF LITERATURE

Scientists consider that the support of IT specialists in reducingkfake news is very important. In order to deal with them, researchers have used data mining as a technique. Data integration is used to detect fake news using datamining. In order to combat clickbait, organizations invest a lot of resources.

### FAKE NEWS DETECTION ASSOCIATED WORK

#### SPAM DETECTION

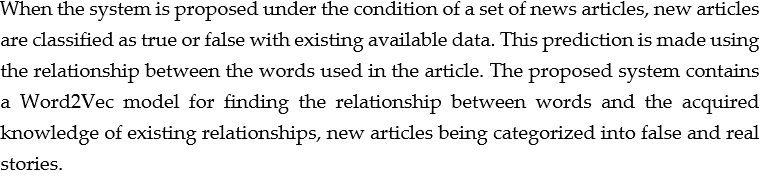
The problem of finding inaccurate sources of info through content-based analysis is considered to be solved at least on a spam detection site, spam detection using mathematical learning techniques to categorize the text as spam or not spam. These strategies include pre-processing of the text, removal of the feature and using the feature selection because of which we can get the best performance on the test. Once these features have been identified, they can be categorized using TF-IDF. They are a feature of supervised machine learning, which means they need some labeled data in order to learn the function.



Both the task of detecting fake news and spam detection aim to distinguish examples of legitimate text from examples of illegitimate, ill-intended text.

# CHAPTER 3: METHODOLOGY

### PROJECTED SYSTEM



### SYSTEM STRUCTURE

### 

Data set is taken from Kaggle. Libraries of python and jupyter notebook the are used in the step above. The vectorizer calculation method is used in the first step. For false news, we have to train the system using a train dataset.kBefore getting into getting fake news , the entire database is divided into two data sets. 75% is used for training and 25% is used for testing. For training, train dataset is used for model training using a train database. For testing, the test data is provided as input and output is predictable. After the test, the confusion matrix is created with the help of the actual and the predicted output. The confusion matrix provides information about the number of correct and.incorrect output in the category of news.

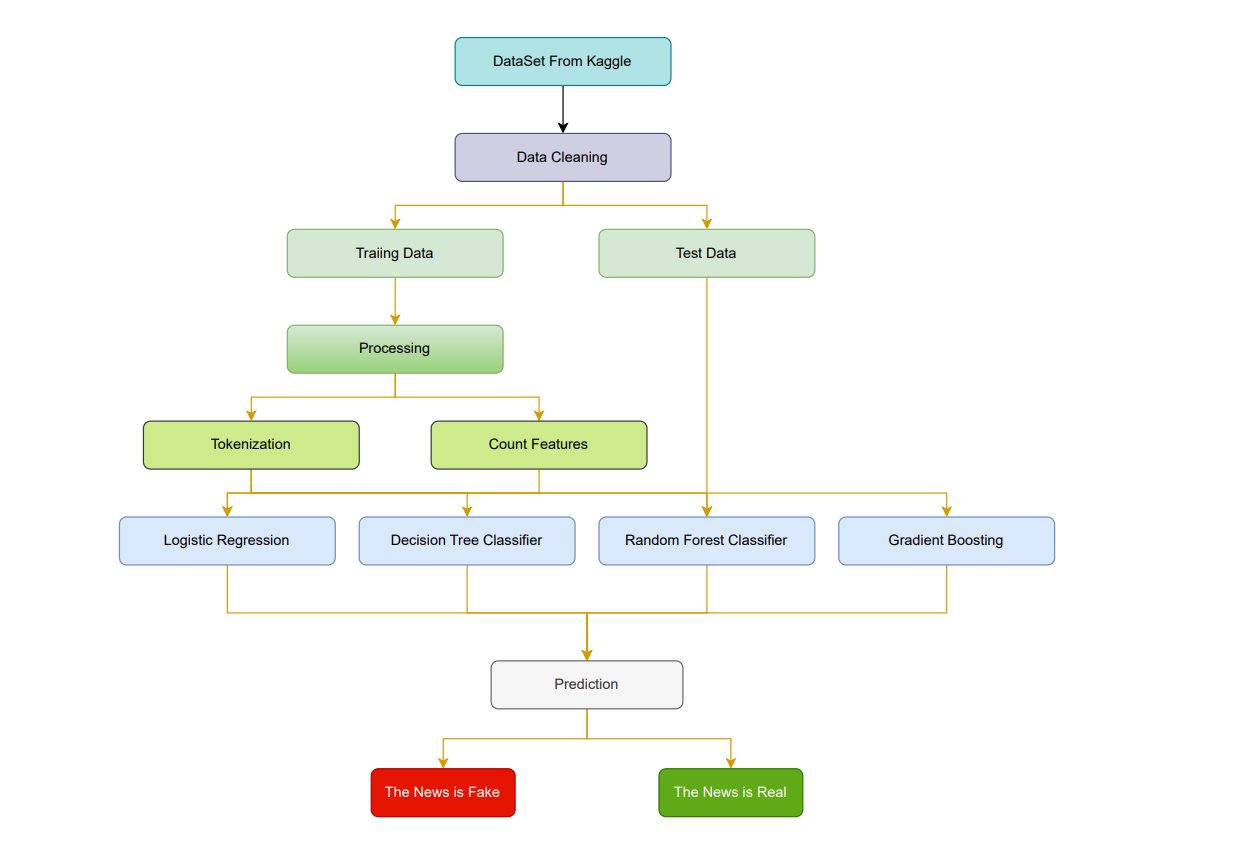


Fig. Work flow diagram

### STEPS FOR THE PROPOSED SYSTEM:

Step 1: Start

Step 2: From Internet sources inputs are collected to create dataset.

