Fake news detection using machine learning and deep learning, on multi-source data.

HRCOS82

63017822 – Brandon Ashley

Supervisor: Professor Etienne Van der Poel

Assessment 4 code: 657452

Abstract – Fake news can be seen daily on social media. The misleading information we see affects us before we can tell if it is true or false. This study will use machine and deep learning algorithms on fake news datasets to define a model that can detect fake news from multiple sources. The experiment used data cleaning, feature creation and machine learning models like XGBoost, Feedforward Neural Networks, logistic regression, Random Forest, LightGBM and CATBoost. The accuracy of the models amounted to 86%, 88%, 85%, 84%, 87%, and 86.9% respectively. These outcomes suggest that despite slightly lower accuracies, the models performed better than expected, demonstrating robust performance.

Keywords: fake news, machine learning, deep learning, feature creation, XGBoost, Logistic Regression, Random Forest, CATBoost, LightGBM, Feedforward Neural Network.

Table of Contents

Table of tables

Table of figures

1 Chapter 1

1.1 Introduction and Background

1.1.1 Introduction

In this paper, we investigate the accuracy of algorithms on fake news datasets. Fake news is characterised as misinformation masked as facts (Lazer *et al.*, 2018; Narwal, 2018). Social media usually spreads fake news, and individuals or organisations are not held to the same accountability as mainstream media (Lazer *et al.*, 2018; Waweru Muigai, 2019). In the thriving social media age, fake news can spread like wildfire. Due to the increased spread and damage that fake news inflicts, it is essential to curb it. Fake news draws social media users' attention, provokes people and gains profit through clicks (Narwal, 2018).

This paper will use algorithms to detect fake news on a dataset and then determine the performance of the algorithms using the accuracy metric. The research will focus on creating a dataset using fake news datasets. The datasets will be cleaned and processed to ensure the best possible results. Algorithms will be chosen to detect and label the test data. The performance of these algorithms will be used to calculate the accuracy of the models. This research will show the effects of having data from different sources on the training of models. The contribution this paper will provide is that future researchers can use the features and algorithms of this paper as a baseline for subsequent research and findings.

1.1.2 Background

A reoccurring theme in the literature was that machine learning and deep learning were crucial in detecting fake news. The gaps that the review showed were that the datasets used were small in size and clustered into specific sources. The datasets focused on certain sources, such as Twitter (alternatively known as X) data or media headlines. In order to create a method to determine fake news automatically, it will be essential to create a machine learning or deep learning algorithm that can make predictions on all possible data.

This study will conduct a literature review using a systematic review. A systematic review investigates a research question by finding and analysing previous studies in a specific domain (Staples and Niazi, 2007). By using a systematic review, the literature will help guide the study's methodology and procedures. For example, the data used features, algorithms, and evaluation metrics.

1.2 Problem statement

Fake news can be found everywhere on emails, WhatsApp, Twitter, and many sites on the internet. Fake news is when a person or persons read an article disguised as the truth or has information that is not “fully” true (Ferreira, Robertson and Kirsten, 2020). However, the information in the article is either false or misleading.

Fake news has a harmful impact on society as a whole. People find these fake articles and assume it is true without a second thought. The situation is made worse because most people share these articles on social media (Belova and Georgieva, 2018). Even more important, these shared articles become even more popular, thus spreading fake news.

Fake news is usually started by someone that wants to provoke or manipulate others into believing something that is not true (Waweru Muigai, 2019). Certain words, descriptions or ideas are used in fake news to inflame the readers. As a result, the reader is angered by these ideas that are not true. However, the reader believes them to be true. The effect of fake news creates a cycle of misinformation where a fake article hooks the reader, and then the reader continues to read fake articles about the subject.

People share more fake news than authentic or factual news (Narwal, 2018; Madrid, 2023). People share fake news with their close friends and family because the misinformation resonates with them. Due to the intimate and trusting relationships between these individuals, close friends and family are more likely to rate the information as true and trustworthy. This leads to fake news reaching a larger audience due to being continuously spread by individuals, often using social media (Narwal, 2018). In order to decrease the spread and hold of fake news, a way needs to be found to inform people and readers that news is fake, and that certain information might be unreliable.

1.2.1 Main research question

Detecting fake news using machine learning and deep learning techniques is a widely researched field. An area of further research is using datasets from various sources and allowing the algorithm to train on all possible data. Furthermore, determining the best features to use in fake news detection is required for the best results. Finally, evaluate the accuracy of the algorithms to showcase the performance.

What is the performance of different machine learning and deep learning techniques for detecting fake news on a heterogeneous dataset composed of data from various sources, considering both evaluation metrics and dataset characteristics?

1.2.1.1 Research sub questions

1.2.1.1.1 Which machine learning or deep learning techniques are best suited for a fake news detection algorithm?

This question aims to find the machine and deep learning techniques with the best accuracy and performance in detecting fake news. Currently, numerous models can be used to find solutions to many problems. However, by selecting the best algorithm for the problem of fake news detection, the research can focus on implementing other data and discovering the best-suited algorithm for various data.

1.2.1.1.2 What are the key considerations for a fake news dataset?

This question aims to find the different factors that affect a dataset in a fake news detection algorithm. The features that help identify if the information is fake news can continuously change as we use machine-learning approaches (Manzoor, Singla and Nikita, 2019). Finding the right features for a dataset; features allow the algorithm to learn and predict at a higher accuracy. Just as important are the data sources, format, and dataset size. For this reason, it is essential to understand the data that needs to be collected for the dataset and which information is essential.

1.2.1.1.3 What is the influence of diverse data sources on the performance and accuracy of a fake news algorithm?

This question sheds light on a diverse dataset's effects on a fake news algorithm. When considering a dataset, where the data is only created out of Twitter data, an algorithm might perform better because of some features or information only available on Twitter. However, the same algorithm might perform poorly on another dataset. For this reason, it is vital to assess the effect of diverse data from diverse sources on current algorithms' performance and accuracy.

1.2.2 Research objectives

First, obtain the best-suited models for a fake news detection algorithm. Evaluate the performance of previously used algorithms to determine the most successful algorithms. This study should test these models on data they have not been tested on before and determine their performance.

Second, assess the datasets that have previously been used and establish the critical factors of these datasets to understand why they increase performance. Furthermore, when creating a dataset of different sources and sizes, it is essential to determine the best features and structure for the algorithm.

Third, assess the impact that a diverse dataset might have on the performance and accuracy of algorithms. Using data from different sources like Twitter or mainstream media and different structures could influence the performance of fake news detection. Recognising the shortcomings of multiple sources could increase the future performance of fake news detection. This study should highlight the crucial factors of data collection and algorithms' implementations on datasets from various sources.

1.2.3 Outcomes

After the research has concluded, this study will show how different algorithms perform on a dataset with a wide variety of data. In addition, the performance of such algorithms is vital for future work to understand the effects of more diverse datasets and the use of different models.

2 Chapter 2

2.1 Literature review

2.1.1 Introduction

The literature will be reviewed using a systematic review. A systematic review is used to answer the research questions using previous studies (Staples and Niazi, 2007). Moreover, the researcher can answer the research questions by finding the research structure and gaps that literature has and identifying possible future research. An essential purpose of a systematic review is to ensure a full grasp of previous research to prevent current research from being repeated (Pollock and Berge, 2018). By understanding the research, other researchers will allow this study to focus on the gaps and avoid repetitive work in the same field.

Journals and articles were searched for on Google Scholar and UNISA Library to find the research of interest. The research focused on fake news, with particular attention to what fake news is and fake news algorithms—using the research question and sub-questions to focus the search on Google Scholar and UNISA Library. Keywords were combined to find the best results. Search phrases were used like “What is fake news”, “Fake news algorithms”, “machine learning + fake news”, and “deep learning + fake news”. Furthermore, using the IEEE site, keywords like “fake news” and “deep learning” and the date set to 2020 and 2023 were used to find journals of interest. Most searches were conducted up to the 5th page. However, some searches required more pages to find relevant journals.

2.1.2 Topic of interest

2.1.2.1 The use of machine learning and deep learning algorithms in the detection of fake news.

This research aims to determine the use of machine learning and deep learning algorithms in fake news detection. The basis of this research is divided into subtopics, namely, the algorithms chosen, the performance of these algorithms, and the datasets used.

An essential aspect of the research depends on the algorithms used in previous research. There are many different machine and deep learning algorithms that can be used. Thus, using the research already conducted, the best algorithms can be chosen to improve on and be further tested. Algorithms of interest are mainly natural language processing (NLP), which will be used to create the dataset and improve the data used. Various algorithms can then be used to make predictions on fake news, for example, Random Forest (RF), K-Nearest (KNN), Logistic Regression (LR), Support vector machines (SVM), Naïve Bayes (NB), Decision trees (DT), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), Artificial Neural networks (GAN) and XGBOOST (XGB). Depending on the performance of these algorithms, the best will be used to make predictions on a dataset.

2.1.2.2 What datasets and features were used?

Additionally, the researcher aims to determine the datasets and features used in previous research. During the research, it is essential to identify the size, format, features and data used in the dataset. In order to get the best performance out of an algorithm, the dataset used needs to be cleaned, meaning no missing data, duplicates and features are labelled. Furthermore, it is vital to determine the features that will be used. Existing research shows that the best features can be used in future datasets to improve data collection. In this literature review, all the dataset information must be found to determine the most suitable features. Whilst researching fake news and datasets, the pre-processing needs to be investigated; this will aid in creating datasets and preparing them for the algorithms.

2.1.3 Main themes

2.1.3.1 Algorithms employed by past research.

Machine learning and deep learning algorithms play an essential role in detecting fake news. However, it is difficult to determine which algorithm works best as many of the algorithms perform well, dependant on the data used (Varma *et al.*, 2021). The accuracy of algorithms was high for much of the literature reviewed see Table 1.

