**FAKE NEWS DETECTION**

**A Project Report**

*Submitted in the partial fulfilment for the award of the degree of*

**BACHELOR OF ENGINEERING**

**IN**

**COMPUTER SCIENCE WITH SPECIALIZATION IN**

**ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING**

**Submitted by:**

**ARYAN CHAUDHARY**

**(20BCS6620)**

**MOHIT GUPTA**

**(20BCS6574)**

**CHIRAG KALRA**

**(20BCS6601)**

**KARTIKEY**

**(20BCS6693)**

**Under the Supervision of:**

**SIDDHARTH KUMAR**



**CHANDIGARH UNIVERSITY, GHARUAN, MOHALI - 140413,**

**PUNJAB**

**February, 2023**

# Abstract

Fake news detection system using machine learning would describe a system designed to automatically identify and classify news articles as real or fake using machine learning techniques. The system would be trained on a large dataset of news articles that have been labelled as either real or fake.

The system would use a variety of machine learning algorithms, such as supervised learning, unsupervised learning, and deep learning, to analyse the text of the news articles and identify patterns and features that distinguish real news from fake news. The system would also incorporate other features, such as the source of the news article, the date and time of publication, and the social media activity around the article, to improve its accuracy

Keywords:

Online fake news, Machine learning, fake news, Text Classification, social media, NLP, Classification.

# Table of Contents

|  |  |
| --- | --- |
| Title Page | i |
| Abstract | ii |
| 1.    Introduction |  |
| 1.1   Problem Definition |  |
| 1.2 Project Overview |  |
| 1.3 Hardware Specification |  |
| 1.4 Software Specification |  |
| 2.    Literature Survey |  |
| 2.1 Existing System |  |
| 2.2 Proposed System |  |
| 2.3 Literature Review Summary |  |
| 3.    Problem Formulation |  |
| 4.    Research Objective |  |
| 5.    Methodologies |  |
|  |  |
| 6.    Conclusion |  |
| 7.    Tentative Chapter Plan for the proposed work |  |
| 8.    Reference |  |

# INTRODUCTION

World is changing fleetly. No doubt we've a number of advantages of this digital world but it also has its disadvantages as well. There are different issues in this digital world. One of them is fake news. Someone can fluently spread a fake news. Fake news is spread to detriment the character of a person or an association. It can be a propaganda against someone that can be a political party or an association. There are different online platforms where the person can spread the fake news. This includes the Facebook, Twitter etc. Machine learning is the part of artificial intelligence that helps in making the systems that can learn and perform different conduct( Donepudi, 2019). A variety of machine learning algorithms are available that include the supervised, unsupervised, underpinning machine learning algorithms. The algorithms first have to be trained with a data set called train data set. After the training, these algorithms can be used to perform different tasks. Machine learning is using in different sectors to perform different tasks. utmost of the time machine learning algorithms are used for vaticination purpose or to descry commodity that's hidden. Online platforms are helpful for the druggies because they can fluently pierce a news. But the problem is this gives the occasion to the cyber culprits to spread a fake news through these platforms. This news can be proved dangerous to a person or society. compendiums read the news and start believing it without its verification. Detecting the fake news is a big challenge because it isn't an easy task (Shu etal., 2017). If the fake news isn't detected beforehand also the people can spread it to others and all the people will start believing it. individualities, associations or political parties can be effected through the fake news. People opinions and their opinions are affected by the fake news in the US election of 2016( Dewey, 2016). Different experimenters are working for the discovery of fake news. The use of Machine learning is proving helpful in this regard. Experimenters are using different algorithms to descry the false news. Experimenters in( Wang, 2017) said that fake news discovery is big challenge. They've used the machine learning for detecting fake news. Experimenters of( Zhou etal., 2019) set up that the fake news are adding with the passage of time. That's why there's a need to descry fake news. The algorithms of machine learning are trained to fulfill this purpose. Machine learning algorithms will descry the fake news automatically formerly they've trained.

## Problem Definition

Fake news detection is a challenging problem that can be addressed using machine learning. Researchers have proposed various approaches to detect fake news based on the words, phrases, sources, and titles of articles. The problem can be framed as a binary classification task where the model predicts whether a piece of news is credible or fake. The LIAR dataset has been used in some studies to train and evaluate machine learning models for fake news detection.

## Problem Overview

Fake news detection is a significant problem that can be addressed using machine learning. Machine learning algorithms can automatically detect fake news once they are trained. Researchers have proposed various methodologies to create models that detect fake news based on the words, phrases, sources, and titles of articles. Deep learning-based techniques have been found to be more accurate in detecting fake news compared to many machine learning techniques. Humans are inconsistent and poor detectors of fake news, as demonstrated by the widespread effects of the large onset of fake news.

The problem of fake news detection involves identifying and classifying news articles as either real or fake. The proliferation of social media and online news platforms has made it easier for people to access news, but it has also made it easier for false or misleading information to be disseminated. Fake news can have serious consequences, including social and political unrest, financial losses, and damage to the reputation of individuals and organizations.

Machine learning techniques can be used to automatically detect fake news by analyzing the content and features of news articles. Some of the challenges associated with this problem include:

1. Lack of labeled data: Collecting a large and diverse dataset of labeled news articles can be challenging, particularly for more recent events.

2. Limited access to features: News articles may contain a wide range of features that could be indicative of their veracity, including linguistic patterns, author reputation, and social media sharing. However, access to these features may be limited due to privacy concerns or technical limitations.

3. Complex relationships between features: The features of news articles may interact with each other in complex ways, making it difficult to identify which features are most important for detecting fake news.

4. Adversarial attacks: Malicious actors may attempt to game the system by intentionally creating news articles that are difficult for machine learning algorithms to classify.

To address these challenges, machine learning models can be trained on large and diverse datasets using techniques such as natural language processing, feature engineering, and deep learning. The models can then be validated and optimized to achieve high accuracy in detecting fake news while minimizing false positives and false negatives.

## Hardware Specification

RAM: 8GB or higher

Processor: i3 8th Gen or higher

## Software Specification

Jupyter notebook

Python 3.5 or higher

# 2. LITERATURE SURVEY

## 2.1 Existing System

1. Language Processing for Fake News Detection" by Ray Oshikawa, Jing Qian, and William Yang Wang provides an in-depth review of the use of natural language processing (NLP) techniques for detecting fake news.

The authors provide a detailed overview of the various datasets and evaluation metrics that have been used for evaluating the performance of NLP-based fake news detection systems. They also discuss the challenges associated with building such systems, such as the lack of annotated data and the dynamFake news detection is a challenging problem that can be addressed using machine learning. Researchers have proposed various approaches to detect fake news based on the words, phrases, sources, and titles of articles. The problem can be framed as a binary classification task where the model predicts whether a piece of news is credible or fake. The LIAR dataset has been used in some studies to train and evaluate machine learning models for fake news detection.