Step 3: After processing the data the dataset is separated into the test and train data.

Step 4: Train data be converted into the vector of real no using Count Vectorization.

Step 5: Different Classification Algorithms are used to train the algorithms.

Step 6: Accuracy is calculated from the confusion Matrix.

## CHAPTER 4: DATASET

### EXISTING DATASETS FOR THIS SYSTEM:

Lack of data for non-self-labeled stories is key to developing more focused, text-based models that cover multiple topics. The counterfeit database is not in line with our intent because it contains the relationship between the text about its truthfulness. Labeled data are available from various sources, such as Twitter, Amazon Updates, for more simple and common NLP segmentation tasks. When it comes to identifying false and true stories, the same does not apply. Using supervised machine learning techniques, researchers and data scientists are faced with a challenge. As part of our research, we explored published databases for sentence-level separation and methods for creating complete sets of documents categorized according to their quality and consistency, along with good and negative examples.

### : PROPOSED DATASET USED:

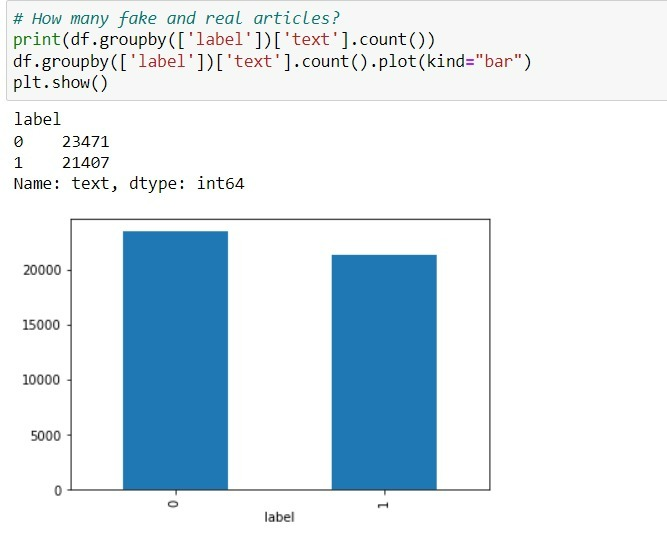
For document level classification of fake news, there is no dataset of comparable quality to the Kaggle dataset on fake news detection. Therefore, we can use document headlines as sentances or create adataset which is mixture of both fake and true news. In this example, two datasets that contain real and unreal news examples are combined for informal and exploratory analysis of the data which helps in better understanding of the data. 

Fig. Dataset description

### : SAMPLES NEWS:

The system uses a fake news data which was taken from the Kaggle.

As clearly we can see in the sample that our data set contains 4 features and these are : ‘title’ , ‘text’ ,’subject’ , ‘date’ .