During the literature review, both machine and deep learning algorithms were used. The machine learning approaches were K-nearest neighbours (KNN), Logistic regression, Support vector machine, Naïve Bayes, XGBOOST, AdaBoost, random forest, Term Frequency-inverse document frequency (TF-IDF), and decision trees. The deep learning approaches were CNN, LSTM, Bidirectional Encoder Representations (BERT), and

Geometric Deep learning. All these models are represented in Table 1; not all models are represented in the table.

Algorithm used	**Source**	**Accuracy**
Random forest	(Khanam *et al.*, 2021)	73%
	(Ahmad *et al.*, 2020)	80%
	(Nasir, Khan and Varlamis, 2021)	92% (ISOTdataset)
	(Abdulrahman and Baykara, 2020)	92.18%
K-nearest	(Khanam *et al.*, 2021)	70%
	(Ahmad *et al.*, 2020)	67%
	(Khan *et al.*, 2021)	70%
KNN+LSTM or ISOT dataset	(Nasir, Khan and Varlamis, 2021)	98% (ISOTdataset)
	(Abdulrahman and Baykara, 2020)	90.12%(+LSTM)
Logistic regression	(Ahmad *et al.*, 2020)	92%
	(Khan *et al.*, 2021)	76% (Combined Corpus)
	(Nasir, Khan and Varlamis, 2021)	52% (ISOTdataset)
	(Abdulrahman and Baykara, 2020)	89.14% (Count Vector)
CNN	(Mishra, Shukla and Agarwal, 2022)	93% (Corpusdataset)
	(Wani *et al.*, 2021)	94%
	(Khan *et al.*, 2021)	93%
	(Nasir, Khan and Varlamis, 2021)	99% (ISOTdataset)
	(Abdulrahman and Baykara, 2020)	100%(+LSTM)
Term Frequency - inverse document frequency (TF-IDF)	(Thota *et al.*, 2018)	94.31%
	(Al Asaad and Erascu, 2018)	95.6%
BERT	(Wani *et al.*, 2021)	98.41%
	(Kaliyar, Goswami and Narang, 2021)	98.90%
Naïve Bayes	(Khanam *et al.*, 2021)	67%
	(Khan *et al.*, 2021)	93%

Algorithm used	Source	Accuracy
	(Abdulrahman and Baykara, 2020)	85.87%
XGBoost	(Khanam *et al.*, 2021)	75%
	(Ahmad *et al.*, 2020)	94%
	(Abdulrahman and Baykara, 2020)	91.39%
Decision Tree	(Khanam *et al.*, 2021)	68%
	(Ahmad *et al.*, 2020)	94%
	(Khan *et al.*, 2021)	67%
Geometric deeplearning	(Monti *et al.*, 2019)	92.70%
LSTM	(Khan *et al.*, 2021)	93%
	(Kannan *et al.*, 2021)	95%
AdaBoost	(Abdulrahman and Baykara, 2020)	100%
SVM	(Khanam *et al.*, 2021)	73%
	(Khan *et al.*, 2021)	71%

Table 1:The accuracy of various method used in the literature to detect fake news.

The models that performed the overall best were deep learning models like CNN, LSTM, and BERT, all receiving an accuracy of no less than 90%. The models were trained on different data and using different features, thusly the accuracies differ.

2.1.3.2 Datasets and features utilised by past research.

Information processing is essential before a model can train on the data (Wani *et al.*, 2021). Preprocessing helps the model understand the meaning of words and the patterns crucial in solving a given problem. The types of preprocessing steps that can be used in a fake news problem are removal of HTML tags, removal of non-characters, nominalisation, removal of stop words and stemming (Al-Ahmad *et al.*, 2021). Removing all unnecessary data and information from the dataset helps the model to focus on the critical information. This includes the sources' names and the writer's email address (Gilda, 2017). Removing information that has a chance of being repeated in other data with no context will help improve a model's accuracy by not being able to map the incorrect data.

Many datasets were encountered during the research, some of which were Liar, fake or real, combined corpus and BuzzFeed datasets (Shu *et al.*, 2017; Khan *et al.*, 2021). Liar is a political dataset with over twelve thousand user-labelled instances. Fake or real news and BuzzFeed datasets are US election datasets. A combined corpus is a dataset created by Khan et al. (2021) that consists of multiple fake news sources and numerous trusted sources. The combined dataset contained eighty thousand entries.

D'Ulizia et al. (2021) did a review of 27 datasets that are widely accepted by other works of literature. They found that most datasets are in English and are concerned with political news. This is a problem since English is not the only language used in the world, which means a vast amount of data is not being used. Second, datasets used for fake news detection consist of less than ten thousand articles or entries. Third, data in a dataset is only created as true or false. Fourth, data is provided by well-known or mainstream media (D'Ulizia *et al.*, 2021).

The features used in a fake news dataset are one of the most integral aspects. Models use features to map the correlation between information and aid the model in predicting fake news. Sahoo and Gupta (2021) defined features for social media as user or news content. The use of both these categories helped the models perform more accurately. The following user content was used Profile ID, Profile name, date joined, Friends, Profile picture groups and pages joined or liked, news posts and the number of stories shared (Sahoo and Gupta, 2021). The news content was source, headline, body, text and images (Sahoo and Gupta, 2021). All models received a 98% or more when run on these features. The problem with these features is that not all datasets can contain all this information, making it difficult to merge with existing datasets. Baarir and Djeffal (2021) on the other hand, used authentic news sources. The features used were 5-word using a bag of words, 3 compound words, N-Gram, Term frequency-inverse Document frequency model (TF-IDF), date, feeling, source, author and the label of true or false (Al Asaad and Erascu, 2018; Baarir and Djeffal, 2021). Features extracted from the data are often possible and are a good starting point for a dataset. Kannan et al. (2021) used a dataset with a title, text, source, and label for fake or true. Furthermore, the features used are easy to implement on other sources. Seddari et al. (2022) determined that the "title, number of words, reading ease, lexical diversity, and sentiment" are the most relevant features.

The above features will be essential when creating a database using multiple sources. Depending on the data used, more relevant features can be created. When using feature reduction, it is essential to check the correlation between features and the information they hold. The use of feature reduction could have an effect on the outcome that a model has (Umer *et al.*, 2020). Feature reduction can remove important information or keep information that holds no correlation to one another.

2.1.4 Importance

During this research, it became clear that we have the technology and the ability to reduce the effects of fake news on society. With machine and deep learning models, we can detect fake news with more than 90% accuracy, depending on the model used. Social media consumers require the ability to perform disinformation and misinformation checks on the post, articles and messages they see (Dame Adjin-Tettey, 2022). Giving individuals the chance to examine the information they are given is vital in the fight against fake news.

Detecting fake news early and informing individuals of the misinformation will decrease the likelihood that fake news is shared and believed. Individuals share the news they find interesting, not knowing that it might not be true. Being able to inform people before they share misinformation is vital to stop the spread of a fake article or post.

For this reason, further research into the use of machine learning and deep learning is essential.

2.1.5 Summary of references

Concept	Number	Reference
Methodology	1.	(Staples and Niazi, 2007)
	2.	(Pollock and Berge, 2018)
Algorithms	3.	(Khanam *et al.*, 2021)
	4.	(Ahmad *et al.*, 2020)
	5.	(Nasir, Khan and Varlamis, 2021)
	6.	(Abdulrahman and Baykara, 2020)
	7.	(Mishra, Shukla and Agarwal, 2022)
	8.	(Wani *et al.*, 2021)

	9.	(Thota *et al.*, 2018)
	10.	(Al Asaad and Erascu, 2018)
	11.	(Kaliyar, Goswami and Narang, 2021)
	12.	(Khan *et al.*, 2021)
	13.	(Monti *et al.*, 2019)
Datasets	14.	(Gilda, 2017)
	15.	(Shu *et al.*, 2017)
	16.	(D'Ulizia *et al.*, 2021)
	17.	(Sahoo and Gupta, 2021)
	18.	(Baarir and Djeffal, 2021)
	19.	(Seddari *et al.*, 2022)
	20.	(Kannan *et al.*, 2021)
	21.	(Umer *et al.*, 2020)
	Prev.	(Al-Ahmad *et al.*, 2021)
	Prev	(Wani *et al.*, 2021).
Discussion/ Elaboration	22.	(Dame Adjin-Tettey, 2022)
	23.	(Mridha *et al.*, 2021)
	24.	(Varma *et al.*, 2021)
	Prev.	(Shu *et al.*, 2017)
	Prev.	(Nasir, Khan and Varlamis, 2021)

Table 2: Literature reviewed.

2.1.6 Literature gaps

The literature has shown that there has been extensive research on the models that have been used and the algorithms that produce the best results. Neural networks have the best performance. However, using traditional models in feature engineering is an important aspect (Nasir, Khan and Varlamis, 2021).

The journal of (Mridha *et al.*, 2021) outlined the challenges and the directions needed for future study:

1. Previous literature needed to focus on feature selection more which could improve the performance of neural networks.
2. Engineering new features like text and statistical features is vital in detecting fake news.
3. There is a need for more image and video fake news detection.
4. Receiving information from multiple sources and creating a dataset with more information could be beneficial.
5. There is a need for more data.

The amount of data currently available is not beneficial to any algorithm.
Shu et al. (2017) suggests that instead of using binary values for determining fake news probabilistic approach could be more beneficial. Furthermore, they suggest using ensemble methods to improve the detection. Ensemble models use multiple models to predict first using a weaker model. However, a second model with predictive power is used; the combination of both models increases the accuracy.
Research that was reviewed showed a lot of research focused on the algorithms and features. The research into algorithms that are best suited is extensive and thorough; however, this does not mean there is no room for improvement. The research that shows the features used and extracted is also numerous. The main concern shown by most research is the data that is used. During the research, it became evident that the datasets used needed to be bigger, or the data was focused on one source. We see multiple data sources from Twitter, WhatsApp, news articles and recommendations throughout our daily lives. Thus, it is vital to research the best way to allow an algorithm to learn from all possible sources. By finding the features and algorithms best suited for fake news detection, we can create larger datasets and find better ways of showing social media users when information is not true.