1. The paper "Detecting Fake News in Social Media Networks" by Ali Alwahedi and Monther Aldwairi proposes a machine learning-based approach for detecting fake news in social media networks.

The authors compare several machine learning algorithms, including logistic regression, decision trees, and support vector machines, and evaluate their performance using various metrics such as accuracy, precision, recall, and F1 score. They also compare their approach with other state-of-the-art methods for detecting fake news.

The results show that the proposed system achieves high accuracy in detecting fake news articles, with an accuracy of up to 98%. The authors also find that using a combination of features improves the performance of the system.

1. The paper "A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities" by Xinyi Zhou and Reza Zafarani provides a comprehensive overview of the problem of fake news, its impact, and various methods for detecting and combating it.The paper concludes by emphasizing the importance of addressing the problem of fake news through a multidisciplinary approach that combines technical, social, and ethical considerations. The authors highlight the need for collaboration between researchers, journalists, policymakers, and other stakeholders to develop effective solutions for detecting and combating fake news.

## 2.2 Proposed System

A proposed system for fake news detection could use logistic regression, decision tree classifier, and random forest classifier machine learning algorithms. The system would be built using a labeled dataset of news articles, where each article is labeled as either "real" or "fake." The dataset would be split into training and testing sets to evaluate the performance of the models.

The logistic regression model would be used to model the probability of an article being fake or real based on features extracted from the text, such as the frequency of certain words or the presence of certain phrases. The model would output a probability score for each article, which could be threshold to make a binary classification decision.

The decision tree classifier would be used to build a tree-like model that uses a series of binary decisions based on features in the text to classify articles as fake or real. The model would recursively split the dataset based on the most informative features, creating decision nodes that branch off into child nodes until a final classification is made.

Finally, the random forest classifier would be used to build an ensemble of decision trees, where each tree is trained on a random subset of the features and examples. The model would use a voting mechanism to combine the outputs of the individual decision trees, resulting in a final classification decision.

The proposed system could be evaluated using standard metrics such as precision, recall, and F1 score on the testing dataset, and the performance of the different models could be compared to determine which one is most effective at detecting fake news. Overall, this system would provide a robust approach to detecting fake news and could be further improved by incorporating additional features or algorithms

## 2.3 Literature Review Summary

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Year and**  **Citation** | **Article/ Author** | **Tools/ Software** | **Technique** | **Source** | **Evaluation Parameter** |
| EUSPN 2018 | Detecting Fake News in Social Media Networks Ali Alwahedi, Monther Aldwairi | Jupyter Notebook | K-Means  Clustering,  Logistic Regression | [Detecting Fake News in Social Media Networks | Semantic Scholar](https://www.semanticscholar.org/paper/Detecting-Fake-News-in-Social-Media-Networks-Aldwairi-Alwahedi/7575e122e51390069b540615a7ff0ac938960cea) | Confusion Matrix: Precision, Recall, F-Measure, ROC |
| [ACM Computing Surveys](https://dl.acm.org/toc/csur/2021/53/5)  2020 | A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities Xinyi Zhou, Reza Zafarani | Jupyter Notebook | NLP, Data mining | [Detecting Fake News – The Data Exchange](https://thedataexchange.media/detecting-fake-news/) | Authenticity, Intention |
| IRJET,2020 | Fake News Detection Using ML Srishti Agarwal, Vaishali Arora | Jupyter Notebook | Natural Language Processing | [Fake News Detection Using ML: Srishti Agrawal, Vaishali Arora, Ruchika Arora, Pronika Chawla, Madhumita Kathuria | PDF | Accuracy And Precision | Statistical Classification (scribd.com)](https://www.scribd.com/document/502399881/IRJET-V7I51102) | Confusion Matrix: Precision, Recall, F-Measure, ROC |
| [ACM Computing Surveys](https://dl.acm.org/toc/csur/2021/53/5),  2017 | CSI: A Hybrid Deep Model for Fake News Detection,Natali Ruchansky, Sungyong Seo and Yan Liu, | Jupyter Notebook | Recurrent Neural Network | [CSI: A Hybrid Deep Model for Fake News Detection | Request PDF (researchgate.net)](https://www.researchgate.net/publication/320885186_CSI_A_Hybrid_Deep_Model_for_Fake_News_Detection) | Accuracy |
| Tokyo,Japan | A Survey on Natural Language Processing for Fake News Detection,Ray Oshikawa, Jing Qian, William Yang Wang, | Jupyter Notebook | NLP, ML | [(PDF) A Survey on Natural Language Processing for Fake News Detection (researchgate.net)](https://www.researchgate.net/publication/328736618_A_Survey_on_Natural_Language_Processing_for_Fake_News_Detection) | Stance Detection |
| Bangladesh University,2019 | A Benchmark Study on Machine Learning Methods for Fake News Detection, Junaed Yonus Khan, Md. Tawkat Islam Khondaker, Anindya Iqbal, Sadia Afroz, | Jupyter Notebook | RNN, CNN | [(PDF) A Benchmark Study on Machine Learning Methods for Fake News Detection (researchgate.net)](https://www.researchgate.net/publication/333077208_A_Benchmark_Study_on_Machine_Learning_Methods_for_Fake_News_Detection) | Performance Matrices |
| TEST,2019 | Fake News Detection on Social Media-A  Review  Steni Mol T S\*, P S Sreeja | Jupyter Notebook | RNN, Naïve Bayes,SVM,Kmeans | [(PDF) Fake News Detection on Social Media-A Review (researchgate.net)](https://www.researchgate.net/publication/342051943_Fake_News_Detection_on_Social_Media-A_Review) | Evaluation on false news unmasking on  social media describe and distinguished various  datasets such as LIAR, PHEME,Fake News Net,  BuzzFeed News[14,35,31] datasets. |

# 3. PROBLEM FORMULATION

The problem formulation for a machine learning-based fake news detection system can be framed as a binary classification task where the system is trained to distinguish between real and fake news articles. The input to the system is a textual document representing a news article, and the output is a binary label indicating whether the article is real or fake.

# 4. OBJECTIVES

The objective of a machine learning-based fake news detection system is to automatically identify and classify news articles as real or fake using machine learning algorithms. The system aims to assist journalists, social media platforms, and news consumers to quickly and accurately identify fake news articles, which can have significant negative impacts on society.

