

**Finding Fake News**

Prangige Shashini Randika Peiris

Supervised by: Dr. Dimitra Zografistou

and Dr. Brian Pluss

A dissertation submitted for the Degree of Master of Science

MSc in Data Science and Engineering

School of Science and Engineering

University of Dundee

17th March 2023

**Executive Summary**

**Project Summary**

The objective of this project is to develop a machine learning model to identify fake news. It has become difficult to spot and distinguish whether the news is real or fake as it has developed to a level where one can produce not just fake text but fake pictures and videos similar or indifferent from original picture / video using digital tools. It has become widespread information available online and it is becoming increasingly difficult to distinguish between real and fake news.

**The work that has been done**

The machine learning model was trained on a dataset of articles labelled as real or fake. The model used features, including textual and metadata, to identify patterns and characteristics of fake news articles. After training, the model was tested on a separate dataset to evaluate its accuracy and performance.

**The main outcomes / results**

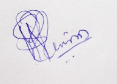
The results of the project showed that the machine learning model can accurately classify fake news articles. This approach could potentially be used by journalists, fact-checkers, and news organizations to quickly identify and flag fake news articles, enabling them to take necessary action to prevent the spread of misinformation.

**Recommendations**

Overall, this project demonstrated the potential for machine learning to be a powerful tool in the combat fake news. With continued development and refinement, machine learning models could help to reduce the bad influences of fake news on the public and support the growth of a more informed and reliable media landscape.

**Declaration**

I declare that the special study described in this dissertation has been carried out and the dissertation composed by me, and that the dissertation has not been accepted in fulfilment of the requirements of any other degree or professional qualification.



Signed[[1]](#footnote-1):

Date: 2023-03-17

Certificate

I certify that Shashini Peiris has satisfied the conditions of the Ordinance and Regulations and is qualified to submit this dissertation in application for the degree of Master of Science.

Signed[[2]](#footnote-2):

Date: 2023-03-17

Acknowledgements

I would like to thank my project supervisors Dr. Dimitra Zografistou and Dr. Brian Pluss for the superior guidance and support given all throughout the short period of project. Mr. Andy Cobley and Dr. Craig Ramsay also must be remembered for the support given during the project.

I would also like to take this opportunity to thank all the past students who shared their valuable work for our reference and all those who supported me in evaluating my project and giving inputs where applicable.

Also, I would like to thank all the students who were generous to support me during the project as well as while I was in Sri Lanka during the Semester 2 (January 2022 - April 2022).

Table of Contents

[**Executive Summary** i](#_Toc133359532)

[**Declaration** ii](#_Toc133359533)

[Certificate iii](#_Toc133359534)

[Acknowledgements iv](#_Toc133359535)

[Table of Contents v](#_Toc133359536)

[List of Tables vii](#_Toc133359537)

[1. Introduction 1](#_Toc133359538)

[1.1 What is Fake News and what are the consequences of fake news? 1](#_Toc133359539)

[1.1.1 Why detecting fake news is important? 1](#_Toc133359540)

[1.1.2 How to find Fake news? 1](#_Toc133359541)

[1.2 Objectives 2](#_Toc133359542)

[2. Background 3](#_Toc133359543)

[2.1.1 Fake News Definition 3](#_Toc133359544)

[2.2 Approaches proposed to detect fake news 3](#_Toc133359545)

[2.3 What is the approach for this project? 4](#_Toc133359546)

[2.4 Have others already done work in this area? 5](#_Toc133359547)

[2.5 How could that help you in your project, or not? 5](#_Toc133359548)

[3. Literature review 6](#_Toc133359549)

[4. Specification 7](#_Toc133359550)

[4.1 Purpose, Scope and Goal 7](#_Toc133359551)

[4.2 Data Collection 7](#_Toc133359552)

[4.2.1 Data collection sources 7](#_Toc133359553)

[4.2.2 Cleaning and pre-processing dataset 7](#_Toc133359554)

[4.3 Feature Extraction 8](#_Toc133359555)

[4.4 Modelling 8](#_Toc133359556)

[4.4.1 Why chose multiple Algorithms? 10](#_Toc133359557)

[4.4.2 Why chose these Algorithms in particular? 10](#_Toc133359558)

[4.5 Performance Metrics 10](#_Toc133359559)

[4.6 Maintenance and Updates 1](#_Toc133359560)

[4.7 Ethical, Legal, Social and Professional Concerns 1](#_Toc133359561)

[4.7.1 Ethical Concerns 1](#_Toc133359562)

[4.7.2 Legal Concerns 2](#_Toc133359563)

[4.7.3 Social Concerns 2](#_Toc133359564)

[4.7.4 Professional Concerns 2](#_Toc133359565)

[4.8 Timeline 2](#_Toc133359566)

[4.9 Weekly Group Meetings (Summary) 3](#_Toc133359567)

[5. Design 4](#_Toc133359568)

[5.1 Environment and Programming Languages 4](#_Toc133359569)

[5.2 Frameworks and Data Structures 4](#_Toc133359570)

[5.3 Design choices that can evolve during the project 5](#_Toc133359571)

[5.4 Process flow 5](#_Toc133359572)

[6. Implementation 6](#_Toc133359573)

[6.1 Functional aspects 6](#_Toc133359574)

[6.2 Non-functional aspects 6](#_Toc133359575)

[6.3 Benchmarking against the work of others 6](#_Toc133359576)

[6.4 Implementation 7](#_Toc133359577)

[7. Evaluation, Testing and Results 13](#_Toc133359578)

[7.1 Evaluation - Model Comparison 13](#_Toc133359579)

[7.1.1 Tabular comparison 13](#_Toc133359580)

[7.1.2 Graphical comparison 13](#_Toc133359581)

[7.1.3 Bar chart comparison 14](#_Toc133359582)

[7.2 Testing and Results - Real World News 15](#_Toc133359583)

[7.3 User Experience 16](#_Toc133359584)

[7.4 Description of the Final Product 16](#_Toc133359585)

[7.4.1 Coding 16](#_Toc133359586)

[7.4.2 Input 17](#_Toc133359587)

[7.4.3 Output 17](#_Toc133359588)

[8. Critical Appraisal, Summary and Conclusions 18](#_Toc133359589)

[8.1 Critical Appraisal 18](#_Toc133359590)

[8.2 Summary 18](#_Toc133359591)

[8.3 Conclusion 18](#_Toc133359592)

[9. Future Work 19](#_Toc133359593)

[10. References / Bibliography 20](#_Toc133359594)

[10.1 Books / Journals 20](#_Toc133359595)

[10.2 Websites 20](#_Toc133359596)

[10.3 Other by project owner 21](#_Toc133359597)

[11. List of Appendices 22](#_Toc133359598)

List of Figures

[Figure 1: Bayes theorem 8](#_Toc133359452)

[Figure 2: Logistic Regression formula 8](#_Toc133359453)

[Figure 3: SVM Classifier Formula 9](#_Toc133359454)

[Figure 4: SGD Classifier Formula 9](#_Toc133359455)

[Figure 5: Random Forest Classifier Formula 9](#_Toc133359456)

[Figure 6: Pseudo Code for Passive-aggressive Classifier 9](#_Toc133359457)

[Figure 7: Accuracy, (Harikrishnan N B, (2019)) 10](#_Toc133359458)

[Figure 8: Precision, (Harikrishnan N B, (2019)) 1](#_Toc133359459)

[Figure 9: Recall, (Harikrishnan N B, (2019)) 1](#_Toc133359460)

[Figure 10: F1-Score, (Harikrishnan N B, (2019)) 1](#_Toc133359461)

[Figure 11: Gantt Chart 3](#_Toc133359462)

[Figure 12: Process flow diagram (Shashini P, (2023)) 5](#_Toc133359463)

[Figure 13: Importing Python libraries 7](#_Toc133359464)

[Figure 14: Read the dataset 7](#_Toc133359465)

[Figure 15: Pre-processed dataset (Refer Appendix A for more) 8](#_Toc133359466)

[Figure 16: Shape of training dataset (Refer Appendix A for more) 8](#_Toc133359467)

[*Figure 17: Data distribution of Training* (Refer Appendix A for more) 8](#_Toc133359468)

[*Figure 18: Dataset quality check - Training* (Refer Appendix A for more) 9](#_Toc133359469)

[*Figure 19: Train count for CountV & TfidfV* 9](#_Toc133359470)

[*Figure 20: Word Clouds for CountV & TfidfV* 9](#_Toc133359471)

[*Figure 21: Word Count and Feature names for CountV and TfidfV* 10](#_Toc133359472)

[Figure 22: Classification Report - Model - Naive Bayes 10](#_Toc133359473)

[Figure 23: Classification Report - Model - Logistic Regression 10](#_Toc133359474)

[Figure 24: Classification Report - Model - Naive Bayes 11](#_Toc133359475)

[Figure 25: Classification Report - Model - Logistic Regression 11](#_Toc133359476)

[Figure 26: K-Fold Cross Validation, (Shanthababu Pandian, (2022)) 11](#_Toc133359477)

[Figure 27: Confusion Matrix - Model - Naive Bayes - CountV 12](#_Toc133359478)

[Figure 28: Confusion Matrix - Model - Logistic Regression - CountV 12](#_Toc133359479)

[Figure 29: Confusion Matrix - Model - Naive Bayes - TfidfV 12](#_Toc133359480)

[Figure 30: Confusion Matrix - Model - Logistic Regression - TfidfV 12](#_Toc133359481)

[Figure 31: Evaluation - F1-score and Accuracy score 13](#_Toc133359482)

[Figure 32: F1-Scores of Different Models (sorted) 14](#_Toc133359483)

[Figure 33: Accuracies of Different Models (sorted) 14](#_Toc133359484)

[Figure 34: Predictions for all models (Refer Appendix A for more) 15](#_Toc133359485)

[Figure 35: A code snippet with input text field 16](#_Toc133359486)

[Figure 36: Final Output - Prediction and Analysis 17](#_Toc133359487)

List of Tables

[Table 1: Confusion Matrix 10](#_Toc133359488)

[Table 2: Timeline 2](#_Toc133359489)

[Table 3: Evaluation - Sort using F1-score 13](#_Toc133359490)

# Introduction

This section addresses the definitions and objectives of this Finding Fake News project.

## What is Fake News and what are the consequences of fake news?

Fake news is false or deceptive information reported as if it is real news, often spread through social media and other online platforms. It can be fabricated intentionally and spread for various motives, such as to deceive or manipulate public opinion, to generate web traffic, or to gain political or financial advantage. Fake news can take many forms, including misleading headlines, falsified images and videos, and fabricated stories or quotes attributed to real people or sources. The spread of fake news can have serious effects, including influencing elections, causing social unrest, and eroding public trust in news media and other institutions.

### Why detecting fake news is important?

Detecting fake news is important because it helps to prevent the spread of misinformation and disinformation, which can have serious consequences. Here are some reasons why:

**Misleading information**: Fake news can contain misleading information that is not accurate or true. This can lead people to make decisions based on false information, which can have serious consequences.

**Public safety**: Fake news can also contain information that is harmful to public safety. For example, false information about natural disasters or health crises can lead people to take actions that put their lives or the lives of others at risk.

**Social division**: Fake news can also be used to fuel social division by spreading false or exaggerated information about certain groups of people. This can lead to hatred and discrimination, which can have long-lasting effects on society.

**Political manipulation**: Fake news can also be used to manipulate political outcomes by spreading false or misleading information about candidates or issues. This can undermine the democratic process and lead to the election of unfit or corrupt leaders.

### How to find Fake news?

Finding fake news is a critical skill in today's digital age, where misinformation and disinformation spread rapidly through social media and other online platforms. This project aims to equip people with the tools and knowledge they need to identify and avoid fake news.