2.1.7 Conclusion

In conclusion, the research on the algorithms and datasets used in fake news detection is extensive. However, there is room for improvement regarding the data used, feature extraction and the algorithms used.

This study should focus on creating a dataset from various sources. Also, choose the best features and cleaning techniques for the data. Furthermore, testing various algorithms on the data and determining the best models for the detection of fake news.

3 Chapter 3

3.1 Research methodology

3.1.1 Methodology

3.1.1.1 Introduction

The purpose of the methodology is to find datasets and clean up and merge the chosen datasets. It is essential to create features as discussed in the literature review as well as to use algorithms that have been fine-tuned to determine if the news in the test set is fake news. Then determine the accuracy and performance of the algorithms.

3.1.1.2 Research design

The research will be conducted using the quantitative approach. Quantitative studies are meticulously organised and go through rigorous credibility checks. The procedures and environment must be defined and well documented when creating a quantitative research report. Thus, other researchers could replicate the study in the future (Kumar, 2011). As shown in the literature review, previous studies used preexisting datasets, some of which were also combined datasets (Khan *et al.*, 2021; Mishra, Shukla and Agarwal, 2022). Further, preprocessing techniques were used to clean data then features were created (Seddari *et al.*, 2022). Previous studies employed algorithms to determine fake news and then measured the algorithms in terms of accuracy (Kannan *et al.*, 2021; Wani *et al.*, 2021).

3.1.1.3 Datasets

This study will use datasets as the primary source of information or data. Datasets will be chosen depending on the type of data; the overall idea is to find data that can be used to create larger datasets. Furthermore, after datasets have been selected, preprocessing will start. First, datasets will be compared to determine the features being used. When datasets are correct and can merge, a new dataset or datasets will be created. Shuffling of the new datasets will be essential to reduce the bias of the sets as well as when the test and train sets are created. It is required that both test and train sets have information on all the datasets that were initially used. The new dataset will be subject to preprocessing. Removal of NAN (Not A Number), NAN is when a data value is missing. NAN could be solved by fillna, bfill, or row removal. Using fillna and bfill will not be advised for datasets as is in the context of fake news articles; all it will do is duplicate the data.

For this reason, the removal of the row will be best suited. Next, a dataset could be filled with quotes, question marks, and other non-character information. The algorithms will not be able to understand the purpose of these non-characters. For this reason, we will remove them. Before removal, they will be counted, and the number of non-characters holds significance in fake news detection. Non-characters can offer additional context increases in quotation marks, which might suggest that data is likely authentic. Moreover, an abundance of exclamation marks could indicate potential fake news. However, it will be discussed later in the study. During preprocessing, all characters must be lowercase letters; different cases overcomplicate the search of terms; thus, it is safer to convert all text to lowercase—next, the removal of stop words. Stop words are commonly used in text and speech, for example, 'the', 'a', and 'an'. These words produce no information but take up a lot of space in data (GeeksforGeeks, 2022). Furthermore, applying stemming to words. Stemming is when a bunch of words have the same root. For example, chocolate can be 'Choco', but the word 'choco' does not improve our data (Yash, 2022). Removing data and preprocessing will help the algorithms perform better and learn more effectively.

3.1.1.4 Variables and measures

Previous research showed that certain features work better than others (Mridha *et al.*, 2021). Creating features such as the sentence word count, sentence length, and number of punctuations will all provide important information about the text. Furthermore, the use of a Bag of words is a new feature. Bag of words is a Natural Language Processing (NLP) that checks the occurrences of words in a particular text. Bag of words does not care about the order of the sentences or where the words occur, only the frequency of the words (Dubey, 2018). 5 Bags of words will choose the five most used words in a specific text.

Additionally, the feature reading ease is another important subject. Reading ease or readability determines the reader's understanding of a particular text. It is determined by the text's length, syntax and complexity (sirjan13, 2021). Also, the lexical diversity of text is a feature. *Lexical diversity* is a score or measure given to a text that shows vocabulary diversity. For example, the number of unique words will improve the score.

Last but not least, sentiment is a feature of fake news. Sentiment analysis of the text shows how a person could "feel" when reading (Arora, 2022). The sentiment of a text could be positive or negative. Usually, a score is between -1 (very negative) and 1 (very positive).

3.1.1.5 Algorithms

The algorithms were chosen depending on how they previously performed and the respective models' advantages on datasets.

3.1.1.5.1 Logistic Regression

Logistic regression is a machine-learning technique to predict a binary variable (Menard, 2014). The model detects the relationship between a dependent variable and one or more independent variables. Furthermore, the model gives a probability value between 0 and 1 for a dependent variable prediction ;this is achieved by the sigmoid function (JavatPoint, 2018).

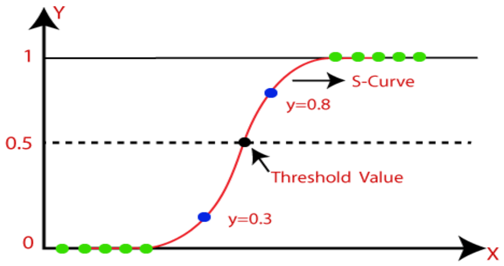


Figure 1: Logistic function (JavatPoint, 2018)

3.1.1.5.2 Random forest

Random forest is an ensemble learning method that uses the decision tree classifier. However, instead of one decision tree, the algorithm takes the majority votes of several trees to make a prediction (Javatpoint, 2021). The decision tree model can yield higher accuracies and reduces the risk of overfitting data (Javatpoint, 2021). Random forest does not consider the features split of all features; rather, it only uses a subset of the features.

3.1.1.5.3 XGBOOST

XGBOOST is a machine-learning algorithm that boosts the gradient. XGBOOST gives better results than other decision tree algorithms. The algorithm is enhanced by regularization and sparsity awareness: weighted quantile sketch and cross-validation (Morde, 2021).

3.1.1.5.4 Feedforward Neural Network (FNN)

A Feedforward Neural Network (FNN) is a simple artificial neural network consisting of an input, hidden, and output layer (Feed, no date). The input layer consists of the data frame features. The data is passed to the hidden layers that calculate the weights for each neuron. Furthermore, the neurons continue calculating the weights for all the hidden layers. The results are sent to the output layer, and the weights are updated based on the output error. An FNN uses the feedforward phase, and the weighted sum of the input is calculated until the final layer is reached. Also, the Backpropagation Phase calculates the error of the predicted and actual output. The error is sent back through the network to adjust the weights. This process is known as gradient descent and is performed until the network performs optimally (Feed, no date).

3.1.1.5.5 LightGBM

Light Gradient Boosting Machine (LightGBM) is a gradient boosting framework (AI, no date; Mondal, 2022). This model works well with large datasets to produce higher accuracy, faster training, low memory usage and parallel learning. LightGBM uses gradient boosting the same as XGBoost and CatBoost; gradient-based sampling is used to split the trees.

3.1.1.5.6 CATBoost

Categorical Boosting (CATBoost) is a gradient boosting algorithm. CATBoost builds a decision tree using gradient-based optimisation (Oppermann, 2023). A loss function reduces by using only features that reduce the loss. CATBoost reduces overfitting and can be used with categorical features.

3.1.1.5.7 Evaluation metric

In this study, the metrics used are accuracy, f1-score and confusion matrix. Each metric can have misleading results (Gupta, 2023). However, combining all the metrics will give a better overview of the performance of the model.

$$Accuracy = \frac{TrueNegatives + TruePositive}{TruePositive + FalsePositive + TrueNegative + FalseNegative}$$

Figure 2: Formula for accuracy (Erika D, 2019)

Accuracy shows the number of predictions a mode classifies as correct (Erika D, 2019). To calculate accuracy, you divide all the correct predictions (True negatives + True Positives) made by all the predictions made.

F1-Score combines recall and precision to determine the model's accuracy when classifying positive or negative labels (Wood, 2019). The metric indicates the balance between precision and recall and how well the model can predict positive or negative labels.

The confusion matrix shows the counts for True positive, false positive, true negative and false negative (Erika D, 2019). The confusion matrix can provide important information on why model predictions are incorrect. For example, when the amount of False Positive is high, we know the model is prediction positive; however, the actual class is negative. The matrix can show us where our model is going wrong.

3.1.2 Procedure

3.1.2.1 Experimental setup

The experiment was conducted on the Python distribution platform of Anaconda. Anaconda creates a new environment where new libraries can be used without being affected by previous programs. The interactive development environment (IDE) used was JupyterLab (3.63). JupyterLab allows the user to follow the code and receive real-time information about the executed code. In JupyterLab, Python 3 (ipykernal) was used to perform experiments on data and to run models. The following libraries were used: NumPy, pandas, matplotlib, wordcloud, seaborn, re (Regular expressions), NLTK, scikit-learn(sklearn), textstat, spaCy, textblob, Keras, XGBoost, spipy, statsmodels.api, lightgmb, catboost and TensorFlow. Throughout the experiment, a random seed value of 42 was used.

3.1.2.2 Data collection

This study will use multiple datasets defined in previous literature and other sources. The first dataset is the Horn dataset (*Horne 2017 Fake News Data*, no date). The horn dataset has text files that are Fake, Real or Satire; which were processed. The dataset has 326 rows and five columns: unnamed:0 (index), Title, Content, Label, and ID. The Horn dataset was chosen for the reliability of the data. A collection of fake news datasets was used to increase the data used. Another data set, Fake News Detection was added

(Jruvika, 2017). This dataset contained URLs, Headlines, Bodies and Label columns containing 4009 rows. Next, the real or fake dataset was added which contained title, text and label columns (George, 2017). It contains 6335 rows of data. Furthermore, the Fake news dataset was added; only the training data was used with 20 800 rows (Lifferth, 2018). It contained the id, title, author, text, and label. ISOT Dataset was added with 44 898 rows and text, data, title, and subject as the columns (Ahmed, Traore and Saad, 2018). A Twitter dataset called Natural Language Processing with Disaster Tweets is also used (Addison Howard devrishi, 2019). This Kaggle dataset contains 7613 rows and three columns: id, text, and target.