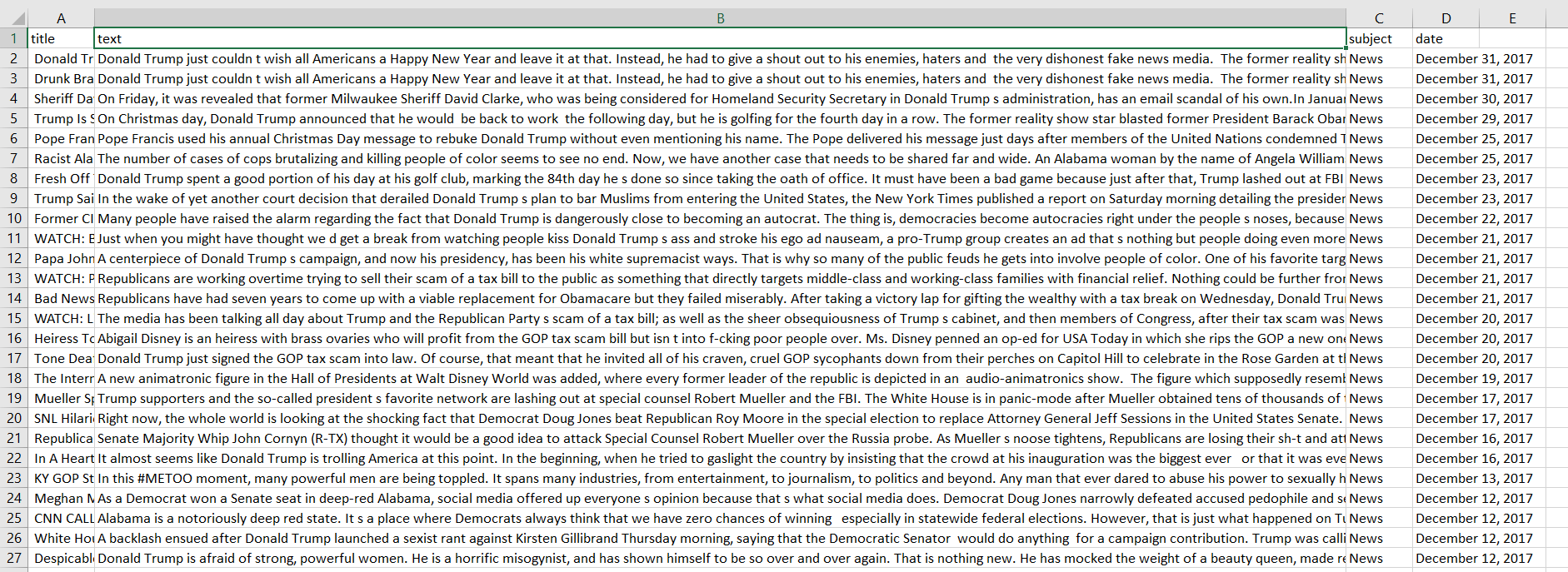


Fig. Fake news samples

# CHAPTER 5: CONCEPTS

### PREPROCESSING:

Data preprocessing is the phase in any Computer Learning process when the data is altered or encoded to make it easier for the machine to parse. To put it another way, the algorithm can now readily comprehend the characteristics of the data.

Pre-processing is the most important thing to perform when it comes to detecting bogus news. To begin, because the data dataset was compiled from a variety of sources, any extraneous data should be eliminated, changed to lowercase, punctuation, signs, and stopwords removed.

### TEXT PRE-PROCESSING STEPS :

#### NORMALIZATION OF TEXT:

The procedure of converting text into a single recognized form is known as text normalization. Normalizing text before st0ring or processing it all0ws the system to separate required data from the rest, allowing it to deliver consistent data as an input to the algorithm's further stages.

#### REMOVAL OF STOP WORD

* + - 1. **StopnWord**:

A Stop Term is common word in a language, such as "a, an, the, for, is, was, which, , your selves, and so on."

The Stop words occurs so many times so they need to be eliminated from the data so that the other factors can be considered. Stop word removal is pre processing of the text data to remove the stop words, punctuations and other symbols which are less important to us. The result is a vector with the length of the text and count of how many times the word appear in the data set.

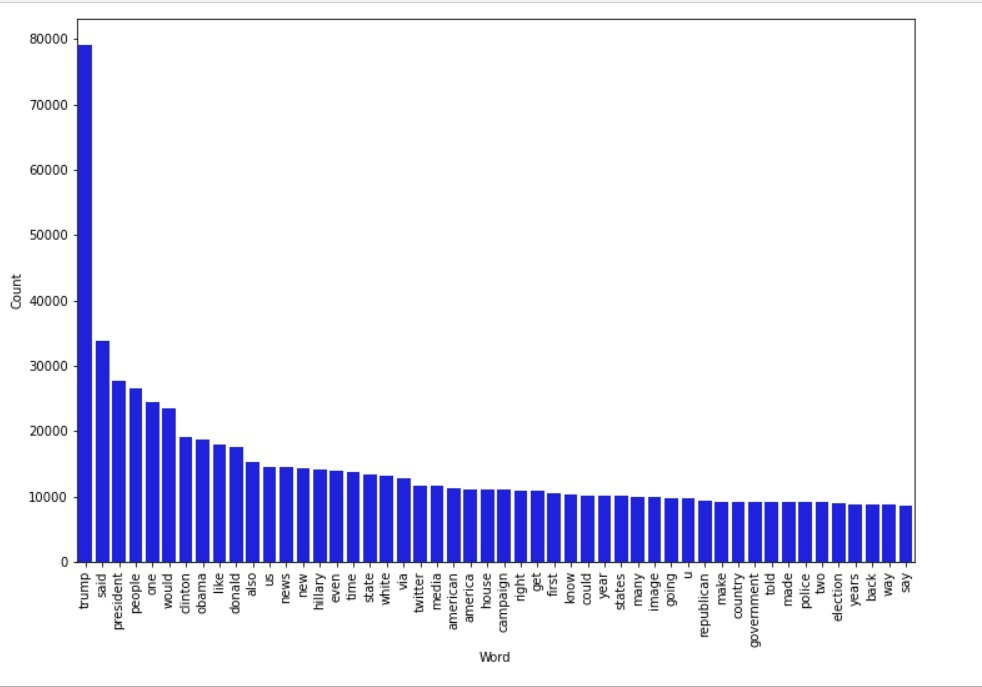
Let’s take a sentence:

"that was the right way of using tabular transformation ." ["right", "method", "tabular", "transformation"] is the o/p after stop word removal.