The first step in finding fake news is to **understand** what it is. Fake news refers to false or misleading information presented intentionally as if it were real news. It can be spread through various channels, such as social media, websites, blogs, and even traditional news media.

To identify fake news, it is essential to **evaluate the source** of the information. Popular news sources have editorial standards and fact-checking processes in place to ensure the accuracy of their reporting. On the other hand, fake news sources often lack transparency and may use misleading headlines, fabricated quotes, or doctored images to manipulate readers.

Another crucial aspect of identifying fake news is to be aware of **bias** and to look for multiple sources to verify the information. Biased sources may selectively present information to support their viewpoint, while reputable news sources strive for objectivity and balance.

("How to Spot Fake News" by FactCheck.org) [[3]](#footnote-3)

## Objectives

The objective of this project is to develop machine learning models to identify fake news. It has become difficult to spot and distinguish whether the news is real or fake as it has developed to a level where one can produce not just fake text but fake pictures and videos similar or indifferent from original picture / video using digital tools. Fake news has become widespread information available online and it is becoming increasingly difficult to distinguish between real and fake news.

The other objective is to learn and master data analysis and programming skills that is pretty much required in the IT industry and Data Engineering.

# Background

Fake news has become more prevalent in the recent past due to the rapid spread of information through social media and the internet. The purpose of fake news is to deceive and mislead readers by looking and feeling like legitimate news, but it is specially crafted to deceive and mislead them. Fake news is having an increasingly significant impact on individuals, organizations, and societies. It may undermine trust in reliable sources of information, cause confusion and misunderstandings, and even result in harm.

### Fake News Definition

**The fake news term originally refers to false and often sensationalist information disseminated under the guise of relevant news**

(Caulfield, M. (2017). Web Literacy for Student Fact-Checkers. Pressbooks) [[4]](#footnote-4)

Methods that can detect fake news:

* The false information in it
* The writing style of the article
* The propagation pattern of articles
* The credibility of its source

(“A Survey of Fake News”, Xinyi Zhou, Reza Zafarani, 2020) [[5]](#footnote-5)

**These 4 tips can be used to check the truthfulness of a news**:

The story: what are they trying to say? Is it an ad or a joke? Look to see if you can find the same story somewhere else

The author: is it someone’s opinion or a fact? Real news will most likely have a link to the writer’s details, but if there’s no author, dig deeper

The website: are there spelling or grammar mistakes? What’s the URL? Check the address bar at the top - most trusted URLs end with “.com”, “.co.uk”, “.net”, “.gov”, “.org”, “.mil” and “.edu”

The date: is the story recent or old? It could be outdated or a copy of something that happened years ago. Computer programs called bots post anytime & often, so be wary of this ("Tips for spotting fake news online" by BBC Bitesize) [[6]](#footnote-6)

## Approaches proposed to detect fake news

1. ***Language approach***: This approach focuses on the use of linguistics by a human or software program to detect fake news. The approach considers all the words in a sentence and letters in a word, how they are structured and how it fits together in a paragraph (Yang et al. 2018). The focus is therefore on grammar and syntax (Burkhardt 2017). The 3 main methods of language approach:
   1. Bag of Words (BOW)
   2. Semantic Analysis
   3. Deep Syntax
2. ***Topic-agnostic approach***: This category of approaches detects fake news by not considering the content of articles bur rather topic-agnostic features. The approach uses linguistic features and web mark-up capabilities to identify fake news (Castelo et al. 2019; Horne and Adali 2017). Examples of topic-agnostic features are:
   1. Many advertisements
   2. Longer headlines with eye-catching phrases
   3. Different text patterns from mainstream news to induce emotive responses
   4. Presence of an author name (Castelo et al. 2019; Horne and Adalı 2017)
3. ***Machine learning approach***: Machine learning algorithms can be used to identify fake news.
   1. Using different types of training datasets to refine the algorithms (Qazvinian et al. 2011)
   2. Rumour identification framework
4. ***Knowledge-based approach***: The knowledge-based approach aims at using sources that are external to verify if the news is fake or real and to identify the news before the spread thereof becomes quicker. There are three main categories:
   1. Expert Oriented Fact Checking
   2. Computational Oriented Fact Checking
   3. Crowd Sourcing Oriented Fact Checking (Hassan et al. 2017)
5. ***Hybrid approach***: There are three generally agreed upon elements of fake news articles (Ruchansky, Seo, and Liu 2017):
   1. The text of an article
   2. The response that the articles received
   3. The source used that motivate the news article

A recent study has been conducted that proposes a hybrid model which helps to identify fake news on social media through using a combination of human and machine learning to help identify fake news (Okoro et al. 2018).

Another hybrid model called CSI (capture, score, integrate) has been developed and functions on the main elements

1. Capture - the process of extracting representations of articles by using a Recurrent Neutral Network (RNN)
2. Score - to create a score and representation vector
3. Integrate - to integrate the outputs of the capture and score resulting in a vector which is used for classification (Ruchansky, Seo, and Liu 2017) [[7]](#footnote-7)

## What is the approach for this project?

A machine learning approach has been used in this project to build different models to identify and compare fake news.

## Have others already done work in this area?

Fake news is not a problem that has been identified recently. This problem has been there since ancient times and there was no solution discovered using technology until recent days.

There are many machine learning projects carried out to identify fake news and there are many sources on the web related to this issue. Some of these projects are in Github and few of such projects are given below:

1. <https://github.com/nishitpatel01/Fake_News_Detection>
2. <https://github.com/Spidy20/Fake_News_Detection>
3. <https://github.com/SaikrishnaPulipati533/Fake-News-Detection_NLP-Project>

Different projects demonstrate different approaches identified to detect fake news.

## How could that help you in your project, or not?

The following project took my attention as it uses multiple machine learning models to identify fake news and Passive Aggressive classifier has been used in this project to create the machine learning model to detect fake news.

<https://github.com/nishitpatel01/Fake_News_Detection>

Some of the codes were re-used in my project but changes were made as I use 6 different classifiers to create 12 models.

# Literature review

Fake news is defined as false or misleading information presented as if it were true, often with the intention of deceiving people. The spread of fake news is a serious problem, as it can lead to a lack of trust in legitimate sources of information and can cause harm to individuals and society. In this literature review, we will explore various methods for finding and detecting fake news, including automated tools and human interventions.

**Automated Tools**:

The authors propose a machine learning approach for detecting fake news using a dataset of real and fake news articles. They use natural language processing techniques to extract features such as sentiment, readability, and topic coherence, and then use these features to train a machine learning model to classify articles as real or fake. (Zhang, Y., & Zhou, D. (2020). Detecting Fake News with Machine Learning. Journal of Information Science Theory and Practice, 8(3), 7-22. doi:10.1633/JISTaP.2020.8.3.1)

This review article provides an overview of the current state of fake news detection using automated tools. The authors discuss various methods, including machine learning, deep learning, and network analysis, and evaluate the strengths and weaknesses of each approach. (Mottaghi, S., & Naderi, H. (2021). A Review of Fake News Detection: Methods, Datasets, and Future Challenges. Journal of Information Science Theory and Practice, 9(1), 33-52. doi:10.1633/JISTaP.2021.9.1.3)

**Human Interventions**:

The authors propose a crowdsourcing approach to identifying fake news, in which participants rate the quality of news sources based on their perceived reliability. They find that this approach is effective at identifying fake news and reducing its spread on social media platforms. (Pennycook, G., & Rand, D. G. (2019). Fighting misinformation on social media using crowdsourced judgments of news source quality. Proceedings of the National Academy of Sciences, 116(7), 2521-2526. doi:10.1073/pnas.1806781116)

This article provides an overview of the current state of research on fake news, including the role of social media in its spread and the effectiveness of various interventions. The authors suggest that a combination of technological and human interventions is necessary to combat fake news effectively. (Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., . . . Zittrain, J. L. (2018). The science of fake news. Science, 359(6380), 1094-1096. doi:10.1126/science.aao2998)

Fake news is a complex problem that requires a complicated approach to address. Automated tools such as machine learning and network analysis can be effective at identifying fake news, but they are not dependable and require ongoing development and improvement.

Human interventions such as crowdsourcing and expert fact-checking can also be effective at identifying fake news, but they are resource-intensive and may not scale well. A combination of these approaches is likely necessary to combat fake news effectively.

# Specification

The project specification describes the Project Overview, Data Collection, Feature Extraction, Modelling, Performance Metrics, User Interface, Deployment, Maintenance and Updates, Ethical Considerations and Timeline and Budget.

A brief description of the project, including its purpose, scope, and goals will be given in this section.

## Purpose, Scope and Goal

The purpose of this project is to develop machine learning models to identify fake news using machine learning algorithms.

The scope of Finding Fake News project includes the following:

1. Find a big dataset
2. Clean the dataset
3. Train multiple models
4. Analyse all the models
5. Run cross validation
6. Test all models on new data
7. Present results using each model

The goal of a "Finding Fake News" project is to develop a machine learning model that can accurately identify and classify news articles as either real or fake. The steps described above in the scope will be carried out to achieve this goal.

## Data Collection

The source of data and pre-processing methods used will be described briefly in this section.

### Data collection sources

Kaggle.com and Github.com was used mainly to find suitable datasets and a dataset was identified from Github use in this project.

Source of dataset: <https://github.com/nishitpatel01/Fake_News_Detection>

### Cleaning and pre-processing dataset

The dataset seems to be a clean dataset, unwanted columns and empty rows were removed already and with the time constraints, I decided to go ahead with it and pre-process using lemmatization and stemming. The purpose of combining lemmatization and stemming in this way is to normalize words and reduce the dimensionality of the textual data, making it easier to process and analyze.

Stemming and Lemmatization are text normalization techniques used in Natural Language Processing. The stemming is faster as it cuts words without knowing the context, while lemmatization is slower as it identifies the context of words before processing. ("Stemming and Lemmatization in Python NLTK with Examples", Daniel Johnson, 2023) [[8]](#footnote-8)

## Feature Extraction

The cleaned and pre-processed data then went through feature extraction using both Count Vectorizer and TFIDF Vectorizer approaches for the purpose of learning as well as to demonstrate the differences between both approaches to evaluate and identify which machine learning model gives the best results.

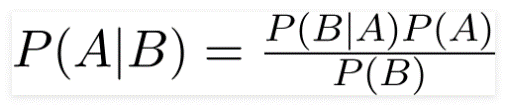
Both Count Vectorizer and TFIDF Vectorizer are methods to convert text data into numeric form so that machine can understand. Machine learning models cannot identify text hence it is required to be converted into numeric form.