# 5. METHODOLOGY

1. Logistic Regression

Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

Logistic regression is a type of statistical analysis used to predict the probability of a binary outcome (e.g. yes/no, true/false, 0/1) based on one or more predictor variables. It is commonly used in machine learning and data analysis for classification tasks, such as predicting whether a customer will buy a product or not based on their demographic information.

The logistic regression model uses a logistic function to transform the output of a linear regression model into a probability value between 0 and 1. The logistic function is an S-shaped curve that maps any real-valued input to a probability value between 0 and 1.

Logistic regression is a parametric model, meaning that it makes assumptions about the distribution of the input data, and it estimates the parameters of the model based on those assumptions. The model can be trained using a variety of optimization algorithms, such as gradient descent, to find the optimal values of the model parameters that minimize the error between the predicted probabilities and the actual binary outcomes.

Overall, logistic regression is a powerful and widely used tool for binary classification tasks, especially when the relationship between the predictor variables and the outcome is non-linear.

II Decision Tree

Decision Tree is a Supervised learning techniquethat can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, whereinternal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

A decision tree is a type of machine learning algorithm used for both classification and regression tasks. It is a tree-like model where each internal node represents a feature or attribute, and each leaf node represents a class label or a numerical value.

The decision tree algorithm recursively splits the dataset into smaller subsets based on the most significant features or attributes, and this process continues until the subset at each node contains only one class label or a single value. The splitting criterion is based on the concept of information gain, which measures the reduction in entropy or uncertainty of the target variable after splitting the dataset based on a particular feature.

Decision trees are easy to understand and interpret, and they can handle both categorical and numerical data. They can also handle missing values and outliers. However, decision trees are prone to overfitting, which means that the model may fit the training data too closely and perform poorly on new data.

To prevent overfitting, techniques such as pruning, setting minimum sample size at leaf nodes, and limiting the depth of the tree can be used. Additionally, ensemble methods like Random Forest and Gradient Boosting can be used to improve the accuracy of the decision tree by combining multiple trees.

Overall, decision trees are a powerful and popular tool for solving classification and regression problems in machine learning and data mining.

III Random Forest

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of **ensemble learning,** which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

Random Forest is a type of ensemble machine learning algorithm that combines multiple decision trees to improve the accuracy of the prediction. Each decision tree in a Random Forest is trained on a random subset of the original dataset, and the final prediction is based on the majority vote of all the individual trees.

The key idea behind Random Forest is to reduce overfitting by introducing randomness into the model. In addition to the random sampling of the data, Random Forest also introduces randomness into the feature selection process by selecting only a subset of features for each tree. This helps to ensure that each tree is different and captures a unique aspect of the data.

Random Forest is a versatile algorithm that can be used for both classification and regression tasks. It is highly effective in handling high-dimensional datasets, noisy data, and missing values. It is also robust to outliers and able to capture complex interactions between the features.

One of the main advantages of Random Forest is that it provides a measure of feature importance, which can help to identify the most relevant features for the prediction. This information can be used to improve the performance of the model by selecting only the most important features or by engineering new features that are more informative.

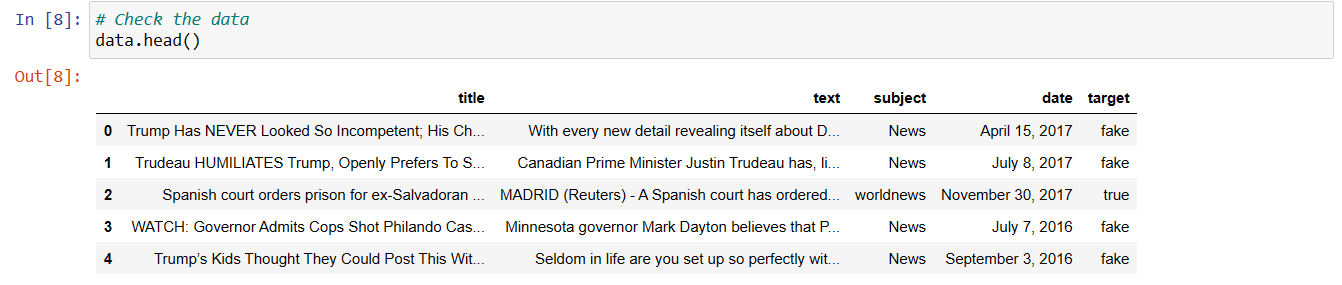
Overall, Random Forest is a powerful and widely used algorithm that can achieve high accuracy in a variety of applications.