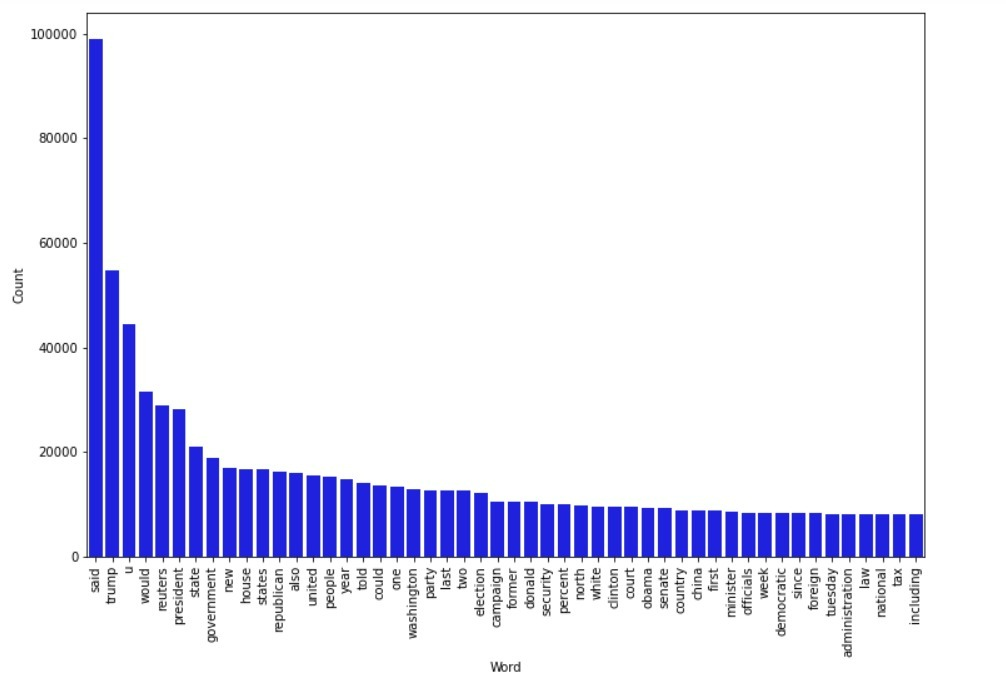
### COUNT VECTORIZER:

Count Vectorizer breaks the expression as well as carrying out very general preparation such as eleminating punctuaton marks, altering all arguments to small letter, and so on. A terminology of recognised arguments is developed, which will then be utilised to encode unknown material. An programmed vector with the complete terminology's length and an integer of total for frequency of each word seems in the text is returned.

### count vectorizer from input:







****

### WORD2VEC MODEL:

WORD2VEC is a type 0f model which depicts a word in a huge text data as a vector in n-dimensi0nal space, carrying related word close. The Skip- Gram model is an example of such a models. It can be applied to big datasets to learn word embeddings. Word2vec-learned embeddings have performed well on a range of downstream NLP tasks.

### WORD2VEC ALGORITHM :

1.Begin with a group of well assigned words as input.

2.Think about the gap size.

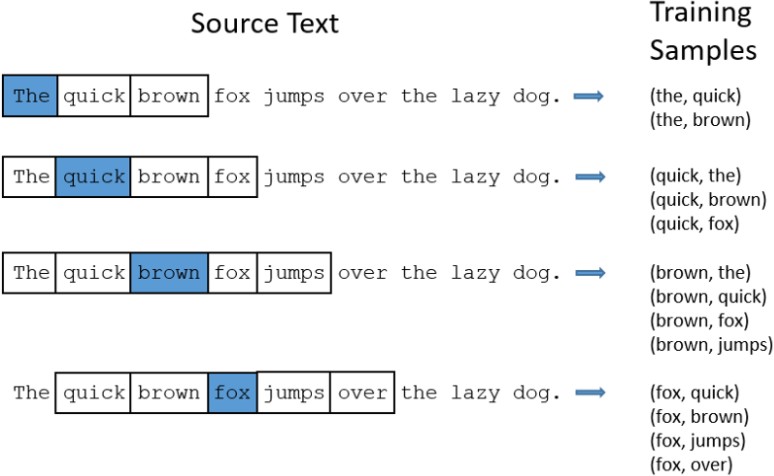
3. For each and every expression in the statement

1.Take the present word as setting.

3. Other arguments in the window to the leftward and right of the word as targets and as a pair .

4. The position of the context and target are determined using the pre-defined vocabulary in the tensorflow library, and those standards are then used in this method .

5. The o/p is sent to the sigmoidal function, yielding the range [-1,1].





### EVALUATION MEASURES:

Every time we create a machine learning model, some kind of statistic is required to assess the model's efficacy. Remember that there are other ways to define the "goodness" of a model, but in a machine learning environment, we typically refer to the model's performance on new instances that weren't included in the training data.

There are two crucial considerations that determine if a model is effective for a certain task:

1 The suitability of the evaluation metric we chose for our issue.

If we are carrying out the evaluation procedure properly.

### TYPES OF METRICS FOR EVALUATION

The assessment metrici we>use is determined on the sort of NLP work we are performing. Furthermore, the project's stage influences the assessment metric we choose. For example, throughout the model development and deployment phases, we would almost always utilise a different assessment measure than when the modeli is in creation. In the first two cases, ML measurements would sufficient; but, in construction, we are concerned with business effect, hence we would want to utilise business metrics to assess the quality of our model.