3.1.2.3 Data preprocessing and analysis

3.1.2.3.1 Preprocessing

The training data was compiled using the following. The horn data has columns of text of data type object, the label column was of type float, and the columns ID was of data type integer. In the horn dataset, the Fake files received a label of 0, a Real label of 1 and Satire was given the label of 0. Twitter and ISOT datasets had no missing data. Twitter dataset columns name target was renamed to label. ISOT was initially split into two data frames called Fake and Real; the datasets were combined; fake was given label 0, and real news was given label 1. In the ISOT dataset, the columns date, title, and subject were removed. Missing data within the Fake News dataset was removed, and the labels initially set to a value of 1 were converted to 0 and, conversely, to align with the requirements of the study. Also, the title and ID were dropped. The real or fake dataset label was changed where “FAKE” was given a value of 0 and “REAL” was given 1. Also, the title and ID columns were dropped. The fake news detection dataset contained missing values, which were removed. Columns “Body” and “Label” were renamed to “text” and “label”; also, ‘URLS’ and ‘Headline’ were dropped.

The data frames were merged, keeping the text and label columns. The new data frame had 60 missing values, which were dropped. The data frame was also checked for data that only contained one character. In the data frame, some entries only contained one blank space. Thus, these rows were dropped for a total of 741 rows. Duplicates were rechecked to ensure no data was the same across the databases; 9875 rows were dropped. The data was split into training and testing data; a split of 8% was conducted. The following tables contain the count for each column: Table 3 & Table 4

Column	Count	Datatype
text	67 440	object
label	67 440	float

Table 3: Data frame columns for the training data

Column	Count	Datatype
text	5 865	object
label	5 865	float

Table 4:Data frame columns for the testing data.

3.1.2.3.2 Feature Creation

New features (See Table 5: New Features Created) give more information about the current data. Features are created for both the training and testing datasets. The training data was shuffled to ensure an even distribution of data from all sources. 'average_sentence_length' is the sentence length of the text column. The punctuation and capitalisation for text were also calculated for both the test and training data. The new features will reveal any distinction in usage between fake news and true news.

Additionally, a critical cleaning technique is the removal of non-characters and capital letters to normalize the text in the dataset as part of the process. A before and after non-character removal is created to indicate the number of non-characters in a text. A feature named 'word_count' was created to show the number of words in each text column. Also, a feature is created, 'character_removal_diff', the difference between 'text_before_character_removal' and 'text_after_character_removal'. The latter feature was then dropped. Both training and testing datasets undergo NLTK word tokenizer to determine the number of words in the text. A new feature for the training and testing data called ‘text_processed’ is created after the character removal, stop word removal and stemmer. The library of NLTK was used to remove stop words and create stemmers. This column will enable models and other algorithms to define more information about the text.

Subjectivity is the presence of opinion and factual information in a text. TextBlob’s subjectivity function calculated the subjectivity of the text for the test and train data. The value of subjectivity was calculated using NLTK’s sentiment subjectivity library, which assigns a value between 0 and 1. Sentiment is how the text will make a person feel negative or positive; sentiment features are created for both the test and train data. The sentiment was calculated using the NLTK library, which determines a value between -1 and 1. The sentiment gives the emotional tone of the text.

In addition, the reading ease of the text was calculated using the textstat library, which assigns a value ranging from negative to positive numbers to indicate how easy it is for a reader to understand. Furthermore, lexical diversity columns were created to show the number of unique words in the text. Lexical diversity returns a value between 0 and 1 that indicates the lexical diversity score (Table 5: New Features Created).

One of the most essential features created was the creation of the bag of words data frame. Using Sklearn’s text feature_extraction, a vocabulary is created, and then a data frame is created limited to the top 1000 words used in the text. The new bag of words data frame contains the 1000 most used words, showing if the word was used in the text (1) or not present in the text (0). The bag of words data frame with 1000 columns is merged into the dataset with the new features. The top 1000 words from the training dataset are utilised to identify and calculate the occurrences of these specific words within the testing data frame, if present in the text.

Column	datatype
average_sentence_length	Float
text_punctuation_frequency	Integer
text_capitilazation_frequency	Integer
word_count	Integer
character_removal_diff	Integer
text_processed	Object
subjectivity_text	Float
text_sentiment	Float
text_reading_ease	Float
text_lexical_diversity	Float

Table 5: New Features Created

3.1.2.3.3 Model Implementation

XGBoost model was implemented using the xgboost library. The parameters were set. The tree_method is 'hist', gpu_id=0, and max_depth is 25. After scaling was done on the data, the model was fitted on the X_train and y_train data. Furthermore, the model's prediction function was used to predict the X_validation and X_test data.

The FNN model was implemented with an input layer of 256 neurons, an activation function of relu, batch normalisation, and a dropout of 0.5. Next are two hidden layers of 128 and 64 neurons: relu activation, batch normalisation and a 0.5 dropout. Table 6: FNN model summary: the FNN model used. The model was compiled using Adam optimiser with a binary cross entropy loss. The metric used was accuracy. The FNN model was fitted to the X_train and Y_train data with an epoch of 25, batch size of 64 and a validation split of 20%. Predictions were first made on the X_validation data, and the model was predicted in probability. Thus, the predictions were converted to binary predictions using (probability predicted>0.5 is one and vice versa). The exact process was followed with the testing data. The summary of the model used in Table 6:

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 256)	258560
batch_normalization (BatchNormalization)	(None, 256)	1024
dropout (Dropout)	(None, 256)	0
dense_1 (Dense)	(None, 128)	32896
batch_normalization_1 (BatchNormalization)	(None, 128)	512
dropout_1 (Dropout)	(None, 128)	0
dense_2 (Dense)	(None, 64)	8256
batch_normalization_2 (BatchNormalization)	(None, 64)	256
dropout_2 (Dropout)	(None, 64)	0
dense_3 (Dense)	(None, 1)	65

Table 6: FNN model summary

The logistic regression model and Random Forest were implemented using the sklearn Logistic Regression and Random Forest Classifier libraries. The default parameters for the logistic regression model other than the random state set to 42 were used. Random forest model used estimators of 100, criterion as entropy and max features used to square root. The models were fitted using X_train and y_train, and predictions were made on the X_validation and X_test data.

The LightGBM and CATBoost models used default parameters. They were fitting the models on the X_train and Y_train data. Predictions were made on X_validation and X_test data.

3.1.2.4 Evaluation

Evaluation will be done using Sklearn's metrics. At first, the training data will be split into a training data and validation dataset, split 85% and 15%, respectively. The model will be tested on the validation dataset before the testing dataset. The models' accuracy will be determined using the validation and test datasets. Furthermore, a confusion matrix will be set using Sklearn, showing the relationship between precision and recall for the testing and validation sets.

3.1.3 Ethical requirements

Researchers are not allowed to falsify their results in any way. When research is completed, the report should reflect the accuracy and completeness of the research. Research data should always be kept safe and secure. Research must be reported even if it does not support the outcomes that the researcher anticipated. Researchers are required to show their methodology and reasoning behind all decisions. All data and information other journals use should be credited (Unisa, 2007).

4 Chapter 4

4.1 Data analysis and results

4.1.1 Data analysis

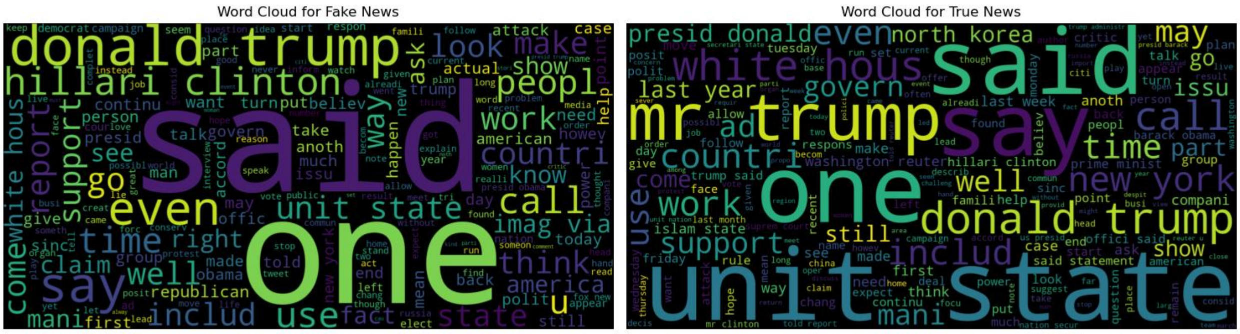


Figure 3: Word Cloud for Fake and True News

Word clouds indicate the frequency of words in the text. Using the word cloud library, we can visualise the number of times a word was used. In Figure 3,the word cloud for fake news; words like ‘donald trump’, ‘call, ‘say’, ‘one’ and ‘hillari clinton’ are used more than others. Figure 3 also shows the word cloud for True News; we can see that ‘one’, ‘said’, ’donald trump’, ‘unit state’, ’say’, and ’even’ are used more often.

Before moving to the models, the features were assets based on the statistical values. Features of object datatype, index, text, and text_processed were dropped. The models will work with numerical data. During the analysis of the data the following created columns were of importance 'label', 'avg_sentence_length', 'text_punctuation_frequency', 'text_capitalization_frequency', 'word_count', 'character_removal_diff', 'subjectivity_text', 'text_sentiment', 'text_reading_ease', 'text_lexical_diversity'. Comparing the mean, standard deviation, minimum value, and maximum value, it was apparent that the data had outliers. Histograms were used to visualize the extent of outliers. Figure 4and Figure 5 illustrate the outliers in the data. In Figure 4, the bell-shaped distribution is skewed to the left. Furthermore, Figure 5 only shows one bar on the far left, indicating the presence of outliers.