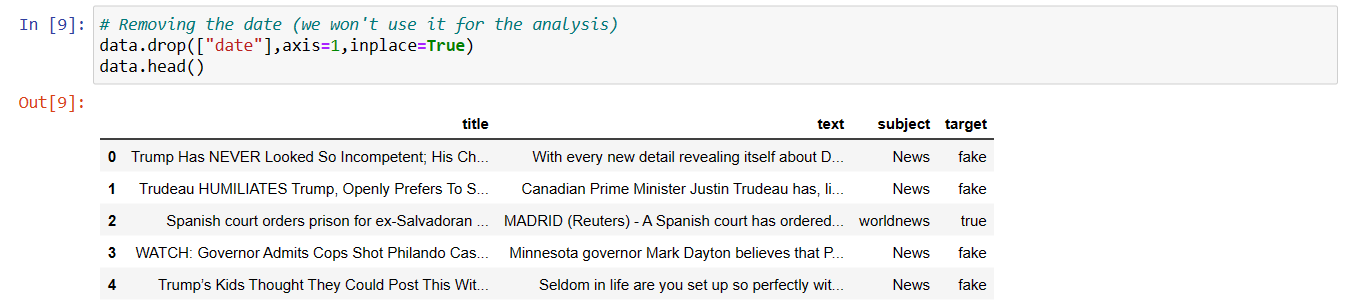
# 6. RESULT & OUTPUT





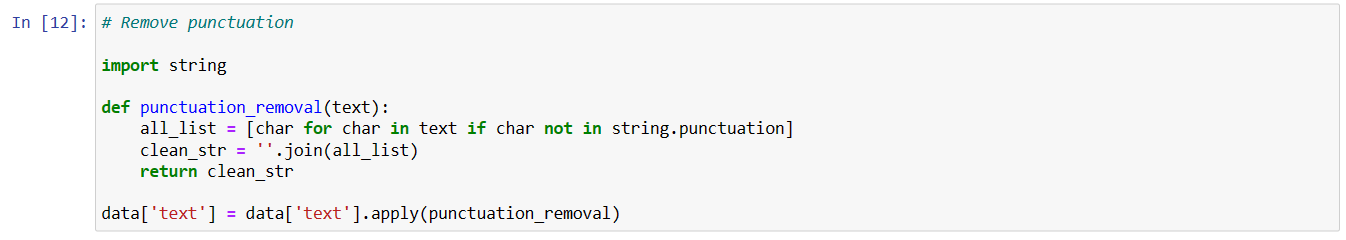




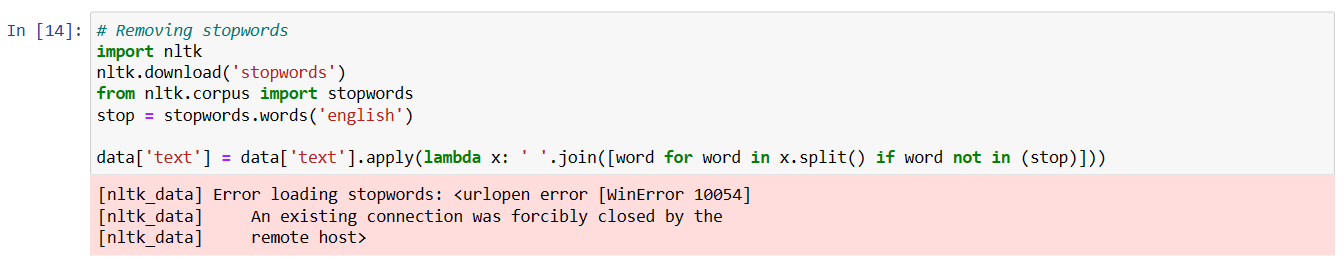


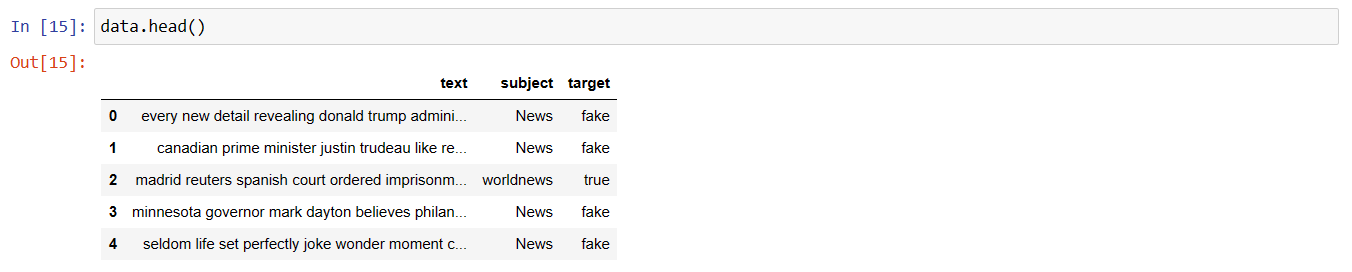


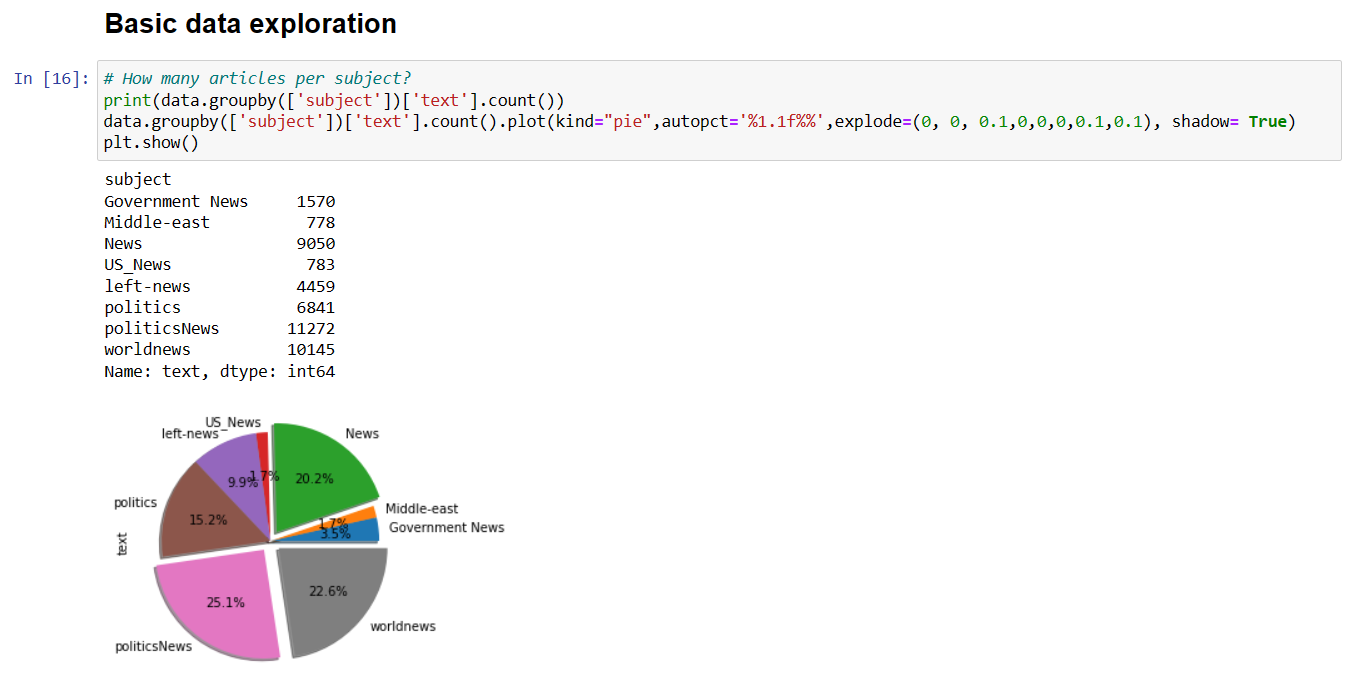


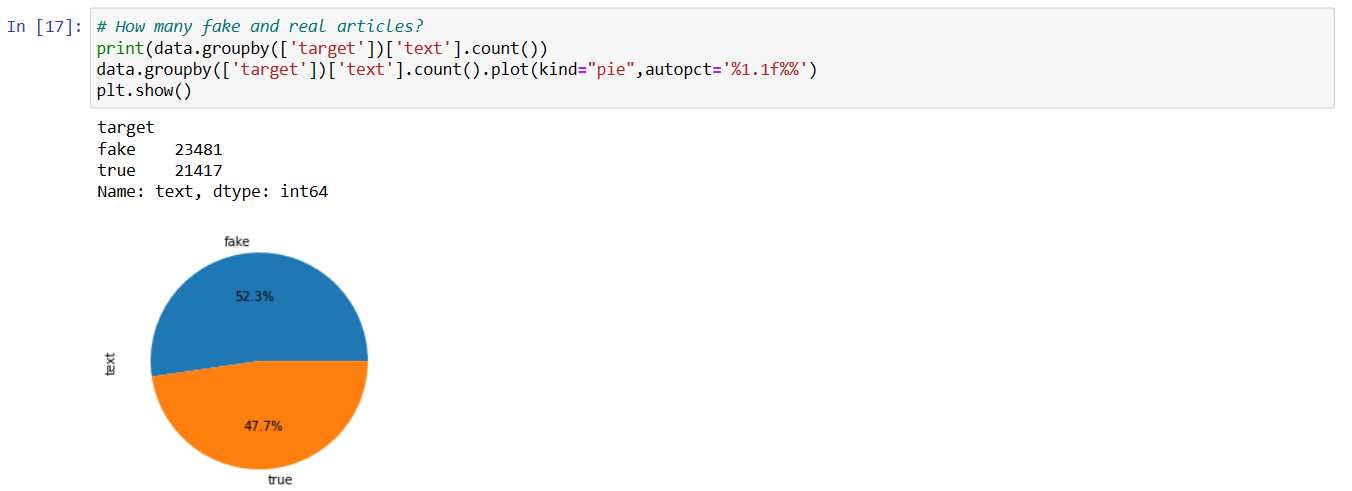


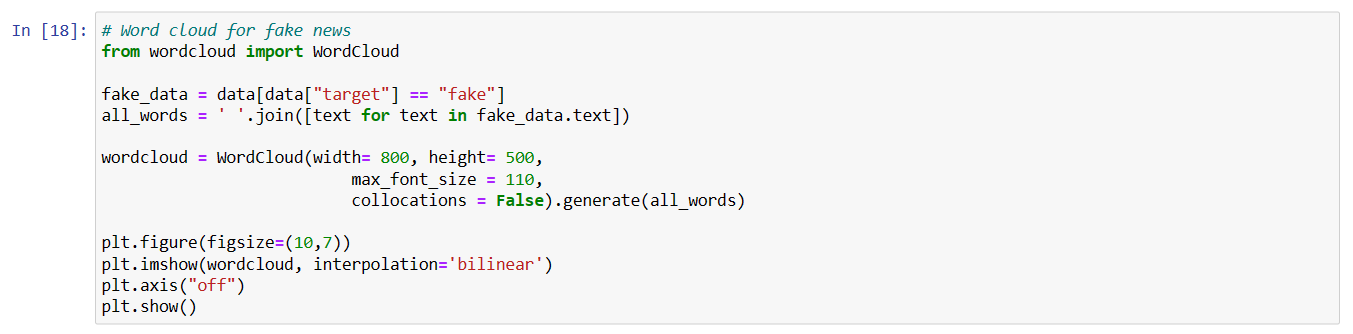




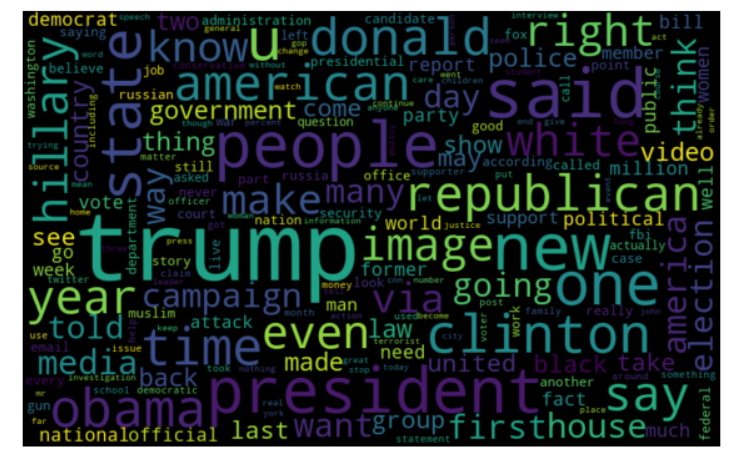


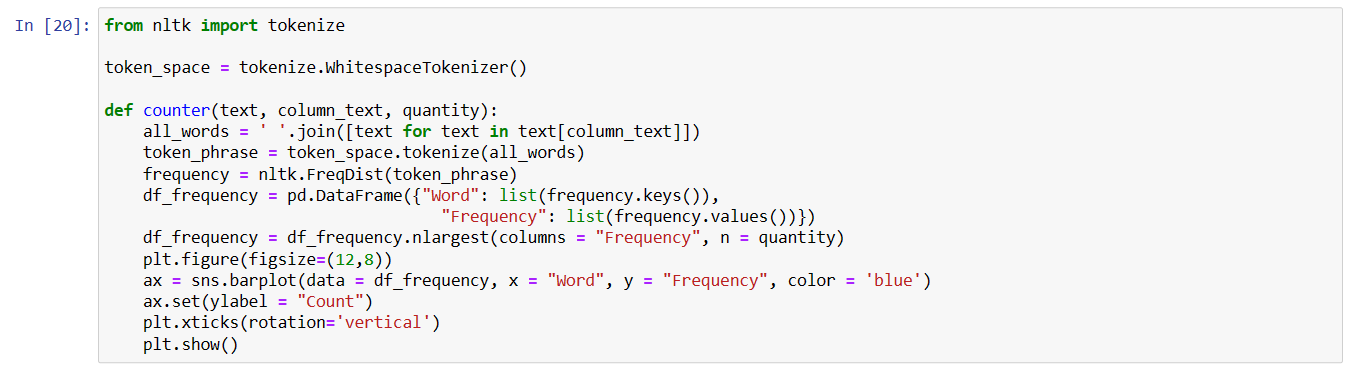


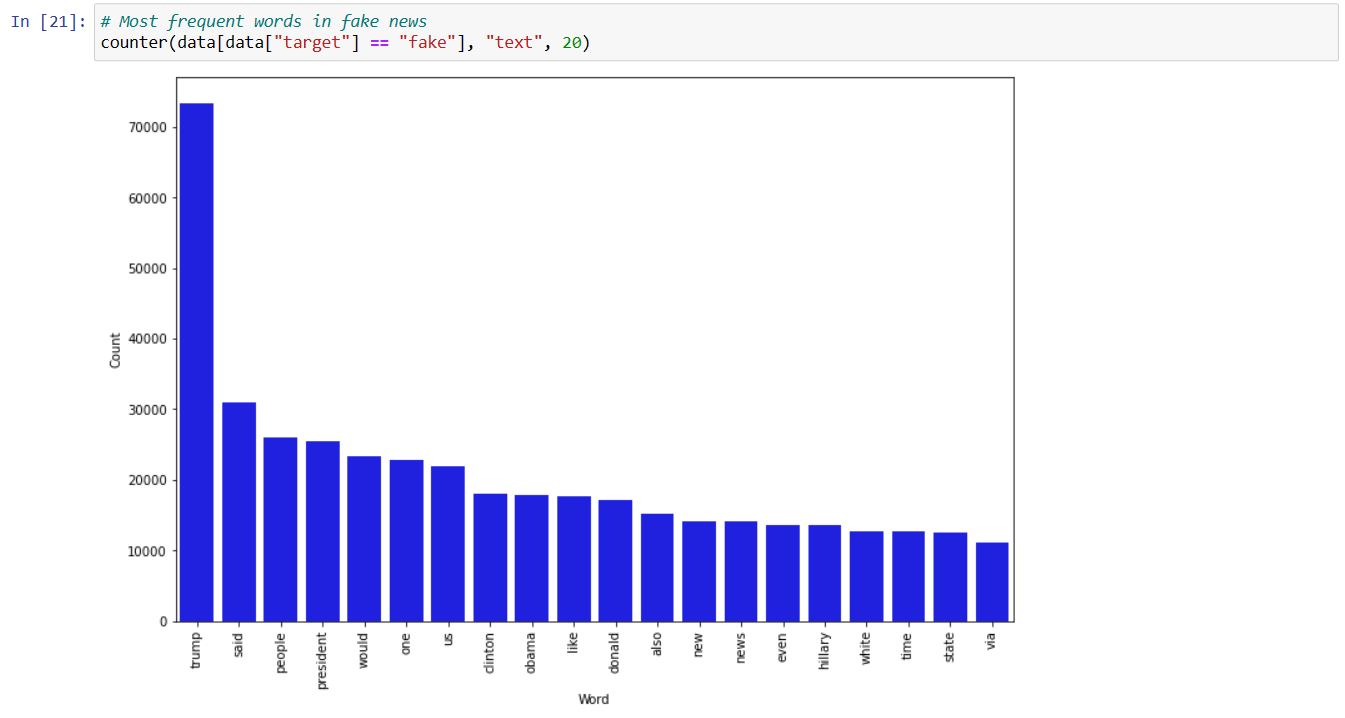


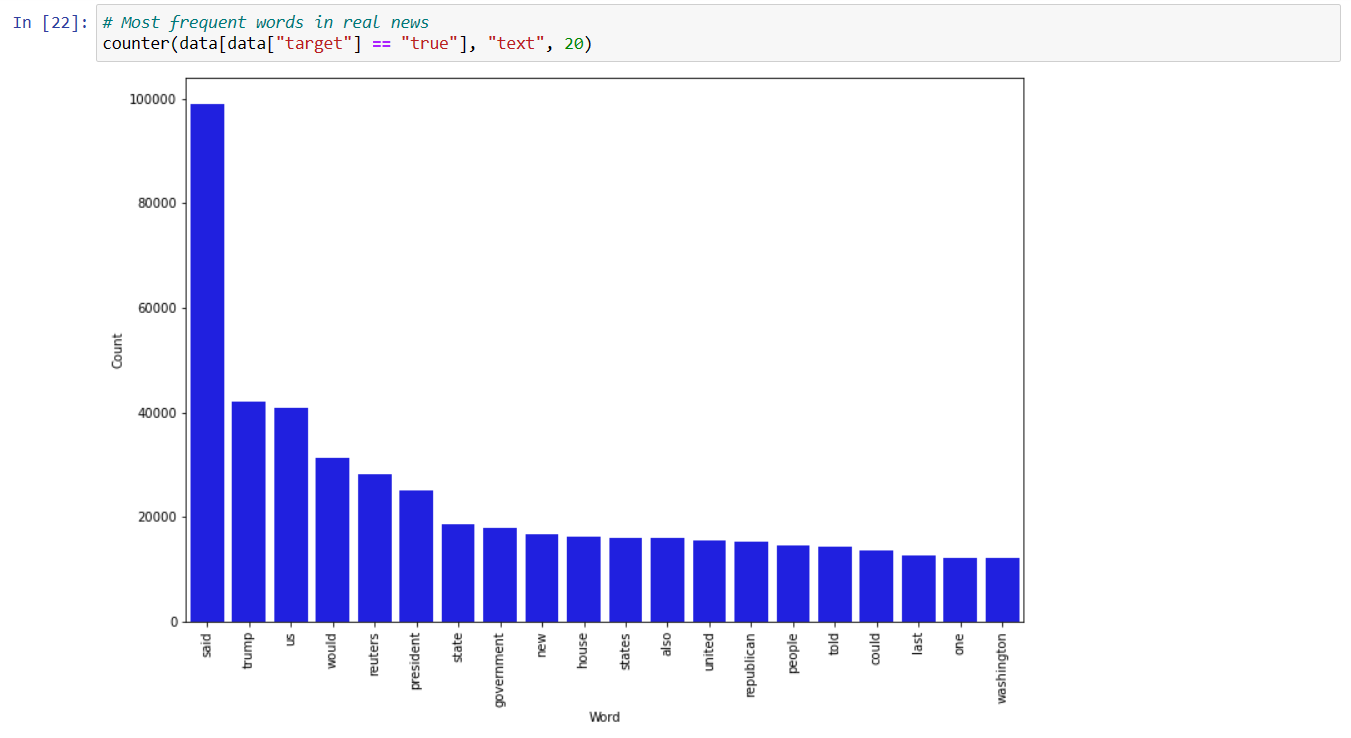




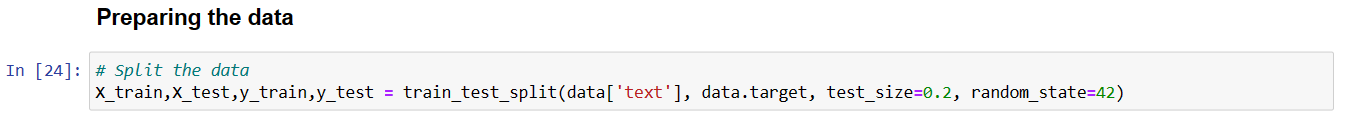


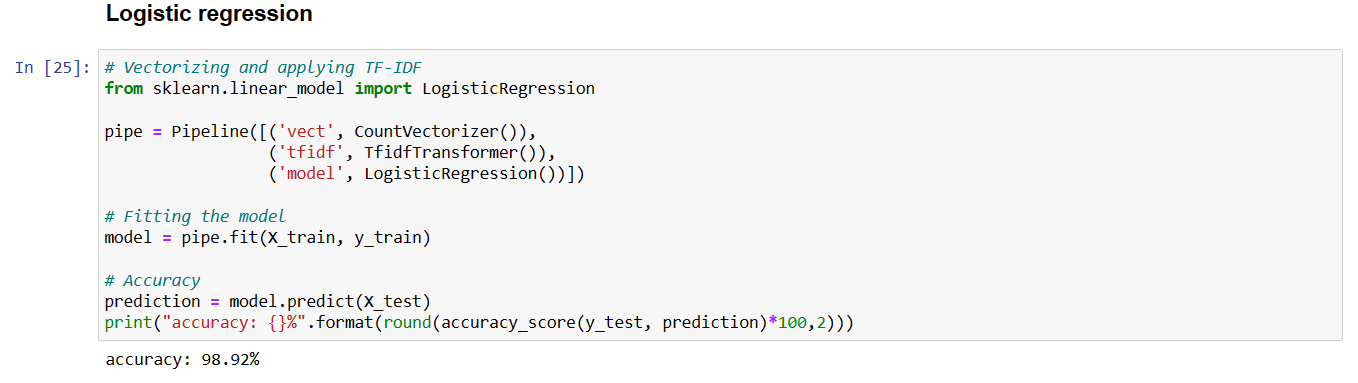


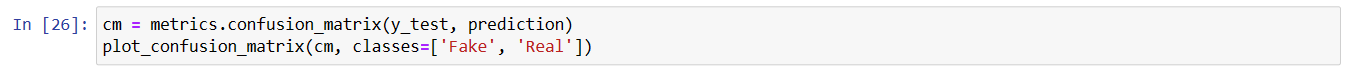


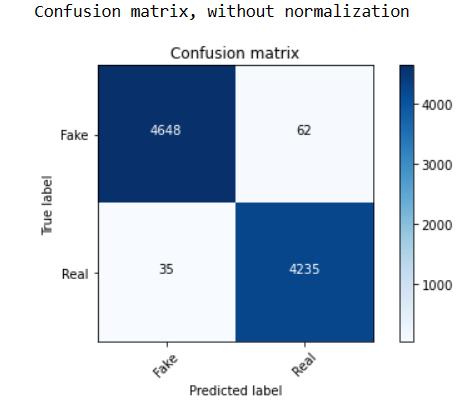


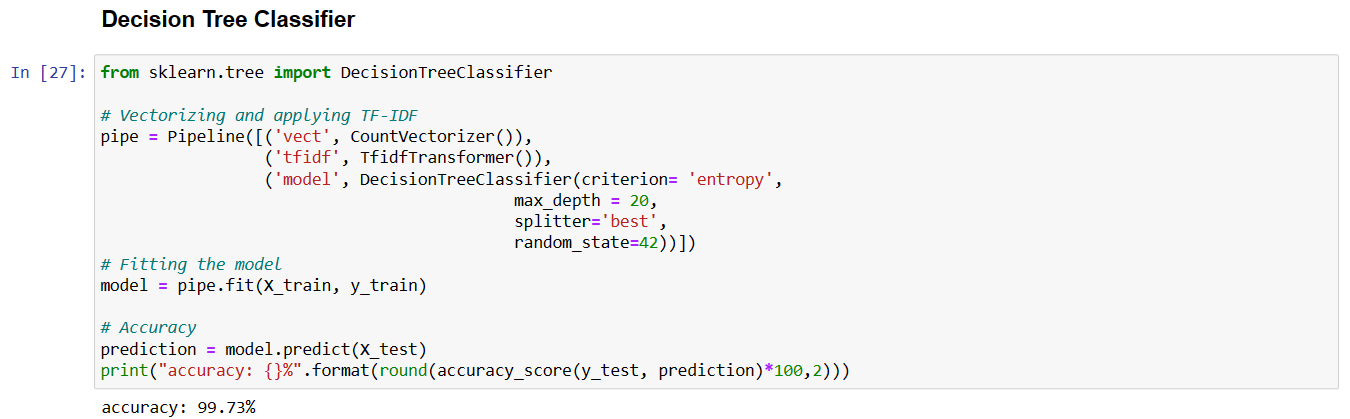


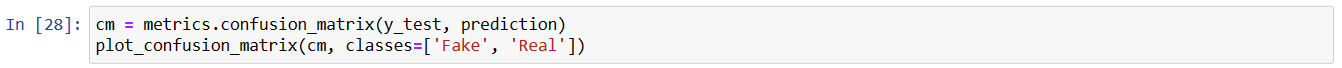


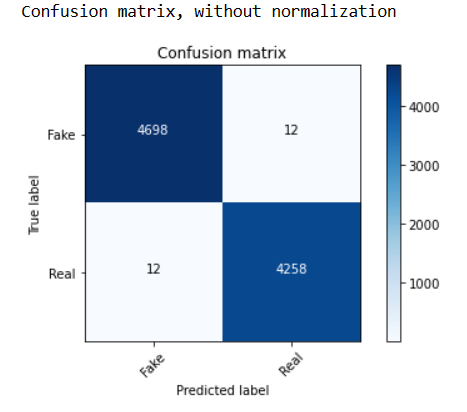


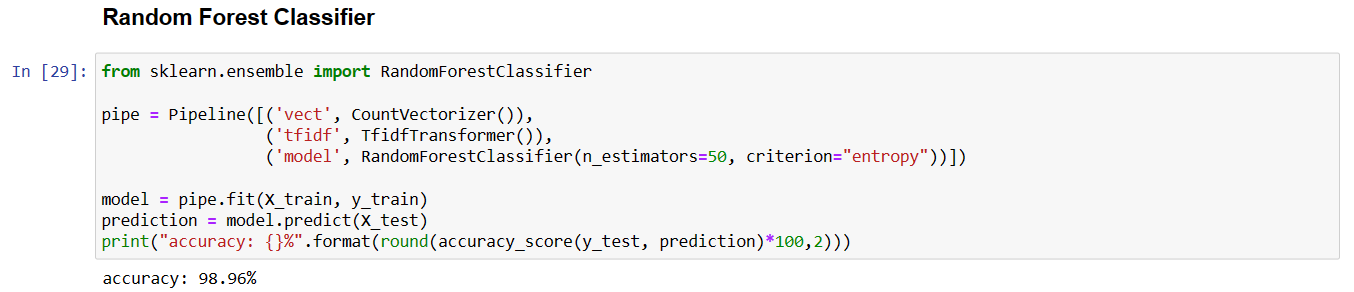


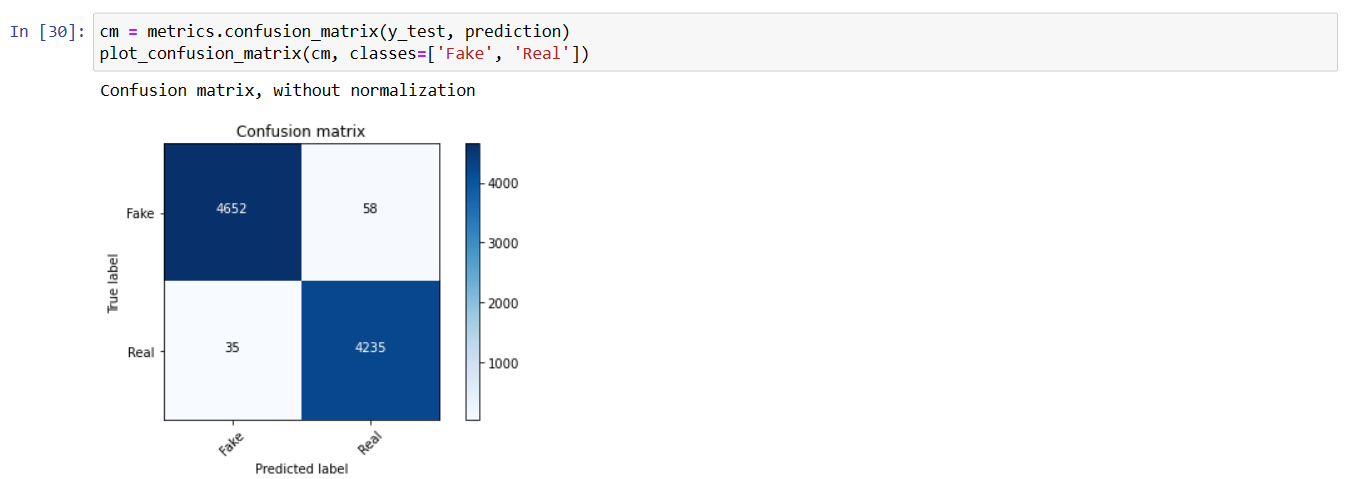












# 7. CONCLUSION

In summary, a machine learning-based fake news detection system can be a valuable tool for combating fake news and promoting truthful and accurate reporting of news articles. However, it is important to recognize that such systems are not perfect and may have limitations and biases. Therefore, it