Count Vectorizer counts the number of instances each word appears in a document and creates a matrix representing the frequency of each word in each document. Tfidf Vectorizer, counts the number of times each word appears in a document as well how common or rare each word is across all documents. Tfidf Vectorizer is generally considered to be more effective than Count Vectorizer for tasks like text classification. ("Machine Learning 101: CountVectorizer Vs TFIDFVectorizer", Dylan Kaplan, 2002) [[9]](#footnote-9)

## Modelling

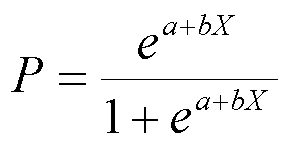
The following supervised learning models will be used in this project to identify truthfulness of news articles:

1. **Naïve bayes classifier** - A Naive Bayes classifier is a probabilistic machine learning model that’s used for classification task which is based on the Bayes theorem. (Nguyen, H. A., & Nguyen, H. L. (2019)) [[10]](#footnote-10)

[[11]](#footnote-11)

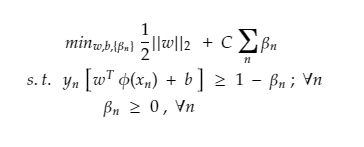
**Figure 1: Bayes theorem**

1. **Logistic Regression classifier** - Logistics regression is a statistical approach and a Machine Learning algorithm that is used for classification problems and is based on the concept of probability. It is used when the dependent variable (target) is categorical. It is widely used when the classification problem at hand is binary; true or false, yes, or no, and it uses the sigmoid function to return the probability of a label. (Nguyen, H. A., & Nguyen, H. L. (2019)) [[12]](#footnote-12)

[[13]](#footnote-13)

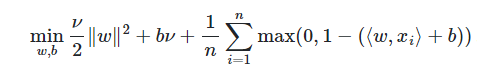
**Figure 2: Logistic Regression formula**

1. **SVM (Support vector machines) classifier** - The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N - the number of features) that distinctly classifies the data points. (Cortes, C., & Vapnik, V. (1995)) [[14]](#footnote-14)

[[15]](#footnote-15)

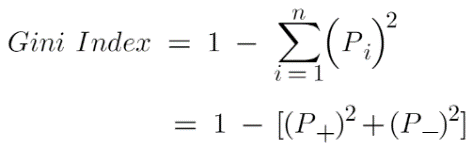
**Figure 3: SVM Classifier Formula**

1. **SGD (Stochastic Gradient Descent) classifier** - Stochastic Gradient Descent (SGD) is an efficient optimization algorithm used to find the values of parameters/coefficients of functions that minimize a cost function. (Bottou, L., Curtis, F. E., & Nocedal, J. (2018)) [[16]](#footnote-16)

[[17]](#footnote-17)

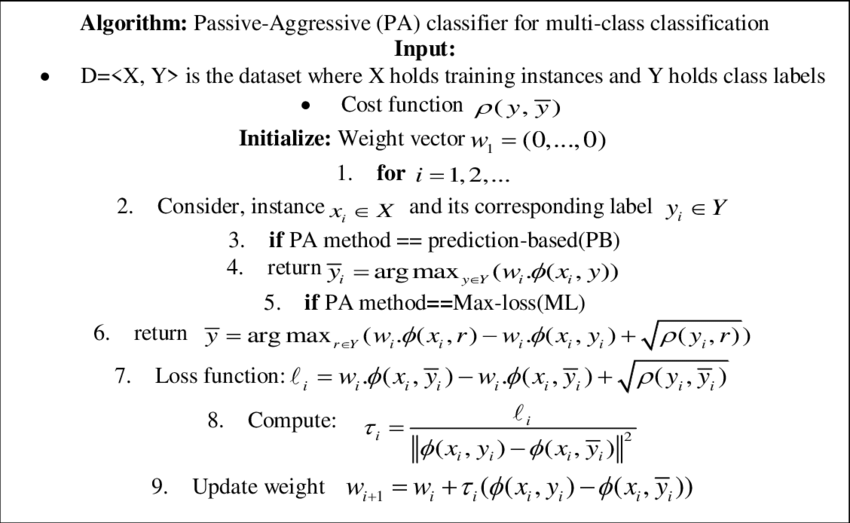
**Figure 4: SGD Classifier Formula**

1. **Random Forest classifier** - A random forest is a meta estimator that fits multiple decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. (Breiman, L. (2001)) [[18]](#footnote-18)

[[19]](#footnote-19)

**Figure 5: Random Forest Classifier Formula**

1. **Passive Aggressive classifier** - The Passive-Aggressive algorithms are a family of Machine learning algorithms and are generally used for large-scale learning. (Alokesh, (2023)) [[20]](#footnote-20)

[[21]](#footnote-21)

**Figure 6: Pseudo Code for Passive-aggressive Classifier**

These classifiers will be used in combination with both Count Vectorizer and Tfidf Vectorizer and 12 models will be developed.

Each model will be stored in a pickle model so that it can be used without training dataset again and again which reduces the time taken to train different models.

### Why chose multiple Algorithms?

Algorithms used:

1. Naive Bayes
2. Logistic Regression
3. SVM (Support Vector Machine)
4. SGD (Stochastic Gradient Descent)
5. Random Forest
6. Passive Aggressive

I decided to choose all the above models to demonstrate the differences between different models and how I can use these differences to explain the outcomes in my project. In general, there are multiple benefits of using multiple models and few of the benefits are:

**Improved accuracy**: it increases the accuracy of predictions and since each model has different strengths and weaknesses, a combined prediction is much healthier and accurate for a system.

**Reducing bias**: the use of one classifier may produce a bias prediction and it can be reduced when multiple classification models are used.

**Handling uncertainty**: different news articles can carry a combination of both true and fake information so use of multiple classifiers may help to handle uncertainty more effectively and avoid false positives or false negatives.

### Why chose these Algorithms in particular?

The above classifiers are Supervised Learning Algorithms and Supervised Learning Algorithms are used majorly to text classification tasks. Supervised Learning Algorithms requires labelled data to train models created using these algorithms to identify the relationship between the features and the outcome, i.e., to identify whether an article is fake or real and a model trained as above can classify and predict new and unseen articles fairly.

## Performance Metrics

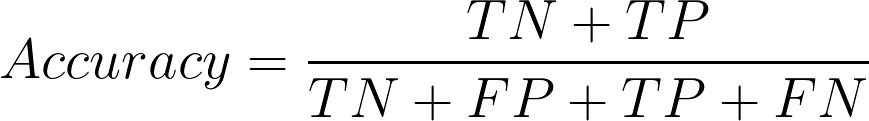
The performance metrics are derived from **Confusion matrix**.

|  |  |  |  |
| --- | --- | --- | --- |
| **Confusion Matrix** | | **Predicted** | |
| ***Positive*** | ***Negative*** |
| **Actual** | ***Positive*** | True positive | False Negative |
| ***Negative*** | False Positive | True Negative |

**Table 1: Confusion Matrix**

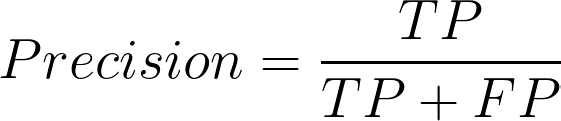
The Accuracy and F1-score are the performance metrics that will be used to evaluate the effectiveness of the models in this project.

1. **Accuracy** - Accuracy represents the number of correctly classified data instances over the total number of data instances.



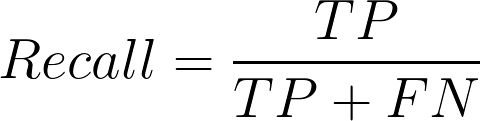
**Figure 7: Accuracy,** (Harikrishnan N B, (2019))

1. **Precision** - This measures the proportion of true positives among all the samples that the model classified as positive.



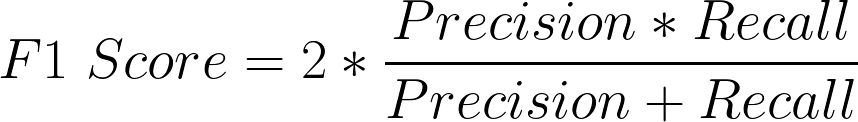
**Figure 8: Precision,** (Harikrishnan N B, (2019))

1. **Recall** - This measures the proportion of true positives among all the actual positive samples in the dataset.



**Figure 9: Recall,** (Harikrishnan N B, (2019))

1. **F1 score** - F1 Score can be used to identify the balance between Precision and Recall.



**Figure 10: F1-Score,** (Harikrishnan N B, (2019)) [[22]](#footnote-22)

## Maintenance and Updates

Maintenance and update procedures for the model, with feedback mechanisms for improving its accuracy will be described in this section.

1. Different datasets also can be used to train these models to improve accuracy over time.
2. New features such as more models built using different classifiers
3. More visualizations can be introduced to ease decision making.
4. Surveys can be introduced to identify user experiences and suggestions of improvements.

These aspects will be discussed again in the Future work. (**Future Work**)

## Ethical, Legal, Social and Professional Concerns

When evaluating a project on finding fake news, it is important to consider the legal, social, and professional issues that may surface.

### Ethical Concerns

This section discusses any ethical considerations associated with the project, such as the potential for bias or unintended consequences, and describe strategies for mitigating these risks.

1. ***Problem***: The training dataset used to develop the machine learning models can be bias to a particular domain.

***Solution***: The news in training data is derived using news from different domains

1. ***Problem***: News can be influenced by personal beliefs, cultural aspects, and political perceptions so there is a risk of falsely identifying true news as fake and it will be controversial.

***Solution***: Provide clear guidelines and ethical principles for identifying and labelling fake news

It is important to continually monitor and evaluate the performance of the machine learning models. (John Doe and Jane Smith, (2020))

### Legal Concerns

It is important to ensure that the project does not violate any laws related to privacy, data protection, or copyright and follow the terms and conditions where appropriate. It should always ensure Citations and References are given where appropriate and if the project is collecting user data, it is important to obtain informed consent and that data protection regulations are followed. (Raza, K., & Elahi, S. (2018))

### Social Concerns

The project should consider its impact on society, including how it may contribute to the spread or prevention of fake news and it should ensure that the project does not contribute to the spread of hate speech, discrimination, or other harmful content and assess the project's potential impact on privacy and individual freedoms. (Van Dijck, J., Poell, T., & De Waal, M. (2018))

### Professional Concerns

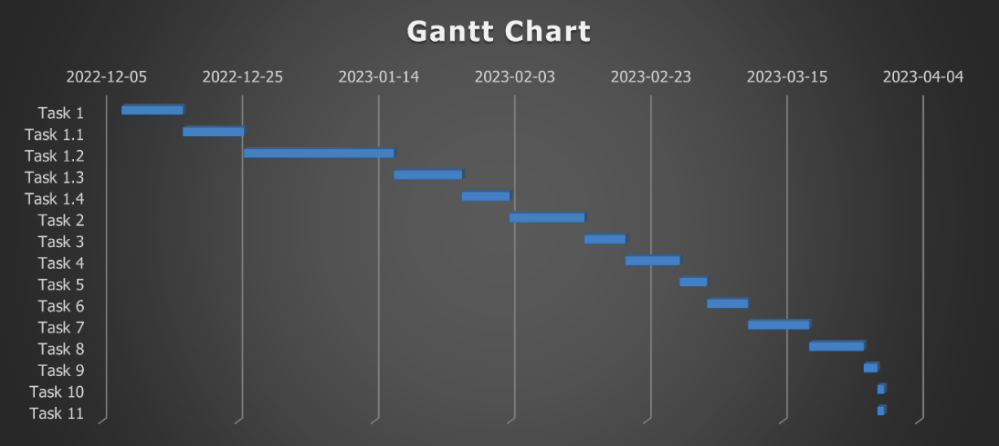
The project should assess the quality and accuracy of the project's findings and methods as well as it should ensure that any conclusions drawn by the project are supported by evidence and can be replicated by others. (Ioannidis, J. P. (2005))

## Timeline

This table consists of the initial project timeline with milestones and deliverables, but it was changed due to unavoidable circumstances.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Task no** | **Task** | **Duration** | **Start Date** | **End Date** |
| Task 1 | Project Initiation | 9 | 2022-12-08 | 2022-12-17 |
| Task 1.1 | Project choices | 9 | 2022-12-17 | 2022-12-26 |
| Task 1.2 | Project allocation | 22 | 2022-12-26 | 2023-01-17 |
| Task 1.3 | Risk Assessment Form | 10 | 2023-01-17 | 2023-01-27 |
| Task 1.4 | Ethics Declaration Form | 7 | 2023-01-27 | 2023-02-03 |
| Task 2 | Literature Review | 11 | 2023-02-03 | 2023-02-14 |
| Task 3 | Data Collection | 6 | 2023-02-14 | 2023-02-20 |
| Task 4 | Data Pre-processing | 8 | 2023-02-20 | 2023-02-28 |
| Task 5 | Model Development | 4 | 2023-02-28 | 2023-03-04 |
| Task 6 | Model Testing | 6 | 2023-03-04 | 2023-03-10 |
| Task 7 | Results Analysis | 9 | 2023-03-10 | 2023-03-19 |
| Task 8 | Report Writing | 8 | 2023-03-19 | 2023-03-27 |
| Task 9 | Presentation Prep | 2 | 2023-03-27 | 2023-03-29 |
| Task 10 | Final Presentation | 1 | 2023-03-29 | 2023-03-30 |
| Task 11 | Project Closure | 1 | 2023-03-29 | 2023-03-30 |

**Table 2: Timeline**



**Figure 11: Gantt Chart**

This Gantt Chart was prepared based on the project timeline as planned at the beginning. But it had to be changed due to unforeseen circumstances.