The following statistics showed the presence of outliers:

- ‘word_count’ maximum of 23 839 and 75% percentile of 625
- The ‘text_punctuation_frequency’ maximum value is 7 295, and the 75% percentile is 88.
- ‘character_removal_diff’ with a mean of 110.6 and a standard deviation of 513
- ‘text_reading_ease’ with a mean of -211.7 and a standard deviation of 326.57

Outliers were dropped using the Interquartile Range (IQR) method with a multiplier of 3. A multiplier of 3 was used to reduce the amount of data dropped. This study uses data from various sources, increasing the likelihood of outliers in the data.

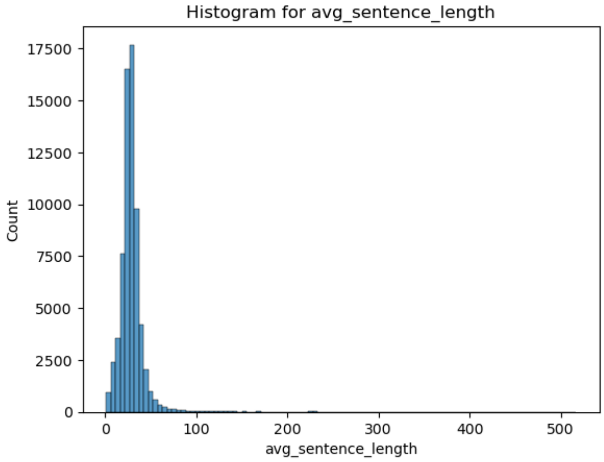


Figure 4: Histogram showing outliers in avg_sentence_length.

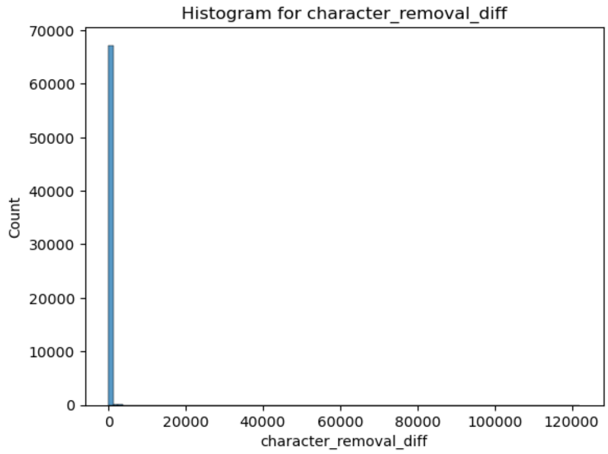


Figure 5: Histogram showing outliers in character_removal_diff.

The training data contained 67 440 rows before the outlier was dropped. After outliers were dropped, 63 291 remained; thus, 4 149 rows of data were removed.

Figure 6 shows the distribution of the avg_sentence_length values after outlier removal. The distribution is bell-shaped. The data in this feature is between 22.5 and 32.45.

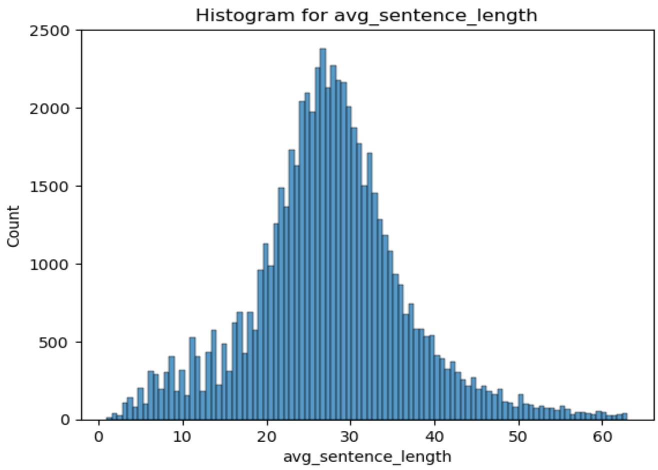


Figure 6: avg_sentence_length after outliers removed.

Figure 7 shows the distribution of the text_punctuation_frequency feature. This figure shows a binomial distribution since multiple datasets were used. Therefore, not all datasets will have the same distribution.

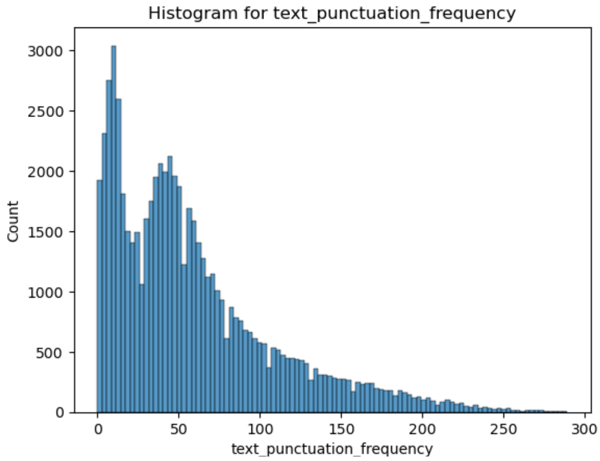


Figure 7: Histogram for text_punctuation_frequency.

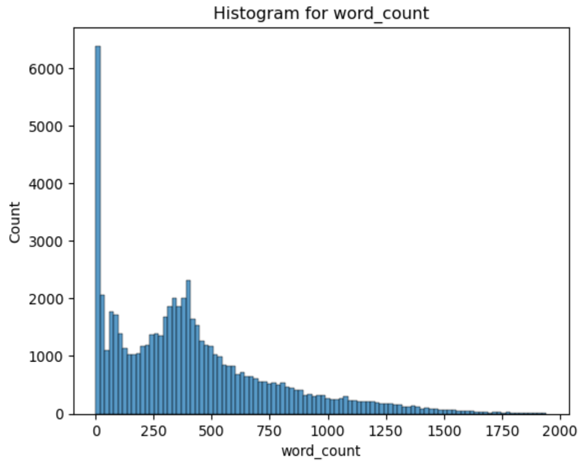
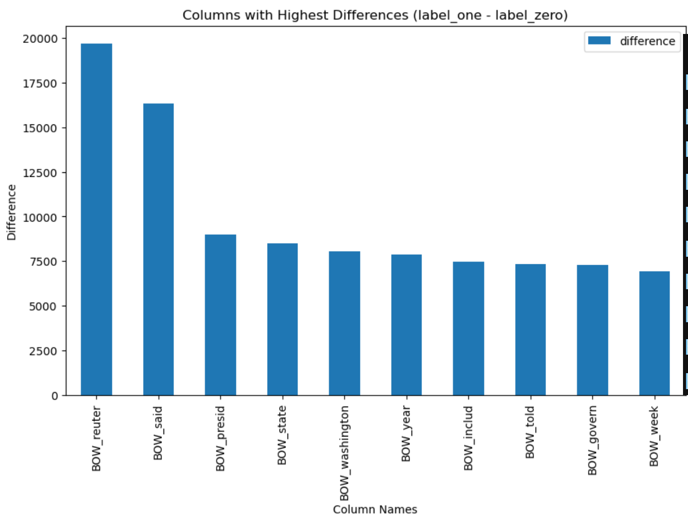


Figure 8: Histogram for word_count.

Figure 8 shows a binomial distribution; one of the peaks is very low with a large amount of data. The low word count is a consequence of Twitter data in the dataset.



	label_zero	label_one	difference
BOW_reuter	333	20047	19714
BOW_said	12171	28486	16315
BOW_presid	9458	18449	8991
BOW_state	9755	18271	8516
BOW_washington	2946	11002	8056
BOW_year	8706	16561	7855
BOW_includ	5369	12845	7476
BOW_told	5478	12792	7314
BOW_govern	4896	12177	7281
BOW_week	4462	11392	6930

Figure 9: BOW column occurrence counts.

Figure 9 is a visual representation of the count of each column. For example, take BOW_reuter. The figure shows that 333 occurrences of the word reuters were fake news, and 20 047 occurrences were true news. The graph shows the differences for the top 10 columns. All these columns were in favour of True News. The numbers suggest that when the word Reuters, Washington, presid, state, include, told, govern, and week is in a text the article or news is more likely to be true news.

Figure 10 is word clouds showing the word presence count. The fake news bag of word presence count shows that fake news uses ‘said’, ‘like’, ‘peopl’, ‘trump’ and ‘time’ more frequently. However, words such as ‘said’, ‘reuter’, ‘state’, ‘presid’, and ’year’ are more frequently used in True News. The two-word clouds have some BOWs in common, for example, “said”, ‘people’, ’time’, ‘trump’ and ‘state’. However, if ‘reuter’ and ’presid’ are used in a text, then there is a higher likelihood that the news is trustworthy. Also, when ‘like’ and ‘peopl’ and is used, the likelihood that news is fake is more significant.

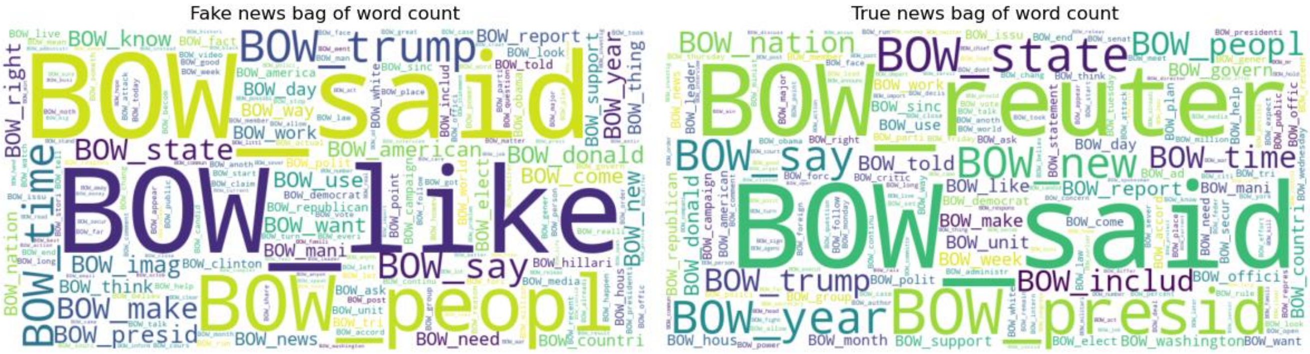


Figure 10: Bag of words column count word cloud.