is crucial to use such systems as a complement to human judgment and critical thinking in order to make well-informed decisions about the accuracy and veracity of news articles.

In conclusion, the "Fake News Detection" project is a valuable initiative that aims to tackle the problem of fake news and disinformation in today's society. With the proliferation of social media and the internet, fake news has become a significant challenge, and this project provides an effective solution to identify and flag false information.

The project uses Natural Language Processing (NLP) techniques and Machine Learning algorithms to analyze text and identify patterns and characteristics that indicate fake news. It has the potential to improve its accuracy and effectiveness by incorporating advanced NLP techniques, multi-lingual support, and the detection of multimedia content along with the help of ML algorithms.

Moreover, this project has a wide range of future development opportunities, including collaborations with social media platforms, fact-checking organizations, and continuous feedback from users. These collaborations can enhance the accuracy of fake news detection and promote responsible journalism.

In conclusion, the "Fake News Detection" project has significant potential to make a positive impact on society by promoting the dissemination of accurate information and combating the spread of false information.

Here are some key points that can be concluded from the conclusion:

* The "Fake News Detection" project is an important initiative that addresses the challenge of fake news and disinformation in today's society.
* The project uses NLP techniques and ML algorithms to analyze text and identify patterns that indicate fake news.
* The project has the potential to improve its accuracy and effectiveness by incorporating advanced NLP techniques, multi-lingual support, and the detection of multimedia content.
* Collaborations with social media platforms, fact-checking organizations, and feedback from users can enhance the accuracy of fake news detection and promote responsible journalism.
* The project has a significant future scope to make a positive impact on society by promoting the dissemination of accurate information and combating the spread of false informationFND systems are becoming increasingly important as the spread of misinformation and disinformation