### METRICES

Some basic metrics used for the calculation are given below:

### ACCURACY

When we use the metric for accuracy, we want to see how near a calculated value is to a known value.

### PRECISION

When the accuracy of the model's predictions is an issue, precision would be employed. The accuracy metric would tell us how many classifications are actually positive vs how many times the classifier classified something good.

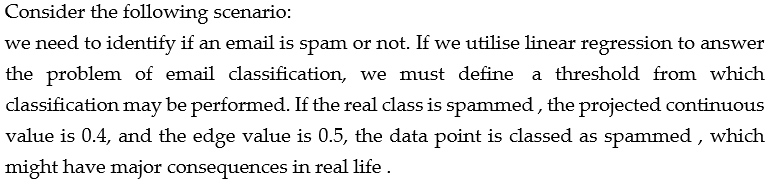
### RECALL

The model's recall of the positive class is measured by recall.

# CHAPTER 6: Model Used

## Logistic Regression

Logistic regression is a method of**modelingi the likelihood of a output achieved from the sigmoid function.**

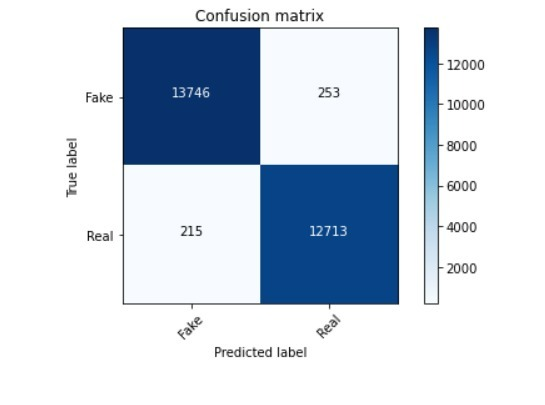


Logistic Regression in Binary

For the categorical response, only two outcomes are possible. Example: Is itnon - spam or

spam?

## 

****

## Fig. Confusion Matrix for Logistic Regression

## Decision Tree

* Decision trees are a supervised. learning method that can be functional to classification and regression complications, but classification problems are where they are most frequently used. It is a tree-structured classifier where each leaf node indicates the outcome and interior nodes contain dataset properties and decision rule branches.
* • The Decision and the Leaf Node makes up a decision tree. While leaf are the outcomes of those previous steps and have no further branches, Decision making nodes are utilised to make decisions and have numerous other decision branches.
* The conclusions or testsi are based on the features of the dataset that has been presented.

.

Fig. Decision tree

## Information Gain: gained Information is used to govern which characteristics offer the most insight into the entropy-based categorization.

## Gini Index: The Gini index is calculated by subtracting the total likelihoods of each class from one.

## 

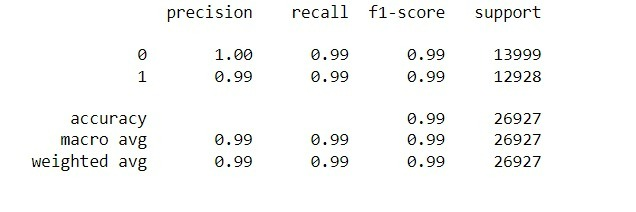
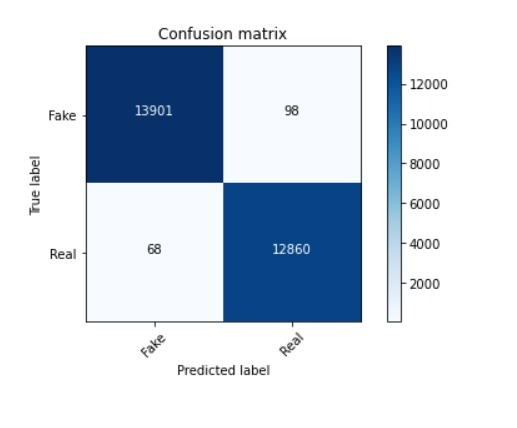


Fig. Results for Decision Tree Classificatio**n**



## Fig. Confusion Matrix for Decision Tree

## Gradient boosting Classifier

An algorithm called a gradient boosting classifier is used when the target column is binary.

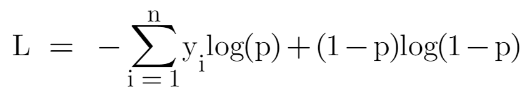
A sort of machine learning approach called gradient boosting classifiers combines many weak learning models to produce a potent prediction model.

A prediction model is produced by gradient boosting from a collection of weak prediction models, frequently decision trees. Gradient boosted trees, the solution used when a decision tree is not a good learner, outperforms random forest in most cases. The model is constructed in the same manner as with previous boosting approaches, but its application is expanded since it permits optimization of any differentiable loss function.

The Gradient Boosting Classifier employs a loss function. Both custom and different standardised loss functions can be used by gradient boosting classifiers, however the loss function must be differentiable.

A prominent technique in classification algorithms is logarithmic loss.