- Figure 11 shows the averages of key features corresponding to the label. The average subjectivity of the text indicates that both true and fake news have low degrees of subjectivity. However, True news has an average subjectivity of 0.335, slightly lower than fake news's subjectivity of 0.373.
- The average reading ease indicates that the reading ease is very low; this indicates that the content is hard to understand, contains jargon, or sentences are complex (Figure 11). Fake news has a reading ease average of -134 and a true news average of -207, which signifies that Fake news is easier to understand and might indicate a sense of misleading the reader. People could find it easier to understand, meaning they are more likely to believe it.
- The average text punctuation frequency and character removal difference indicate that true news has a higher average in punctuations and non-characters (Figure 11). These features could indicate if the news is true or false since true news sources use more quotes to reference a source.
- Figure 11 shows that the average word count for fake news is 362.4, and true news is 468.4. The difference of 106 indicates that the feature could indicate true and fake news. Using more words in a true news article could indicate that more detail is given to elaborate on the subject. Furthermore, in fake news the number of words is less possibly because it gives vague information to the readers.
- Figure 12 shows the distribution of the labels (0: Fake News, 1: True News). Indicating that 44.9% of the data pertains to fake news and 55.1% represents true news. This analysis indicates no substantial class imbalance, and the data is distributed evenly.

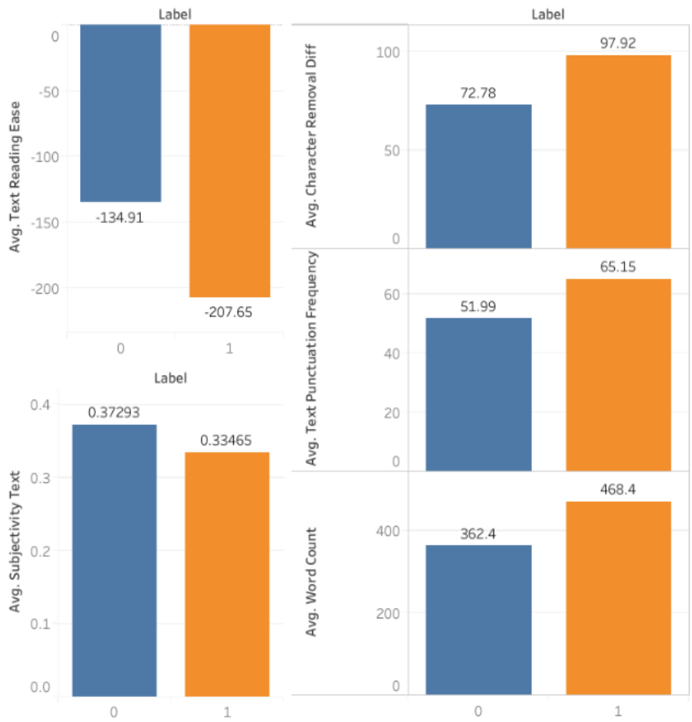


Figure 11: Averages for key features.

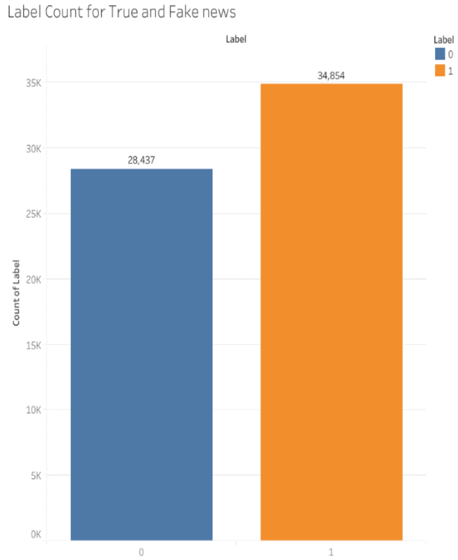


Figure 12: Label Count for True and Fake News.

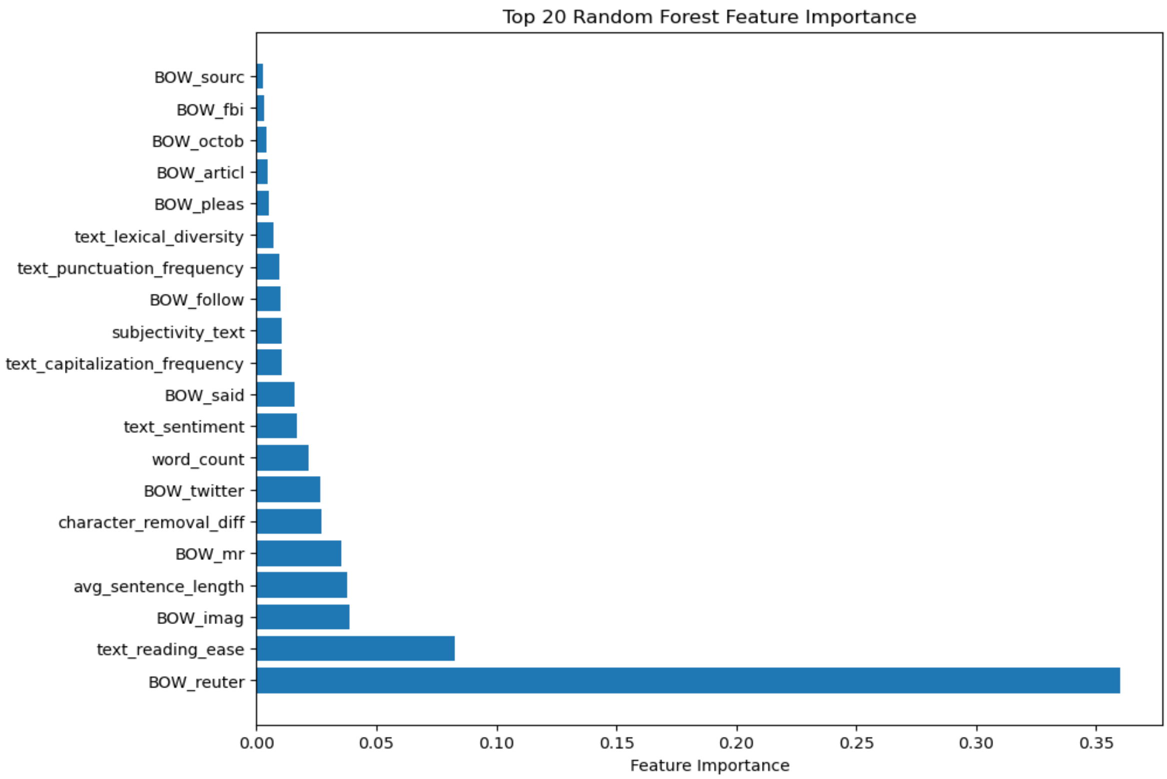


Figure 13: Feature importance for a random forest model.

A random forest model was employed to assess feature importance for testing a hypothesis related to fake news. In Figure 13, a random forest model feature importance demonstrates that BOW_reuter holds the highest significance, followed by text_reading_ease. Words such as 'imag', 'mr', 'twitter', and 'said' aids in determining if news is true or fake. These features were further analysed using hypothesis testing. In Table 7, all the coefficients for the top 12 features and all the features had a p-value of close to zero. For instance, BOW_reuter exhibited a coefficient of 5.1, suggesting that the presence of the word Reuter increases the likelihood of the news being true.

Furthermore, the word image (BOW_imag) indicates that news is fake, potentially due to the rarity of image captions in true news articles. The word 'said, and 'twitter' can be used to quote a person either on twitter or a direct quote. Meanwhile, text_reading_ease and word_count are indicators that news is fake even though the margins are minimal. More positive texts are more likely to be true, indicating that fake news tends to evoke negativity. Shorter sentence lengths indicate that news is True where avg_sentence_length has a coefficient of -0.056. Moreover, true news exhibits fewer capitalisations and lower subjectivity levels.

Feature name	Coefficient
BOW_reuter	5.1554
Text_reading_ease	-0.0032
BOW_imag	-2.3525
Avg_sentence_length	-0.0560
BOW_mr	1.3996
Character_removal_diff	0.0067
BOW_twitter	1.1987
Word_count	-0.0011
Text_sentiment	0.0856
BOW_said	1.0378
Text_capitalization_frequency	-0.0762
Subjectivity_text	-1.6202

Table 7: Hypothesis testing Coefficients.

Before modelling the ‘text’ and ‘text_processed’ columns were dropped. These columns are object types and provide no significant information.

Next, the data was split into train and validation sets before the models received them. They use Sklearn ‘train_test_split’ with a validation size of 0.15 and a random state of 42.

Table 8 shows the resulting variables. Based on the Labels 0 and 1 split, we are satisfied with the test and train data distribution.

split	shape	Label 0	Label 1
X_train	53797, 1009	24100	29697
y_train	53797		
X_validation	9494, 100	4337	5157
y_validation	9494		
X_test	5865, 1009	2673	3192
Y_test	5865		

Table 8: Data split.