## Weekly Group Meetings (Summary)

Weekly group meetings were conducted with the supervisors with valuable inputs and guidance as well as to identify any misunderstandings and clear them. E-Mails conversations were carried out where required to clarify issues risen during the process. The points discussed mainly are:

1. What is the project and how to identify requirements with next steps
2. Analysing and designing the prototype
3. Demonstrations of developed prototypes and improvements
4. Report writing and how it should describe the system
5. Submission requirements

Due to unplanned and unavoidable personal circumstances, I had to apply for mitigation circumstances, and I obtained extension, and my submission deadline is set as 27th April 2023. During this time, my supervisors were kind enough to conduct weekly meetings personally and guide me as well as encourage me about the project to end it successfully.

# Design

The design choices made in this project is described in this section.

## Environment and Programming Languages

This section describes the resource requirements of this project.

1. Windows laptop with 16GB RAM
2. Google Chrome browser to run development tools
3. Google docs Drawings tool has been used as the tool to draw process flow (<https://docs.google.com/drawings/>)
4. Google Colab (updated to Python 3.8) is used as the development tool as well as to deploy the system (<https://colab.research.google.com/>). Google Colab allows to write and execute arbitrary python code through the browser, and it facilitates machine learning and data analysis.
5. Python programming language used in this machine learning project

## Frameworks and Data Structures

The following machine learning frameworks of Python programming language has been used in this project:

* Pandas: A library for data manipulation and analysis.
* NumPy: A library for numerical computation.
* NLTK: A library for natural language processing.
* Seaborn: A library for data visualization.
* Scikit-learn: A machine learning library that provides a wide range of algorithms and tools for data analysis and modelling.
* Wordcloud: A library for generating word clouds from text.

The following algorithms and tools from Scikit-learn is used in this project as well:

* Count Vectorizer: A tool for vectorizing text data.
* Tfidf Transformer: A tool for transforming text data using the term frequency-inverse document frequency (TF-IDF) algorithm.
* Pipeline: A tool for constructing and running machine learning pipelines.
* F1\_score: A tool for evaluating the performance of classification models.
* Confusion Matrix: A tool for visualizing the performance of classification models.

These Python libraries are also used:

* Matplotlib & Seaborn: Libraries used for data visualization.
* Multiprocessing: A library for parallel computing.
* Pickle: A library for serialization and deserialization of Python objects.
* Regular expressions (re): A library for pattern matching and string manipulation.

Efficient data structures are important for storing and manipulating large datasets in fake news detection systems. The following Data Structures are used in this project:

* Arrays
* Lists
* Dictionaries
* Graphs

## Design choices that can evolve during the project

It was decided to use one supervised learning model at the beginning of the project but later identified that comparison of multiple different models using different classifications would be ideal to make this project unique.

During the design it was decided to train multiple models and identify the best model to use in identifying fake news. While discussing with the supervisors, the idea of using all the developed models to be used in the final result was decided, i.e. train multiple models and use all of the models to predict the truthfulness of an article and predict an overall result.

## Process flow

This is the process flow for detecting real or fake news using machine learning:

**Collect a few datasets of news articles**: A couple of datasets of news articles were collected from sources such as Github.com or Kaggle.com.

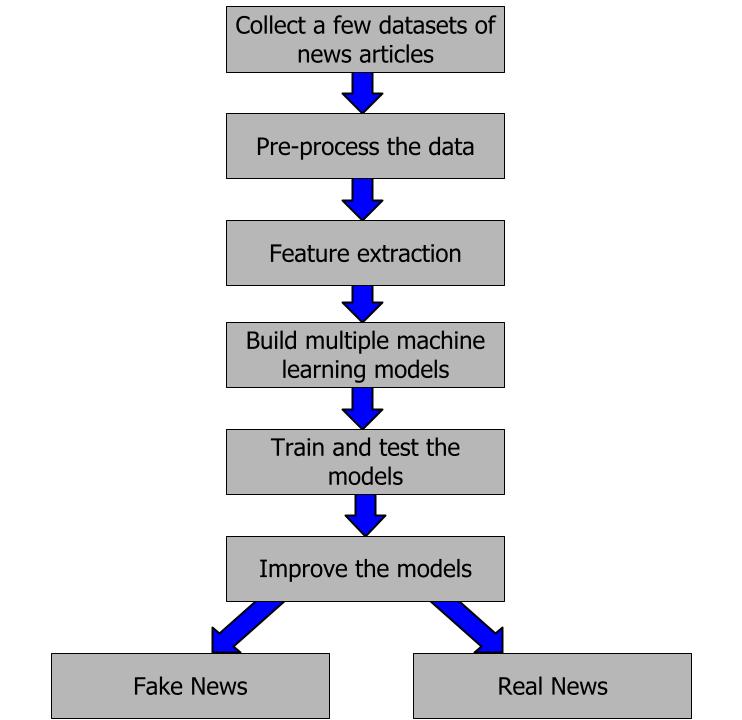
**Pre-process the data**: Then the data in the dataset were cleaned and pre-processed. This included tasks like removing special characters, removing stop words, stemming/lemmatizing the text, and converting the text into lowercase.

**Feature extraction**: Next, features extraction step was carried out from the pre-processed data using the technique TF-IDF Vectorizer and Count Vectorizer both.

**Build multiple machine learning models**: After feature extraction, I had to build machine learning models using different classification algorithms like Naive Bayes, Logistic Regression, SVM, SGD, Random Forest and Passive Aggressive to train the models.

**Train and test the models**: Once a model is developed, the dataset needs to be trained. Then the model should be tested using a separate test dataset to see how it generalizes to new data.

**Result - Real news or Fake news**: Finally, the trained models can be used to classify new news articles as either real or fake. The output of the models will be a binary decision: **Real or Fake**.



**Figure 12: Process flow diagram (**Shashini P, (2023)**) [[23]](#footnote-23)**

# Implementation

This section describes the important aspects of implementation, testing and debugging that went through to produce the system: (Further information can be found in the Appendix A)

The functional aspects and non-functional aspects of the Finding Fake News project is also described below.

## Functional aspects

These are some of the functional aspects to consider when implementing a Finding fake news project:

**Data collection**: The system uses datasets from Github as well as Kaggle that has reliable and comprehensive datasets of news articles.

**Data visualization**: The system uses ***Matlab*** and ***Seaborn*** to generate visualizations in a way that is easy to understand and helps users identify trends and patterns.

**User feedback**: A simple survey prepared using Google Form or Microsoft Form can be used to collect user feedback. This will be implemented in the future. (**Future Work**)

## Non-functional aspects

When implementing a project to find fake news, there are several non-functional aspects to consider, such as Security, Scalability, Performance, Usability, Accessibility and these aspects are critical to ensuring the effectiveness, reliability, and usability of a fake news detection system

**Security**: The security requirements should be researched, identified, and evaluated to protect it against cyber threats and prevent unauthorized access. (**Future Work**)

**Accessibility**: The accessible requirements also must be identified and addressed so that system is available to users with disabilities and provide support for multiple languages and cultural contexts. (**Future Work**)

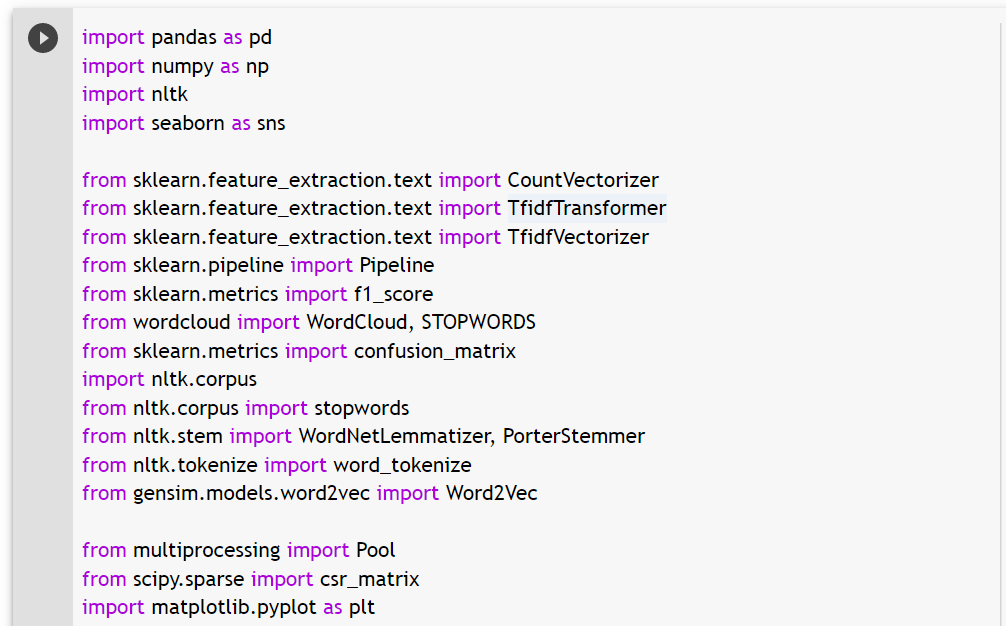
## Benchmarking against the work of others

As mentioned in the Have others already done work in this area?, each developer has used either Count Vectorizer or Tfidf Vectorizer as well as only one classification model, mostly Passive Aggressive classifier, or Logistic Regression classifier to detect fake news in their projects.

But this project uses multiple classification models to predict whether some news article is True or Fake while the work done by other developers only one classification model. This is a significant difference in this system although some re-uses it’s features and coding with significant improvements.