The classification problem's loss function is shown below.:



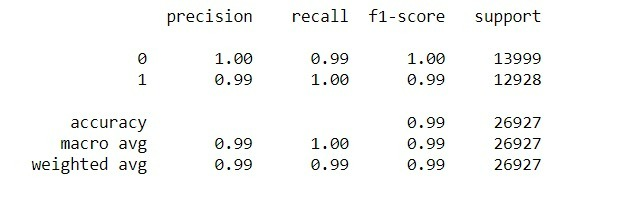
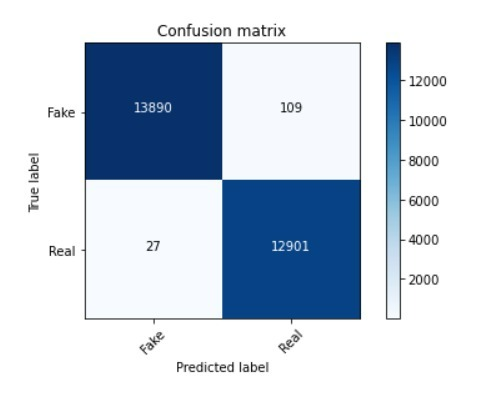


Fig. Results for Gradient Boosting Classifier

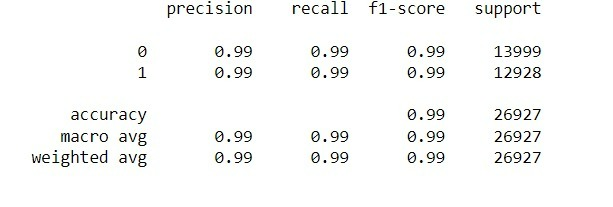


## Fig. Confusion Matrix for Gradient boosting classifiers

## Random Decision Forest

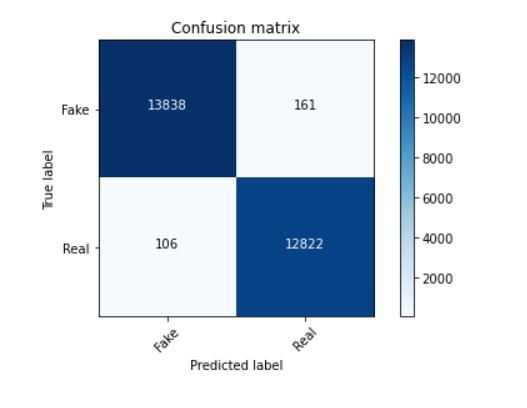
In order to increase the projected accuracy of a dataset, the classifier Random Forest uses many decision trees on various subsets of that dataset and averages their findings.

## To create a more precise and reliable forecast, random forest combines numerous decision trees. Being adaptable to both classificationjand regression issues, which make up the lmajority of modern machine jlearning systems, is a key advantage of random forest.



29

Fig. Results for Random Forest Classifier



# CHAPTER 7: EXPERIMENT ANALY

# Fig. Confusion Matrix for Random Forest Classifier

# CHAPTER 7 Expeimental Setup

#### Experimental Setup

We have used the 'WindowslO' operating system for our experimental setup.

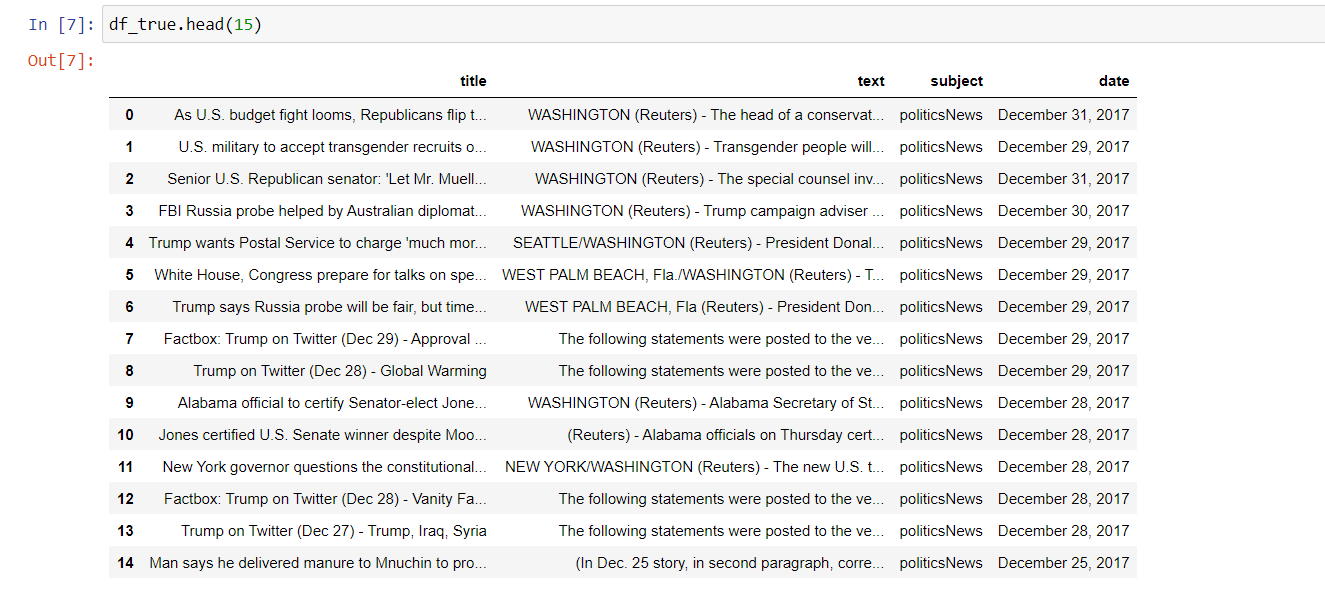
The specifications are as follows:

Processor: Intel® Core TM i5-9300H CPU @ 2.40GHz Installed memory (RAM): 8.0 GB

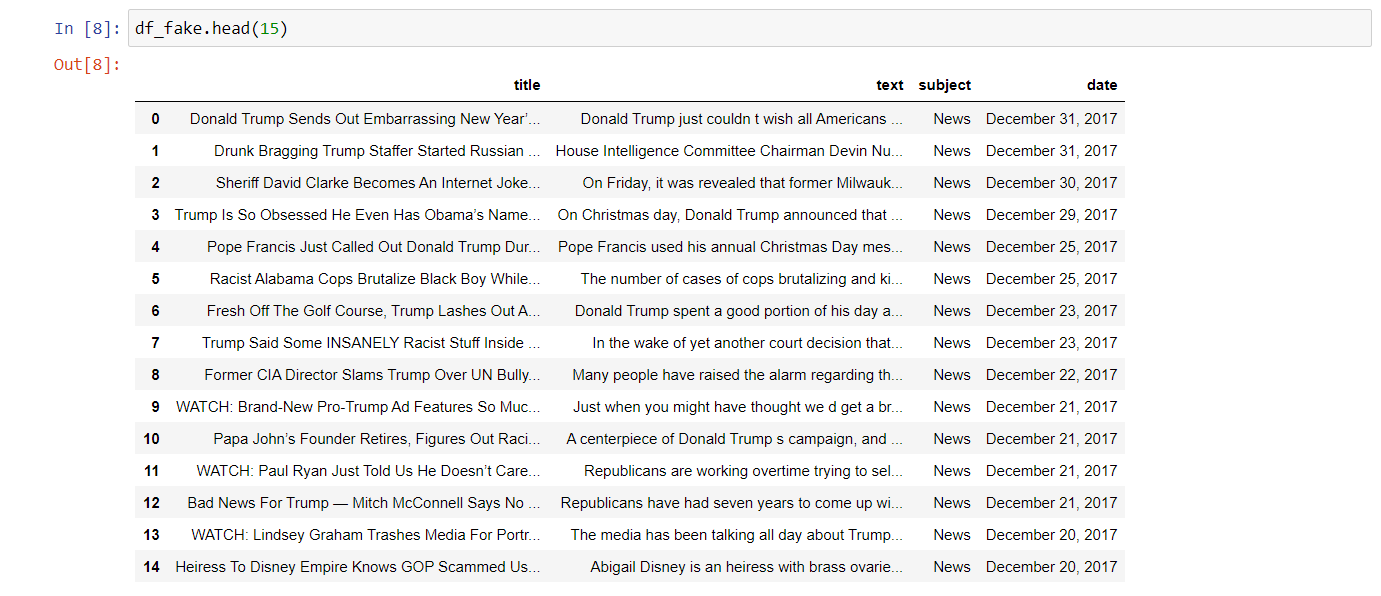
For training of neural network models we used Google Colab and Kaggle Notebook service from Google Research.

Both services provided a much more powerful system specification to work upon without buying them.