4.1.2 Results

Model		Accuracy	F1-score	Confusion Matrix
XGBOOST	Validation	0.92289	(0) 0.92 (1) 0.93	[[4070 267] [465 4692]]
	Test dataset	0.86138	(0) 0.86 (1) 0.86	[[2452 221] [592 2600]]
FNN	Validation	0.91721	(0) 0.91 (1) 0.92	[[3971 366] [420 4737]]
	Test dataset	0.88456	(0) 0.87 (1) 0.90	[[2215 458] [219 2973]]
Logistic Regression	Validation	0.89593	(0) 0.89 (1) 0.90	[[4020 317] [671 4486]]
	Test dataset	0.85490	(0) 0.83 (1) 0.88	[[2017 656] [195 2997]]
Random Forest	Validation	0.90183	(0) 0.89 (1) 0.91	[[3954 383] [549 4608]]
	Test dataset	0.84296	(0) 0.84 (1) 0.85	[[2403 270] [651 2541]]
LightGBM	Validation	0.91615	(0) 0.91 (1) 0.92	[[4051 286] [510 4647]]
	Test dataset	0.87365	(0) 0.87 (1) 0.88	[[2478 195] [546 2646]]

CATBoost	Validation	0.92732	(0) 0.82 (1) 0.93	[[4079 258] [432 4725]]
	Test dataset	0.86922	(0) 0.86 (1) 0.87	[[2441 232] [535 2657]]

Table 9: Results for models.

XGBoost has an accuracy of 0.923 on the validation set. Additionally, an f1-score of 0.92 for fake news and 0.93 for true news demonstrates an exceptional balance between precision and recall. The confusion matrix shows that 267 data points were misclassified as true, and 465 were misclassified as false. The model exhibited an accuracy of 86% on the testing set, indicating a deviation from the performance seen in the validation set. However, the accuracy drop is because the test set consists of data the model has never seen before. The test data's f1-scores for fake and true news was 0.86. According to the confusion matrix, 221 fake articles were incorrectly classified as true, and 592 true articles were incorrectly classified as false. The matrix suggests that the model is more inclined to predict a true article as fake. However, in the realm of fake news, this is an acceptable result.

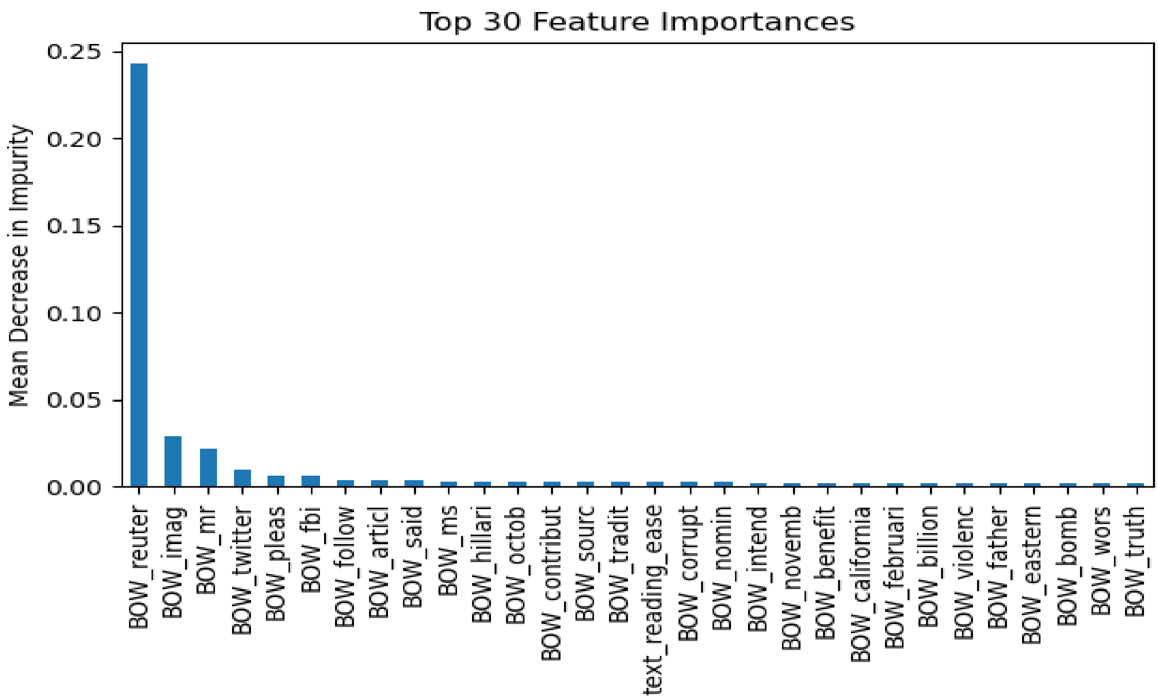


Figure 14: XGBoost- Top 30 Feature Importance

In Figure 14, the feature importance of the XGBoost model is depicted. The Figure illustrates the mean decrease in impurity, indicating each feature's contribution to the model's decision-making accuracy. The most significant feature is 'BOW_reuter'; this feature is a component of the bag of words, and the word 'reuter' is a good indicator of the authenticity of the news. Additionally, words like 'image', the stemmed word for image,' 'mr',' twitter', 'pleas', and 'fbi', aided the XGBoost model's predictions.

The FNN model achieved 91% accuracy on the validation and 88% on the test sets. The model performs very well on data that has not been seen before. The confusion matrix shows that the validation set classifies 366 fake news articles as true and 420 true news as false. In the test set, 458 false articles were classified as true, while 219 correct articles were incorrectly classified as false. The model demonstrates exceptional performance in predicting fake news.

The logistic regression model achieved an accuracy of 89.5% on the validation set. The f1-score indicates 0.89 for fake news data and 0.90 for true news data. The relationship between recall and precision is excellent. The confusion matrix shows 317 false positive data points and 671 false negative data points, which is acceptable in fake news.

Figure 15 displays the coefficient values representing the top 20 features for the logistic regression model. The coefficients of BOW_reuter (4.77) and 'character_removal_diff' (4.2) indicate that the news is true. Conversely, features like 'average_sentence_length' (-3.189), 'text_capitalization_frequency' (-2.16), 'BOW_imag' (-2.04), and 'text_punctuation frequency' (-1.87) serve as reliable indicators of potentially fake news.

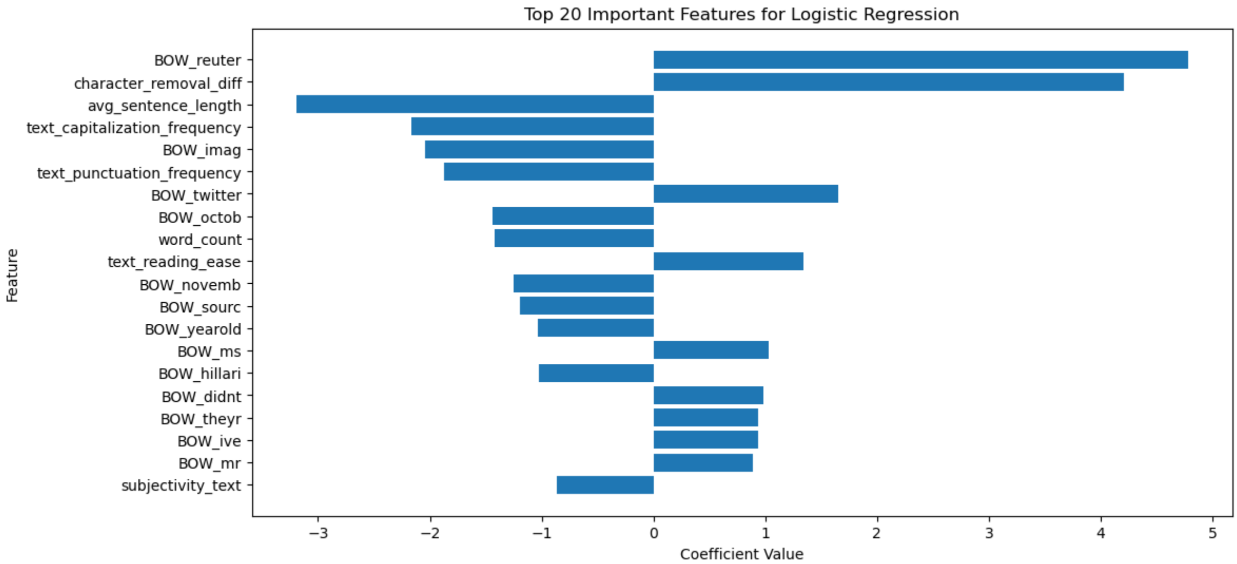


Figure 15: Top 20 important features for logistic regression.

The Random Forest model achieved an accuracy of 90% on the validation set and 84% on the test set. Additionally, the f1-score shows the model has a score of 0.89 for fake news and 91% for true news correctly for validation; the model has a reliable balance between precision and recall. The test set f1-score is 0.84 and 0.85 for fake and true news, respectively. The confusion matrix displays 389 false positives and 549 false negatives in the validation set. The test set has 270 false positives and 651 false negatives. In the context of fake news, these values are considered acceptable.

LightGBM model was employed on both the validation and test set, achieving accuracies of 91.6% and 87.4%, respectively. Both sets achieved a reliable balance between precision and recall. Even though the test data predicts 456 true news as fake, this is acceptable for the scenario. LightGBM proves to be an excellent model for predicting fake news.

In the CATBoost validation set, an accuracy of 92.7% was achieved with f1-scores of 0.82 (fake) and 0.93 (true), indicating a non-ideal precision-recall balance. However, the test set had an f1-score of 0.86 (fake) and 0.87 (true), showing an improved ratio on unseen data. Despite dropping to 86.9, the accuracy remains optimal. The confusion matrix on the test set showed that the model misclassified 232 fake news articles as true and 535 true articles as false, which is acceptable.

Figure 16 presents the Receiver Operating Characteristic (ROC) Curve. An AUC score of 0.98 indicates that the model possesses excellent class discrimination. The ROC curve is towards the top-left corner, suggesting the model achieves higher true positives than false positives.