## Implementation

1. **Import Python Libraries**: The required libraries for the project are imported in this first step of the coding. More info: Refer Appendix section

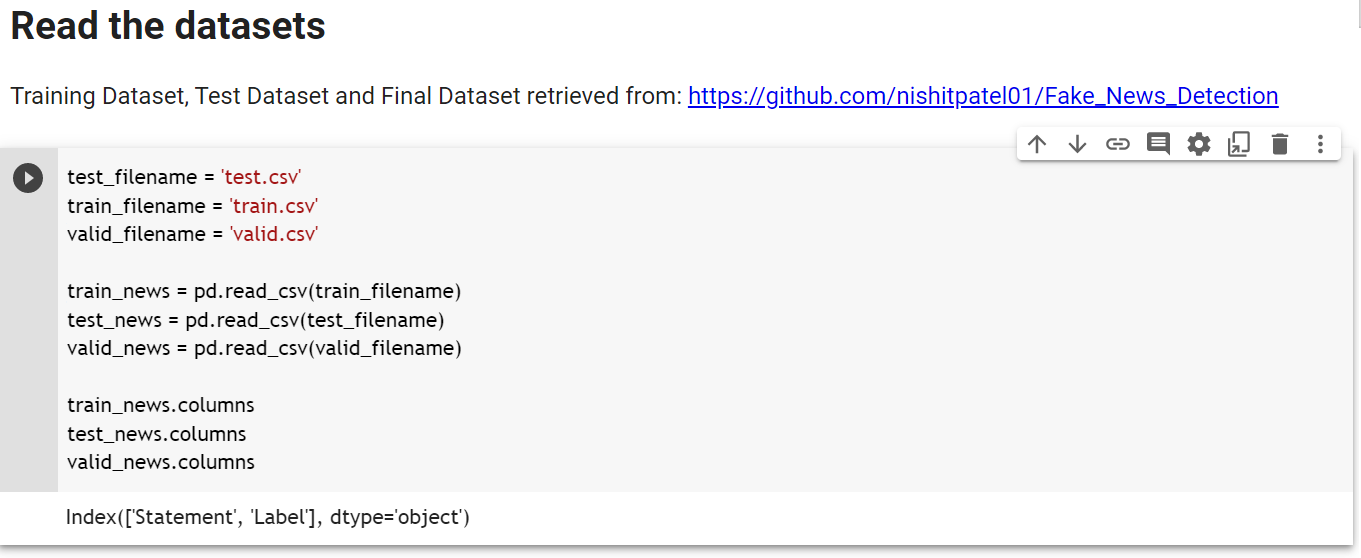


**Figure 13: Importing Python libraries**

1. **Read the datasets**: The dataset is read into the system using Pandas framework. Then it will be saved into an array.

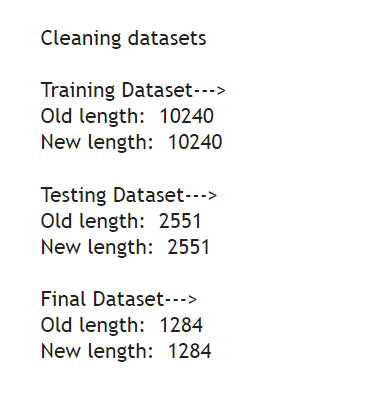
* train.csv - This csv file contains all the data that is used to train the models.
* test.csv - This csv file contains all the data that is used for testing the models.
* valid.csv - This csv file contains the data that can be used to predict using the trained models.

The columns of the datasets will be displayed.



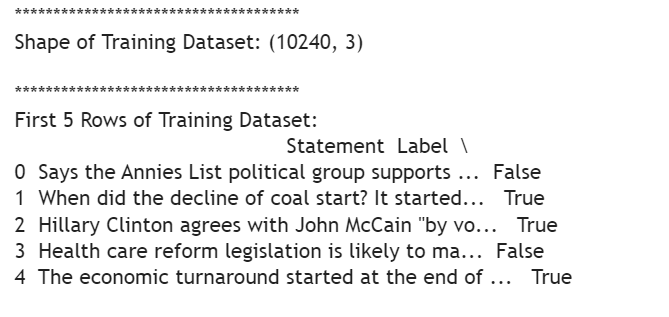
**Figure 14: Read the dataset**

1. **Data Pre-processing**: The next step is pre-processing the data in the dataset that has been read and loaded from 3 different CSV files. The techniques used in this step is described above in 4.2.2 section.



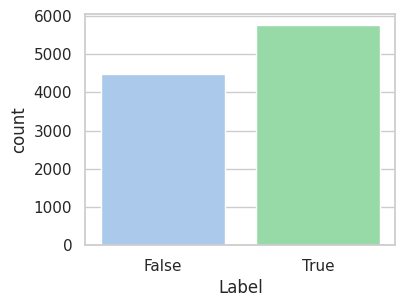
**Figure 15: Pre-processed dataset** (Refer Appendix A for more)

1. **View the Datasets**: The shapes of each of the datasets will be displayed in this step:



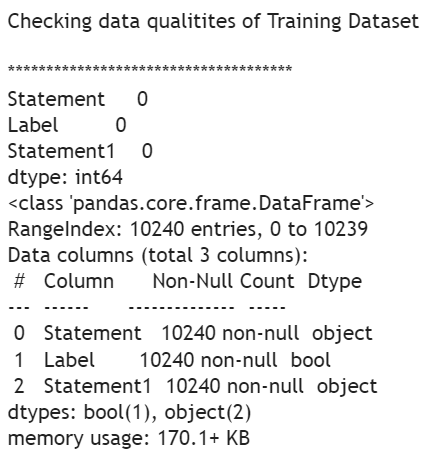
**Figure 16: Shape of training dataset** (Refer Appendix A for more)

1. **Data Distribution**: These steps are carried out to get an idea about distribution of true data and fake data in all 3 datasets. These graphs indicate that the data distribution is fair and balanced.



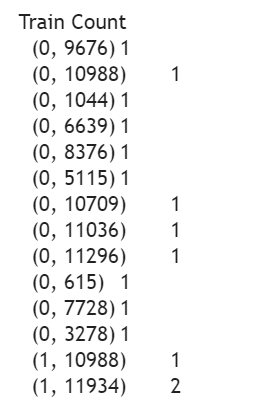
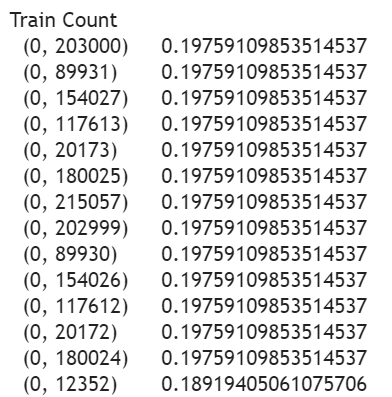
***Figure 17: Data distribution of Training*** (Refer Appendix A for more)

1. **Data integrity check**: This is to check whether there are any null values in all 3 datasets.

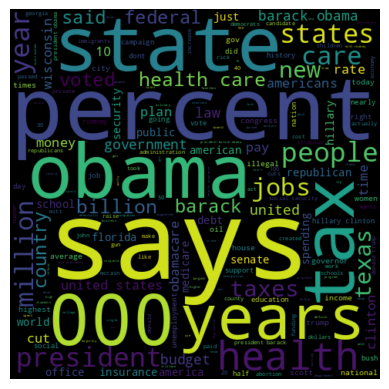
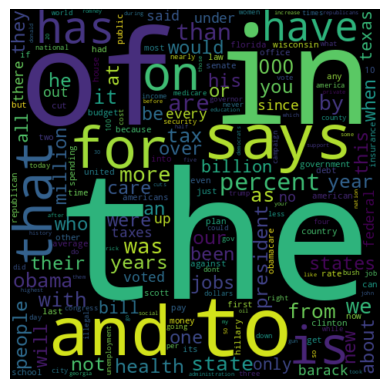


***Figure 18: Dataset quality check - Training*** (Refer Appendix A for more)

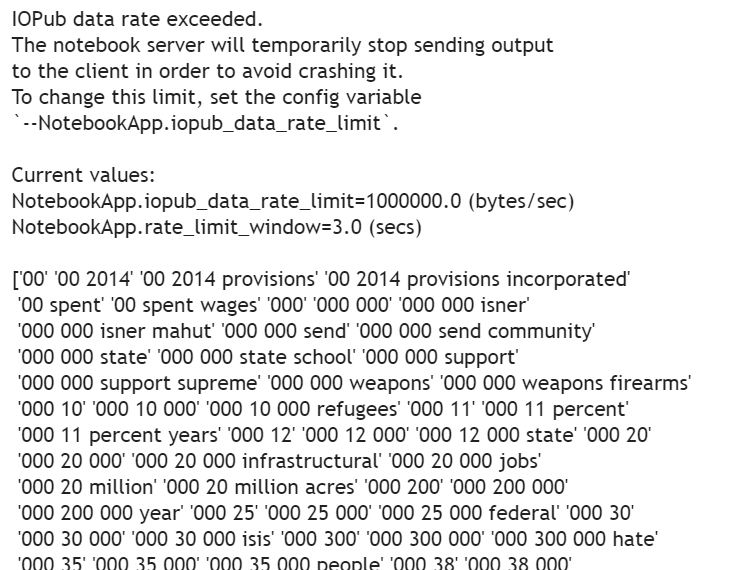
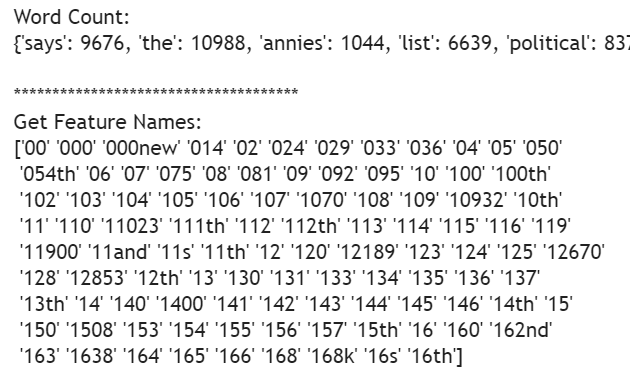
1. **Count Vectorizer and Tfidf Vectorizer**: This is the step to convert the text data into numeric data. These techniques are described in the section 3.3. above. Train count matrix and Word Clouds are used to show how the words are distributed in the datasets. Also Word count and Feature names can be viewed.

***Figure 19: Train count for CountV & TfidfV***



***Figure 20: Word Clouds for CountV & TfidfV***



***Figure 21: Word Count and Feature names for CountV and TfidfV***

***Note***: The pipeline function in Python is used to eliminate multiple steps of code and it will run multiple transformation in one go. The pipeline object is then fit on the training set using the 'fit' method, which trains the model on the training data. The fitted pipeline is then used to make predictions on the test set using the 'predict' method.

A pickle file for each model is built at the end of each function which can be re-used without training datasets again and again.

1. **Building Models using Bag of Words Technique (Count Vectorizer)**

Naïve Bayes, Logistic Regression, SVM, SGD, Random Forest and Passive Aggressive binary classifiers are being used in building the machine learning models using Count Vectorizer.

The F1-Score and Accuracy score is calculated for the purpose of comparing models in the evaluation.

A classification report is produced that shows the precision, recall, f1-score and accuracy and weighted average.

|  |  |
| --- | --- |
| **Figure 22: Classification Report - Model - Naive Bayes** | **Figure 23: Classification Report - Model - Logistic Regression** |

(Refer Appendix A for more Classification reports)

1. **Building Models using N-Grams Technique (TfidfVectorizer)**

Naïve Bayes, Logistic Regression, SVM, SGD, Random Forest and Passive Aggressive binary classifiers are being used in building the machine learning models using Tfidf Vectorizer.

The F1-Score and Accuracy score is calculated for the purpose of comparing models in the evaluation (section 5.2.).

A classification report is produced that shows the precision, recall, f1-score and accuracy and weighted average.

|  |  |
| --- | --- |
| **Figure 24: Classification Report - Model - Naive Bayes** | **Figure 25: Classification Report - Model - Logistic Regression** |

1. **K-fold Cross Validation for all classifiers**

K-Fold cross-validation is a technique used to assess the performance of a model by splitting the data into k equal-sized folds and training the model on k-1 folds while testing it on the remaining fold. This process is repeated k times, with each fold being used as the test set once.

The K-Fold cross-validation technique helps to address the problem of overfitting by using all the available data for both training and testing, without sacrificing the ability to evaluate the performance of the model. It also provides a more accurate estimate of the model's performance than a single train-test split, as it uses multiple splits of the data.