on the internet continues to be a major concern. Natural language processing (NLP) and machine learning methods can help to automate the process of detecting fake news, which can save time and resources compared to manual fact-checking.

## 8. TENTATIVE CHAPTER PLAN FOR THE PROPOSED WORK

**CHAPTER 1: INTRODUCTION**

Provides a general insight about the project.

**CHAPTER 2: LITERATURE REVIEW**

This section contains the detailed study of the existing system as well as the proposed system over the existing system and the literature survey table giving a summarized outlook of the papers referred.

**CHAPTER 3: PROBLEM FORMULATION**

How we are solving the problem.

**CHAPTER 4: OBJECTIVE**

Analyse the major objectives of the project.

**CHAPTER 5: METHODOLOGIES**

Provides a brief description of the steps in the project, to enlighten about the dynamics and workflow of the proposed work

**CHAPTER 6: RESULT & OUTPUT**

Glimpse of our project

**CHAPTER 7: CONCLUSION**

It concludes the synopsis.

## 

## REFERENCES

[1]. Jiawei Zhang, Bowen Dong, Philip S. Yu, “FAKE DETECTOR: Effective Fake News Detection with Deep Diffusive Neural Network” published in August 2019.

[2]. Steni Mol T S and Shreeja P Sin, “Fake News Detection on Social Media-A Review” published in April 2020.

[3]. Monther Aldwairi and Ali Alwahedin, “Detecting Fake News in Social Media Networks” published in 2018.

[4]. Natali Ruchansky, Sungyong Seo and Yan Liu, “CSI: A Hybrid Deep Model for Fake News Detection”.

[5]. Ray Oshikawa, Jing Qian, William Yang Wang, “A Survey on Natural Language Processing for Fake News Detection” published in March 2020.

[6]. Kai Shu, Suhang Wang, Jiliang Tang, and Huan Liu, “Fake News Detection on Social Media: A Data Mining Perspective”.

[7]. Junaed Yonus Khan, Md. Tawkat Islam Khondaker, Anindya Iqbal, Sadia Afroz, “A Benchmark Study on Machine Learning Methods for Fake News Detection” published in May, 2019.

[8]. Xinyi Zhou, Reza Zafarani, Kai Shu, Huan Liu, “Fake News: Fundamental Theories, Detection Strategies and Challenges” published in January 2019.

[9]. Nichlar Olivirea, Pedro Pisa, Martin Lopez, Diogo Mattos, “Identifying Fake News on Social Networks Based on Natural Language Processing: Trends and Challenges” published in December 2020.

[10]. Srishti Agrawal, Vaishali Arora, Ruchika Arora, Pronika Chawla, Madhumita Kathuria. “FAKE NEWS DETECTION USING ML.