## True.csv

****

**Fake.csv**

****

# CHAPTER 8

## USER INPUT INTERFACE



Enter the article

Fig. User interface

## OUTPUT

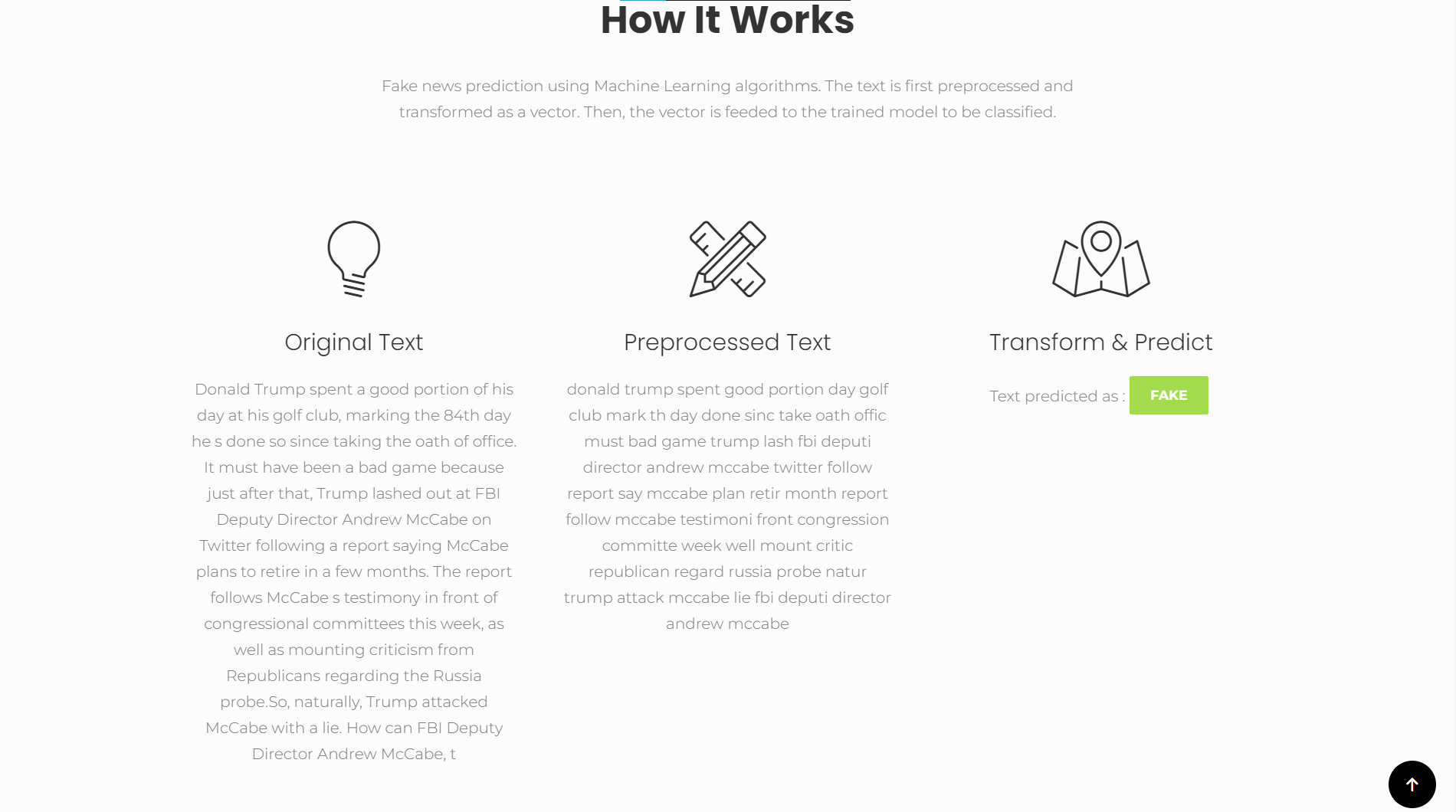


Fig. Output

**CHAPTER 9 CONCLUSION AND FUTURE WORK**



## CONCLUSION:

With the use of Machine Learning algorithms we are able to classify the given news is fake or true . At present the code implements work for satisfied results.



## FUTURE WORK:

These can be pursued to further this project which might lead to new possibilities:

* Application could be turned into an android/ios application so that it covers up more market and is much easier to use.
* We are going to convert this product into multiple languages.

## REFERENCES

1. International journal of recent technology and engineering (IJRTE) ISSN: 2277-3878, volume-7, issue-6, march 2019
2. Building a fake news classifier using natural language processing BY NATHAN ([https://towardsdatascience.com/building-a-fake-news-classifier-using-natural-](https://towardsdatascience.com/building-a-fake-news-classifier-using-natural-language-processing-83d911b237e1) [language-processing-83d911b237e1](https://towardsdatascience.com/building-a-fake-news-classifier-using-natural-language-processing-83d911b237e1))
3. Fake news detector: NLP project by ishant juyal (<https://levelup.gitconnected.com/fake-news-detector-nlp-project-9d67e0177075>)
4. Shloka Gilda,“Evaluating Machine Learning Algorithms for Fake News Detection”

,2017 IEEE 15th Student Conference on Research and Development (SCOReD).

1. Mykhailo Granik, Volodymyr Mesyura, “Fake News Detection Using Naive Bayes Classifier”, 2017 IEEEFirst Ukraine Conference on Electrical and Computer Engineering (UKRCON).
2. Gravanis, G., et al., Behind the cues: A benchmarking study for fake news detection. Expert Systems with Applications, 2019. 128: p. 201- 213.
3. Bondielli, A. and F. Marcelloni, A survey on fake news and rumour detection techniques. Information Sciences, 2019. 497: p. 38-55.
4. Ko, H., et al., Human-machine interaction: A case study on fake news detection using a backtracking based on a cognitive system. Cognitive Systems Research, 2019. 55: p. 77- 81.
5. Zhang, X. and A.A. Ghorbani, An overview of online fake news: Characterization, detection, and discussion. Information Processing & Management, 2019.
6. Robbins, K.R., W. Zhang, and J.K. Bertrand, The ant colony algorithm for feature selection in high-dimension gene expression data for disease classification. Journal of Mathematical Medicine and Biology, 2008
7. Alirezaei, M., S.T.A. Niaki, and S.A.A. Niaki, A bi-objective hybrid optimization algorithm to reduce noise and data dimension in diabetes diagnosis using support vector machines. Expert Systems with Applications, 2019. 127: p. 47-57.
8. Zakeri, A. and A. Hokmabadi, Efficient feature selection method using real-valued grasshopper optimization algorithm. Expert Systems with Applications, 2019. 119: p. 61-72.