Here's how K-Fold cross-validation works:

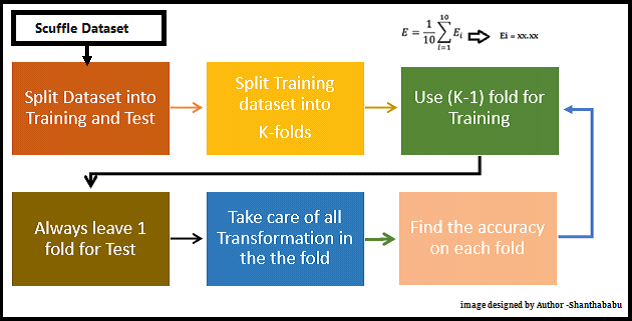
Split the data into k equal-sized folds.

For each of the k folds:

1. Use k-1 folds for training the model.
2. Use the remaining fold for testing the model.
3. Evaluate the performance of the model on the test set.

Repeat the process k times, using each fold as the test set once.

Calculate the average performance across all k folds to get an estimate of the model's performance. (Kohavi, R. (1995)) [[24]](#footnote-24)



**Figure 26: K-Fold Cross Validation,** (Shanthababu Pandian, (2022)) [[25]](#footnote-25)

|  |  |
| --- | --- |
| **Graphical representation of Confusion Matrices generated for different Models with Count Vectorizer** | |
| **Figure 27: Confusion Matrix - Model - Naive Bayes - CountV** | **Figure 28: Confusion Matrix - Model - Logistic Regression - CountV** |

|  |  |
| --- | --- |
| **Graphical representation of Confusion Matrices generated for different Models with Tfidf Vectorizer** | |
| **Figure 29: Confusion Matrix - Model - Naive Bayes - TfidfV** | **Figure 30: Confusion Matrix - Model - Logistic Regression - TfidfV** |

(Refer Appendix A for more Confusion Matrices)

# Evaluation, Testing and Results

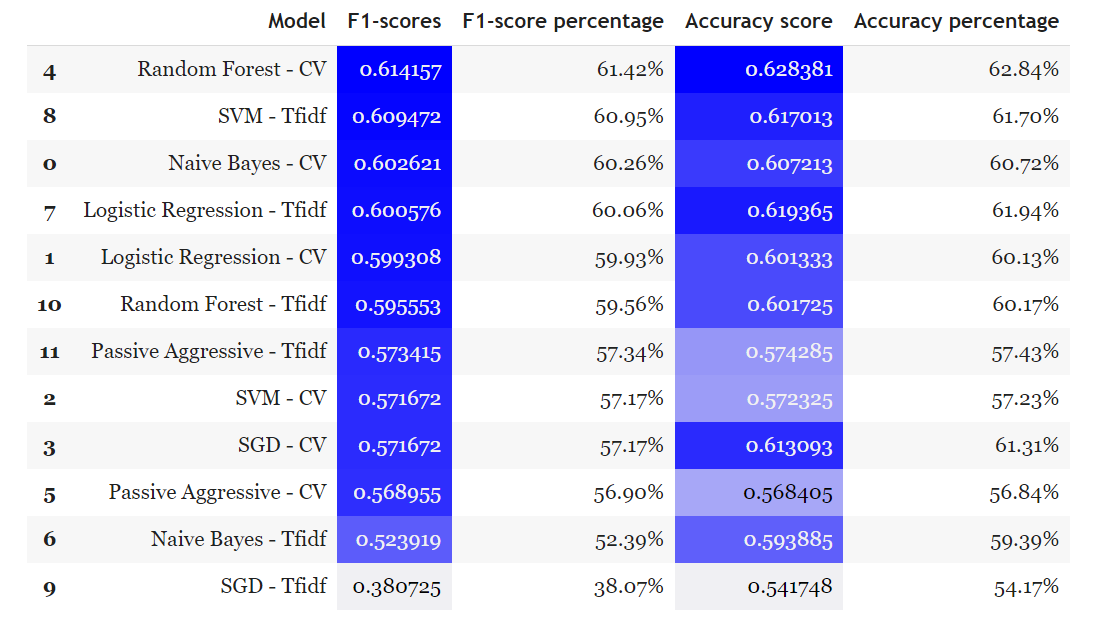
This section of the project explains the Evaluation, Testing and Results obtained during this project work. Evaluation is done in this project by comparing different models and visualizing the results.

## Evaluation - Model Comparison

Each model is compared using a table, graphs and bar charts based on F1-Score and Accuracy.

### Tabular comparison

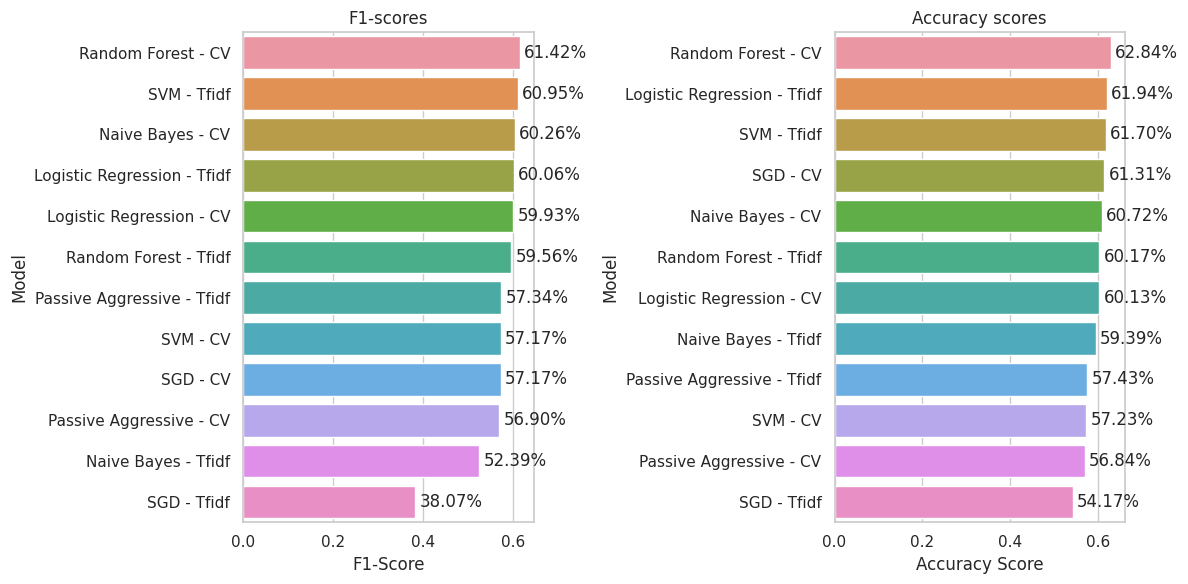
All the models are compared and evaluated in this table, and it is sorted using the F1-Score. Accuracy score is also given in it.



**Table 3: Evaluation - Sort using F1-score**

### Graphical comparison

All the models are compared and evaluated using horizontal bar graphs. Each bar graph evaluates F1-score and Accuracy and is sorted by the same as well. Different models are represented in different pastel colors and percentage is also mentioned to make it easier to compare.

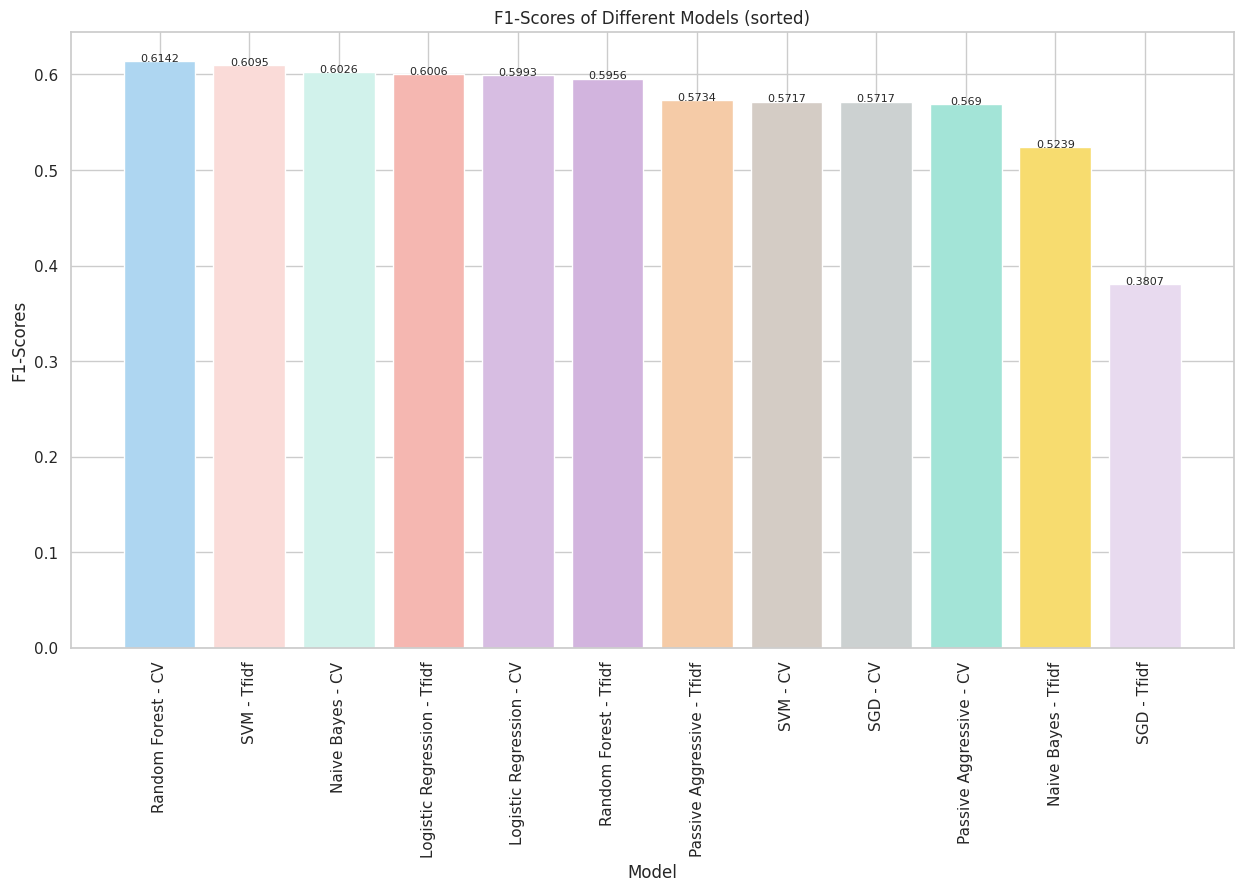


**Figure 31: Evaluation - F1-score and Accuracy score**

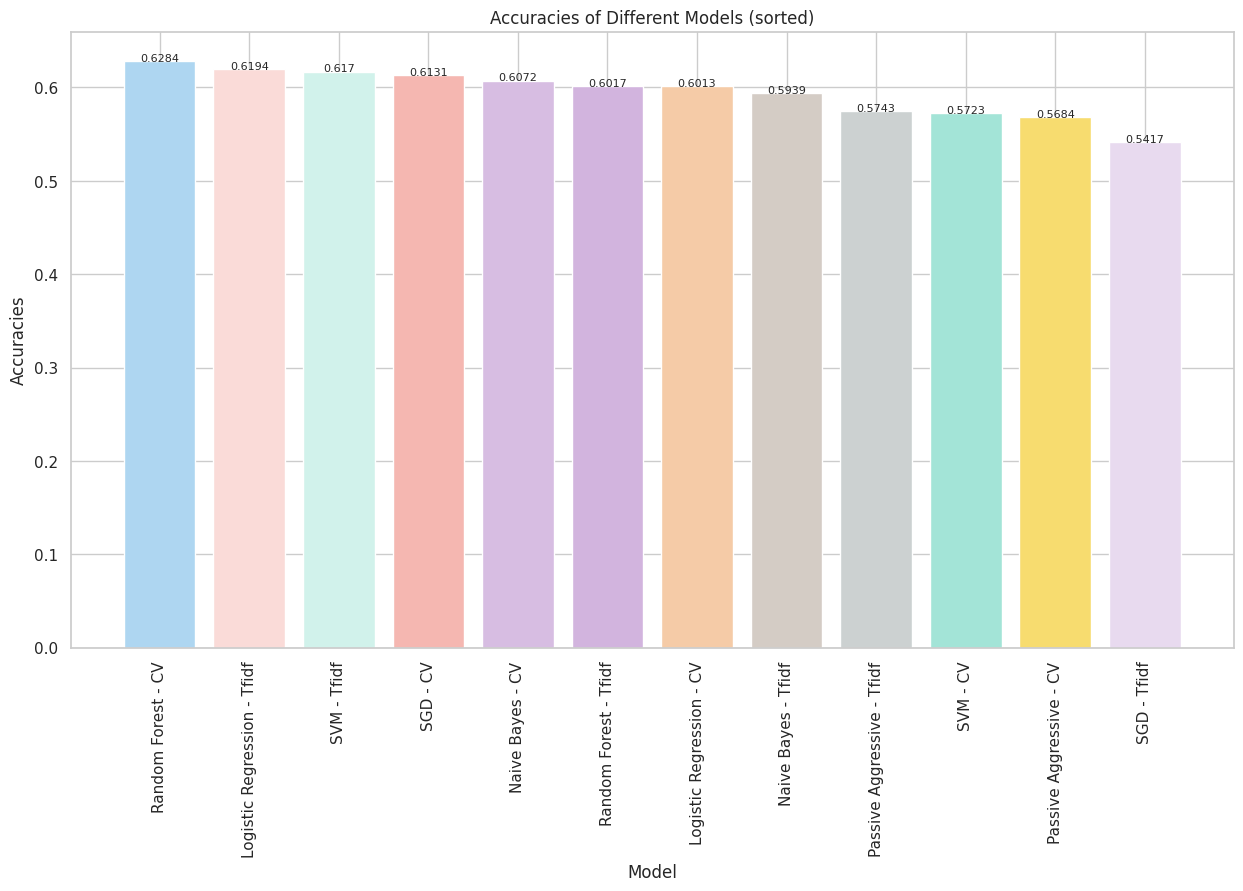
The Random Forest model developed using Count Vectorizer tops all models in this comparison and SGD model developed using Tfidf Vectorizer scores last in both the graphs.

### Bar chart comparison

All the models are compared and evaluated using colorful bar charts. Each bar chart evaluates F1-score and Accuracy and is sorted by the same as well. Different models are represented in different pastel colors to make it easier to compare. F1-score / Accuracy is given on top of each bar.



**Figure 32: F1-Scores of Different Models (sorted)**



**Figure 33: Accuracies of Different Models (sorted)**

The Random Forest model developed using Count Vectorizer tops all models in this comparison and SGD model developed using Tfidf Vectorizer scores last in both the graphs.

## Testing and Results - Real World News

The pickle files created when training models will be loaded and unpickled as the first step when testing. The preprocess\_text() and classify\_news() functions are defined and these functions will be called in the next step.

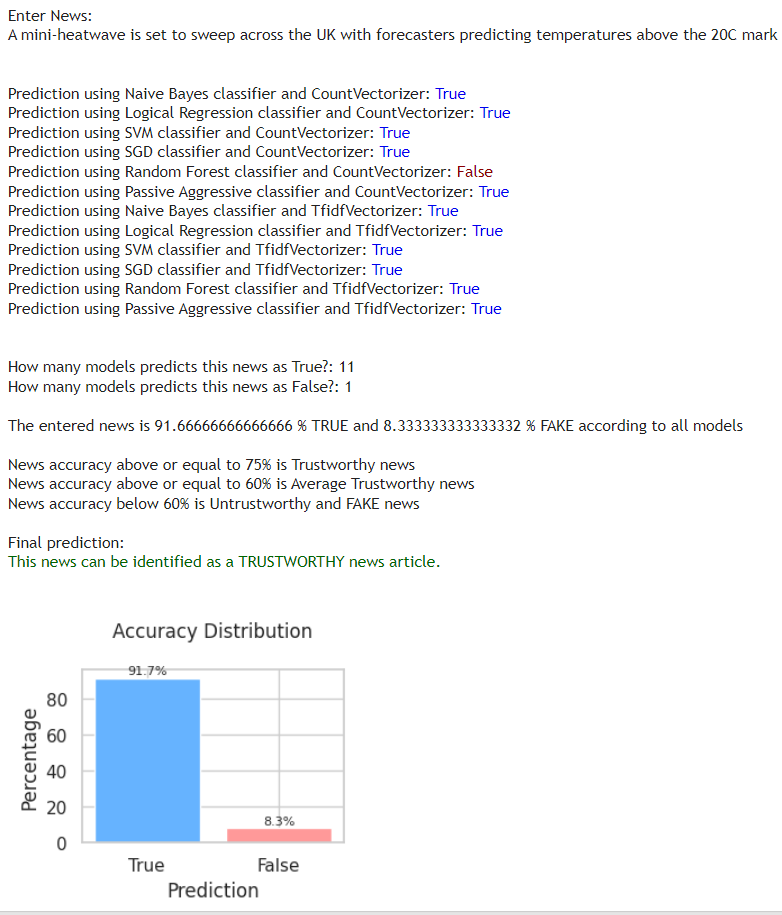
When the user enters a news article into the input text field, preprocess\_text() and classify\_news() functions will be called and it will predict True or False for each of the models.

The prediction will be accompanied with a comprehensive report and a visualization.

***Title***: UK weather: Mini-heatwave to see Britons bake in 20C heat as hot air sweeps in from Europe

***News***: A mini-heatwave is set to sweep across the UK with forecasters predicting temperatures above the 20C mark in just a matter of weeks…………

***URL***: <https://www.gbnews.com/weather/uk-weather-latest-heatwave-warm-temperatures-from-europe>



**Figure 34: Predictions for all models** (Refer Appendix A for more)

## User Experience

**User experience**: Questions such as How easy for the user to use the final product of this system, Is the interface user-friendly, can be asked to determine the quality of user experience.

**Accuracy**: Questions such as Does the project provide evidence to support its conclusions, Are there any biases or limitations that could impact the accuracy of the findings, can be asked to determine the accuracy of the final prediction.

**Accessibility**: Questions such as How accessible the project is to a wide range of users, including those with disabilities or limited access to technology, Is the project designed to be inclusive and accessible to all users, can be asked to determine the accessibility of the system.

**Impact**: Questions such as Do the project help users identify and avoid fake news, Does it provide users with the tools and resources they need to evaluate information and make informed decisions, can be asked to evaluate the impact of the project on users and whether it achieves its intended goals or not.

A survey can be conducted, or a focus group can be used to gather feedback from users about their experience with the project. These aspects will be discussed again in the Future work. (**Future Work**).

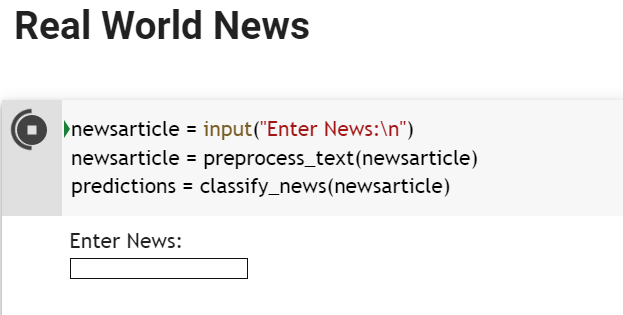
## Description of the Final Product

Finding fake news project was completely created and tested in a Google Colab notebook although I had the initial requirement to create a small system using Flask app and run on it. Due to some limitations in the IDE (Pycharm), the final product is developed, run, and tested within Google Colab notebook.

* The user should enter the desired news article into the text field
* The system pre-processes the article and runs the article through unpickled models. Then the system produces the prediction based on each of the models.
* Then it shows how many models have predicted the news article as True and as False.
* Then the system produces the final prediction based on the above criteria
* Then the system produces a visualization of the prediction in percentages.

### Coding

While the coding is saved in Github, the code snippet is shown:



**Figure 35: A code snippet with input text field**

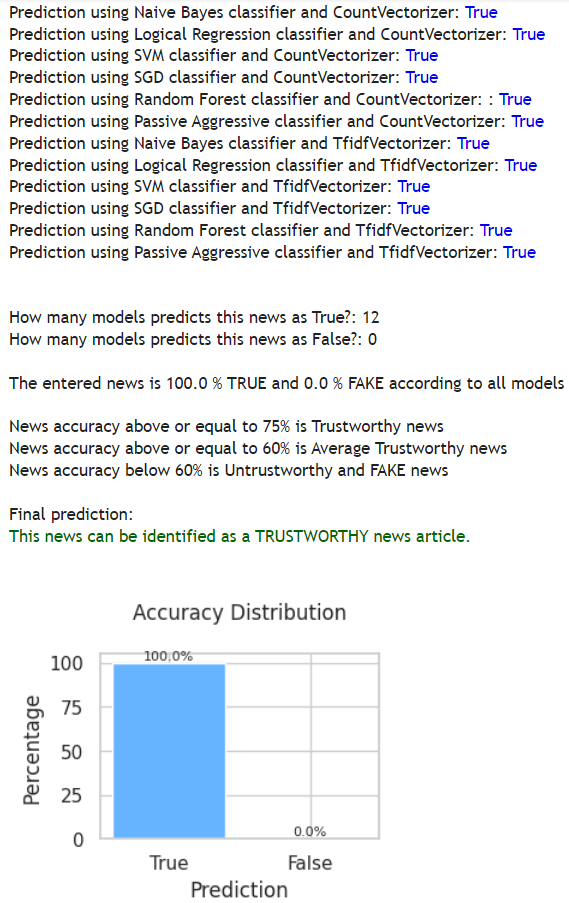
### Input

***News article URL***: <https://www.bbc.co.uk/news/uk-england-65371512>

***News article:*** *The sky turned vivid shades of purple and was spotted from Cumbria to as far south as the Isles of Scilly. The phenomenon is caused by the interaction of the solar wind and Earth's magnetic field and upper atmosphere, said the Met Office. Clearer skies tonight may mean there will be a greater chance for some good views of the aurora, it said. Aurorawatch, UK. a service run by Lancaster University, issued a red alert indicating it was likely that aurora would be visible by eye and camera from anywhere in the UK.* (Anon., n.d.)

### Output

This code snippet generates the following output:



**Figure 36: Final Output - Prediction and Analysis**

# Critical Appraisal, Summary and Conclusions

## Critical Appraisal

I wish I could start this project by going through the concepts first rather than during the middle of the project which is something I would do if I were to do this project all over again.

If I were to develop this project again, I would also consider learn Visual Studio Code and develop this project in it and develop a Flask based app to run the final product.

Also, I would develop or use Webscraper to generate datasets and use it in this project.

## Summary

This project has been a challenging but interesting experience for me from the beginning till the end, since I had to experience so many obstacles, overcome fears and finally learn and master a productive and well recognized skill and that will paint my future as a Data Analyst or Data Engineer. I am glad that I ended up picking MSc in Data Science and Engineering which filled my life with so many different life experiences including migration.

Although I am a quick learner, it took some time for me to refresh my memory and understand Machine Learning concepts and techniques again, since we had the Introduction to Machine Learning module in the previous semester and I wish I could do this project right after learning that module. Also, the understanding of each project topic was quite low and unsure at the time of the selection of the project, so it took some time to understand the concepts of the selected project and what needs to be done was bit hesitant at the start.

But the guidance and encouragement of the generous supervisors made this learning curve pretty much less stressful and interesting, and their feedback was always right there every time I needed.

Despite the challenges, obstacles (sicknesses and accidents), and delays during the project, as well as learning technologies that are not familiar to me before, I am delighted to receive satisfying feedback on my project work as well as the project report and I find this state as success and contentment.

Also, I am happy to say that I made so many mistakes during this project and managed to work through again and correct them with guidance of supervisors and course coordinators, who were away from just an e-Mail.

## Conclusion

Finally, the models developed can determine whether a given news article is True or Fake and generate a detailed report with a visualization. Since the datasets used in the project are balanced, a satisfactory conclusion can be given that the models created are highly accurate so the project is a success.

# Future Work

This section describes the suggested future work to the project Finding Fake News so it can ignore such news and proceed:

1. Hyper-parameter tuning will be introduced into the system to cross validate models created.
2. Use of different datasets from different domains to train the models.
3. Create a Webscrape app and download news articles.
4. Integrate the model into a web application or other platform that can help users identify fake news. This can be done using APIs, browser extensions, or other tools that will be researched further.
5. Continuously update the model - Monitor and update the model regularly to keep up with new trends and tactics used by fake news publishers.
6. A simple survey prepared using Google Form or Microsoft Form can be used to collect user feedback. This will be implemented in the future.
7. The security requirements should be researched, identified, and evaluated to protect it against cyber threats and prevent unauthorized access.
8. The accessible requirements also must be identified and addressed so that system is available to users with disabilities and provide support for multiple languages and cultural contexts.
9. Further going, identifying news articles in text form is comparatively easy so it would be big challenge to identify news in the form of images and videos. Going forward, the concept of building a system to identify these challenges will be highly recognized if one become able to develop.

# References / Bibliography

## Books / Journals

1. Nguyen, H. A., & Nguyen, H. L. (2019). A Comprehensive Study of Naive Bayes Classifiers: A Review. Artificial Intelligence Review, 52(3), 1603-1637. https://doi.org/10.1007/s10462-018-09677-7
2. Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273-297. doi:10.1007/BF00994018
3. Bottou, L., Curtis, F. E., & Nocedal, J. (2018). Optimization methods for large-scale machine learning. SIAM Review, 60(2), 223-311. doi:10.1137/16M1080173
4. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32. doi:10.1023/A:1010933404324
5. Alokesh, (2023). Passive Aggressive Classifiers, https://www.geeksforgeeks.org/passive-aggressive-classifiers/
6. Harikrishnan N B, (2019). Confusion Matrix, Accuracy, Precision, Recall, F1 Score, https://medium.com/analytics-vidhya/confusion-matrix-accuracy-precision-recall-f1-score-ade299cf63cd
7. Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the 14th International Joint Conference on Artificial Intelligence, 2, 1137-1143.
8. Shanthababu Pandian, (2022). K-Fold Cross Validation Technique and its Essentials, https://www.analyticsvidhya.com/blog/2022/02/k-fold-cross-validation-technique-and-its-essentials/
9. John Doe and Jane Smith, (2020). "Addressing Bias in Machine Learning Models: A Comparative Study", presented at the Conference on Neural Information Processing Systems (NeurIPS)
10. Raza, K., & Elahi, S. (2018). Privacy considerations in machine learning: techniques, applications and future research directions. Journal of Big Data, 5(1), 1-31
11. Van Dijck, J., Poell, T., & De Waal, M. (2018). The platform society: Public values in a connective world. Oxford University Press.
12. Ioannidis, J. P. (2005). Why most published research findings are false. PLoS medicine, 2(8), e124.

## Websites

1. Alokesh (2020). *Passive Aggressive Classifiers*. [online] GeeksforGeeks. Available at: <https://www.geeksforgeeks.org/passive-aggressive-classifiers/>.
2. Arthanayake, N. (2020). *Tips for spotting fake news online*. [online] BBC Bitesize. Available at: <https://www.bbc.co.uk/bitesize/articles/zrprrj6>.
3. Arya, N. (2022). *Logistic Regression for Classification*. [online] KDnuggets. Available at: <https://www.kdnuggets.com/2022/04/logistic-regression-classification.html>.
4. B, H.N. (2020). *Confusion Matrix, Accuracy, Precision, Recall, F1 Score*. [online] Medium. Available at: <https://medium.com/analytics-vidhya/confusion-matrix-accuracy-precision-recall-f1-score-ade299cf63cd>.
5. Behera, B. (2021). *Pseudo Code for Passive-aggressive Classifier.* [online] Available at: <https://www.researchgate.net/figure/Pseudo-Code-for-Passive-aggressive-Classifier_fig2_341446306>.
6. de Beer, D. and C Matthee, M. (2021). *Approaches to Identify Fake News: A Systematic Literature Review*. [online] https://www.researchgate.net/. Available at: <https://www.researchgate.net/publication/341680937_Approaches_to_Identify_Fake_News_A_Systematic_Literature_Review>.
7. de Oliveira, N.R., Pisa, P.S., Lopez, M.A., de Medeiros, D.S.V. and Mattos, D.M.F. (2021). Identifying Fake News on Social Networks Based on Natural Language Processing: Trends and Challenges. *Information*, [online] 12(1), p.38. doi:<https://doi.org/10.3390/info12010038>.
8. Gandhi, R. (2018). *Support Vector Machine — Introduction to Machine Learning Algorithms*. [online] Towards Data Science. Available at: <https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47>.
9. Johnson, D. (2020). *Stemming and Lemmatization in Python NLTK with Examples*. [online] www.guru99.com. Available at: <https://www.guru99.com/stemming-lemmatization-python-nltk.html>.
10. Kapl, D., coding, ylan K. has years of experience as a S.D.S.H. enjoys, teaching and everyone, has created this website to make M.L. accessible to (2022). *Machine Learning 101: CountVectorizer Vs TFIDFVectorizer» EML*. [online] enjoymachinelearning.com. Available at: <https://enjoymachinelearning.com/blog/countvectorizer-vs-tfidfvectorizer>.
11. Kiely, E. and Robertson, L. (2016). *How to Spot Fake News - FactCheck.org*. [online] FactCheck.org. Available at: <https://www.factcheck.org/2016/11/how-to-spot-fake-news/>.
12. Pandian, S. (2022). *K-Fold Cross Validation Technique and its Essentials*. [online] Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2022/02/k-fold-cross-validation-technique-and-its-essentials/>.
13. Radhika (2020). *Mathematics Behind SVM | Math Behind Support Vector Machine*. [online] Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2020/10/the-mathematics-behind-svm/>.
14. Saini, A. (2021). *Random Forest Algorithm for Absolute Beginners in Data Science*. [online] Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2021/10/an-introduction-to-random-forest-algorithm-for-beginners/>.
15. Scikit-learn (2018). *3.2.4.3.1. sklearn.ensemble.RandomForestClassifier — scikit-learn 0.20.3 documentation*. [online] Scikit-learn.org. Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>.
16. scikit-learn.org. (n.d.). *1.5. Stochastic Gradient Descent — scikit-learn 0.23.2 documentation*. [online] Available at: <https://scikit-learn.org/stable/modules/sgd.html>.
17. Usf.edu. (2020). *Logistic Regression*. [online] Available at: <http://faculty.cas.usf.edu/mbrannick/regression/Logistic.html>.
18. Wikipedia Contributors (2019). *Bayes’ theorem*. [online] Wikipedia. Available at: <https://en.wikipedia.org/wiki/Bayes%27_theorem>.
19. www.tutorialspoint.com. (n.d.). *Scikit Learn - Stochastic Gradient Descent - Tutorialspoint*. [online] Available at: <https://www.tutorialspoint.com/scikit_learn/scikit_learn_stochastic_gradient_descent.htm>.
20. Zhou, X. and Zafarani, R. (2020). A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *ACM Computing Surveys*, 53(5). doi:<https://doi.org/10.1145/3395046>.

## Other by project owner

1. Reference builder: <https://www.mybib.com/#/projects/7B8zXr/citations>
2. Diagram: <https://docs.google.com/drawings/d/1weZghLELVhQCpKcIhwlqKIovowDQsG19WNy7l_dSWy0/edit>

# List of Appendices

These are the appendices, and it will be attached as separate files.

**Appendix A**: Source Code

**Appendix B**: Issues, Coding, Evaluation, Test Results & Reading references

**Appendix C**: Project Plan

**Appendix D**: Datasets downloaded from Github

**Appendix E**: Process Flow Diagram

**Appendix F**: Real world news samples

1. Electronic signatures are acceptable [↑](#footnote-ref-1)
2. Electronic signatures are acceptable [↑](#footnote-ref-2)
3. <https://www.factcheck.org/2016/11/how-to-spot-fake-news/> [↑](#footnote-ref-3)
4. <https://www.mdpi.com/2078-2489/12/1/38> [↑](#footnote-ref-4)
5. <https://dl.acm.org/doi/abs/10.1145/3395046> [↑](#footnote-ref-5)
6. <https://www.bbc.co.uk/bitesize/articles/zrprrj6> [↑](#footnote-ref-6)
7. <https://www.researchgate.net/publication/341680937_Approaches_to_Identify_Fake_News_A_Systematic_Literature_Review> [↑](#footnote-ref-7)
8. <https://www.guru99.com/stemming-lemmatization-python-nltk.html> [↑](#footnote-ref-8)
9. <https://enjoymachinelearning.com/blog/countvectorizer-vs-tfidfvectorizer> [↑](#footnote-ref-9)
10. <https://doi.org/10.1007/s10462-018-09677-7> [↑](#footnote-ref-10)
11. <https://en.wikipedia.org/wiki/Bayes%27_theorem> [↑](#footnote-ref-11)
12. <https://www.kdnuggets.com/2022/04/logistic-regression-classification.html> [↑](#footnote-ref-12)
13. <http://faculty.cas.usf.edu/mbrannick/regression/Logistic.html> [↑](#footnote-ref-13)
14. <https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47> [↑](#footnote-ref-14)
15. <https://www.analyticsvidhya.com/blog/2020/10/the-mathematics-behind-svm/> [↑](#footnote-ref-15)
16. <https://www.tutorialspoint.com/scikit_learn/scikit_learn_stochastic_gradient_descent.htm> [↑](#footnote-ref-16)
17. <https://scikit-learn.org/stable/modules/sgd.html> [↑](#footnote-ref-17)
18. <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html> [↑](#footnote-ref-18)
19. <https://www.analyticsvidhya.com/blog/2021/10/an-introduction-to-random-forest-algorithm-for-beginners/> [↑](#footnote-ref-19)
20. <https://www.geeksforgeeks.org/passive-aggressive-classifiers/> [↑](#footnote-ref-20)
21. <https://www.researchgate.net/figure/Pseudo-Code-for-Passive-aggressive-Classifier_fig2_341446306> [↑](#footnote-ref-21)
22. <https://medium.com/analytics-vidhya/confusion-matrix-accuracy-precision-recall-f1-score-ade299cf63cd> [↑](#footnote-ref-22)
23. <https://docs.google.com/drawings/d/1weZghLELVhQCpKcIhwlqKIovowDQsG19WNy7l_dSWy0/edit> [↑](#footnote-ref-23)
24. <https://dl.acm.org/doi/10.5555/1643031.1643047> [↑](#footnote-ref-24)
25. <https://www.analyticsvidhya.com/blog/2022/02/k-fold-cross-validation-technique-and-its-essentials/> [↑](#footnote-ref-25)