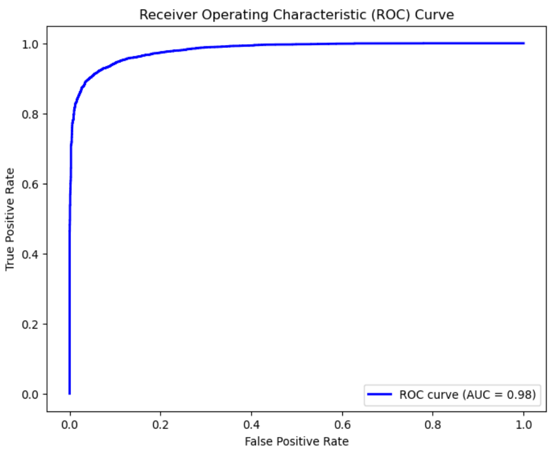


Figure 16: CatBoost ROC Curve

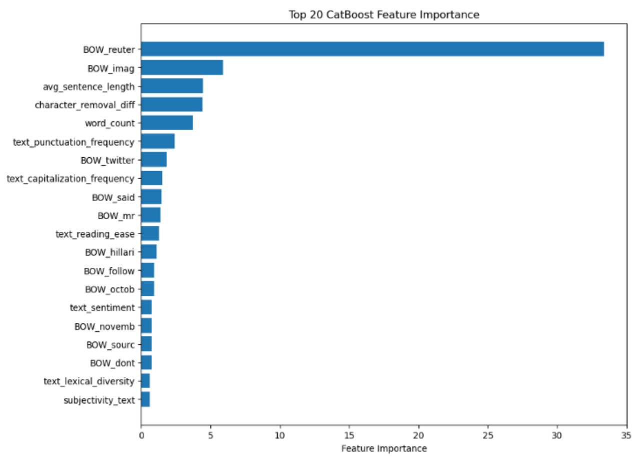


Figure 17: CatBoost feature importance.

Figure 17 again emphasizes the significance of ‘BOW_reuter’ and ‘BOW_imag’. Additionally, it highlights the contributory role of ‘avg_sentence_length’, ’character_removal_diff’, ’word_count’, and ‘text_punctuation_frequency’ in determining fake news.

The research question posed by this study: What is the performance of different machine learning and deep learning techniques for detecting fake news on a heterogeneous dataset composed of data from various sources, considering both evaluation metrics and dataset characteristics?

Machine and deep learning models performed very well on the merged data. Deep learning models like FNN performed the best overall. They achieved an accuracy of 88%. Machine learning models performed very well, such as XGBoost, CatBoost, and LightGBM. The algorithms mentioned are, thus far, the best techniques for detecting fake news.

The results showed how data from different datasets and sources can be merged into a single dataset comprising multiple sources with different contexts. Then, creating features based on the text of the new dataset. The most essential part of detecting fake news is creating the features. In the analysis, it is apparent that each feature created had a vital part to play. A bag of words produces the words that help different models predict fake news.

Features like average sentence length, character removal and bag of words showed how fake news and true news can be identified. Using all these features increases the likelihood of identifying fake news.

4.1.3 Limitations and future work

Limitations in this study are the time-consuming nature of verifying that the dataset contains true news, and that the news is fake. The data is assumed to be true because the data was used in previous studies during the literature review. However, the use in previous studies only guarantees that some of the data is correct. Furthermore, more data should be added from Twitter and WhatsApp datasets. The Twitter and WhatsApp datasets are difficult to verify. The need for computer resources to perform the necessary hyper-tuning is also a study limitation. Hyper-tuning can help models perform better and reduce overfitting.

Future work includes the hyper-tuning of the algorithms and creating more complex features. Incorporating compound words or using embedding techniques on the text data could reduce the models' performance but increase the accuracy. These enhancements can lead to more accurate analysis and increase the performance of models.

5 Chapter 5

5.1 Conclusion

Fake news is in our daily lives; we see it on social media and everywhere we look. Fake news is spread to cause havoc and deceive the reader. Furthermore, spotting fake news is very difficult; the news is created to mislead the reader and compel the user to refrain from asking questions. There is an obvious need for fake news detection. Additionally, with the data and algorithms at our disposal, we must find new ways to detect fake news and reduce its effects on people.

The literature in this study showed that machine learning algorithms were a reliable way to detect fake news. The effects of data cleaning and feature creation became apparent in the literature. Cleaning data using non-character removal, stop words and stemming allowed the data to be normalised before creating features. Features were created in the literature to show the most used words, reading ease and sentiment of the data. In essence, cleaning data and creating meaningful features on our data is vital before attempting machine learning techniques. The literature reviewed showed that numerous models can be used to predict if news is true or fake. In the literature, XGBoost achieved 75%, 94% and 91.39% accuracies, while Random Forest attained 73%, 80%, 92%, and 92.18% accuracy across various studies. Most techniques had 90 % accuracy, although specific models displayed lower accuracies in the literature. The reviewed literature did not encompass Feedforwards Neural Networks, CatBoost or LightGBM models.

This study was conducted to clarify the performance of models on various datasets. Datasets were merged; the data was cleaned as described by the literature. New features were created to give insight into the data using guidance from the literature.

This study utilised XGBoost, FNN, LightGBM, CatBoost Logistic Regression and Random Forest models to predict fake news. Based on the results of the models, the models performed similarly to the literature models.

The increase in data and amalgamation of datasets from various sources did introduce some variations in accuracy. However, the models ultimately achieve a satisfactory overall performance. The XGBoost model in this study had an accuracy of 86%, lower than in the literature; however, the result exceeded expectations. Random forest predicted fake news with an accuracy of 84%, and Logistic regression had an accuracy of 85%. Both models performed lower than the literature, yet the results were deemed acceptable. The FNN, LightGBM and CatBoost results were 88%, 87.4% and 86.9%, respectively, displaying the highest accuracy among the models in this study. The confusion matrix indicated that all models predicted an average of 339 articles as fake, which should have been true and 456 as true, which should have been fake. These misclassifications are considered acceptable within the realm of fake news algorithms.

In conclusion, increased data and utilisation of various sources in a single database slightly reduced the accuracy. However, these reductions were not substantial, and the overall performance remained within acceptable ranges. Features derived from the text of the news article, such as the article's length, the non-characters and specific words used, played a crucial role in fake news detection.

Fake news can be detected using machine learning and deep learning models namely, FNN, LightGBM and CatBoost with accuracy of 88%, 87.4% and 86.9%, respectively.

6 References

Abdulrahman, A. and Baykara, M. (2020) 'Fake News Detection Using Machine Learning and Deep Learning Algorithms', in *3rd International Conference on Advanced Science and Engineering, ICOASE 2020*, pp. 18–23. doi:10.1109/ICOASE51841.2020.9436605.

Addison Howard devrishi, P.C.Y.G. (2019) 'Natural Language Processing with Disaster Tweets', *Kaggle* [Preprint]. Available at: https://www.kaggle.com/competitions/nlp-getting-started/data.

Ahmad, I. *et al.* (2020) 'Fake News Detection Using Machine Learning Ensemble Methods', *Complexity*, 2020, p. 11. doi:10.1155/2020/8885861.

Ahmed, H., Traore, I. and Saad, S. (2018) *Detecting opinion spams and fake news using text classification*, *Journal of Security and Privacy*. Available at: https://onlineacademiccommunity.uvic.ca/isot/2022/11/27/fake-news-detection-datasets/.

Al-Ahmad, B. *et al.* (2021) 'An evolutionary fake news detection method for covid-19 pandemic information', *Symmetry*, 13(6), p. 1091. doi:10.3390/sym13061091.

AI, M. (no date) *Light Gradient Boosting Machine. Light GBM (Light Gradient Boosting… | by AI Maverick | Medium*. Available at: https://samanemami.medium.com/light-gradient-boosting-machine-b4f1b9e3f7d1 (Accessed: 23 December 2023).

Arora, S. (2022) *Sentiment Analysis Using Python*, *Analytics Vidhya*. Available at: https://www.analyticsvidhya.com/blog/2022/07/sentiment-analysis-using-python/#h-what-is-sentiment-analysis (Accessed: 28 June 2023).

Al Asaad, B. and Erascu, M. (2018) 'A tool for fake news detection', in *Proceedings - 2018 20th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, SYNASC 2018*, pp. 379–386. doi:10.1109/SYNASC.2018.00064.

Baarir, N.F. and Djeffal, A. (2021) 'Fake News detection Using Machine Learning', in *2020 2nd International Workshop on Human-Centric Smart Environments for Health and Well-Being, IHSH 2020*. Institute of Electrical and Electronics Engineers Inc., pp. 125–130. doi:10.1109/IHSH51661.2021.9378748.

Belova, G. and Georgieva, G. (2018) 'Fake News as a Threat to National Security', *International conference KNOWLEDGE-BASED ORGANIZATION*, 24(1), pp. 19–22. doi:10.1515/kbo-2018-0002.

D'Ulizia, A. *et al.* (2021) 'Fake news detection: A survey of evaluation datasets', *PeerJ Computer Science*, 7, pp. 1–34. doi:10.7717/PEERJ-CS.518.

Dame Adjin-Tettey, T. (2022) 'Combating fake news, disinformation, and misinformation: Experimental evidence for media literacy education', *Cogent Arts & Humanities*, 9(1). doi:10.1080/23311983.2022.2037229.

Dubey, P. (2018) *An introduction to Bag of Words and how to code it in Python for NLP*, *FreeCodeCamp*. Available at: https://www.mygreatlearning.com/blog/bag-of-words/ (Accessed: 28 June 2023).

Erika D (2019) *Accuracy, Recall & Precision. When it comes to evaluating how well a… | by Erika D | Medium*, *Medium*. Available at: https://medium.com/@erika.dauria/accuracy-recall-precision-80a5b6cbd28d (Accessed: 30 June 2023).

Feed, I. (no date) 'Feed Forward Neural Network Definition | DeepAI'. Available at: https://deepai.org/machine-learning-glossary-and-terms/feed-forward-neural-network (Accessed: 23 December 2023).

Ferreira, C.C., Robertson, J. and Kirsten, M. (2020) 'The truth (as I see it): philosophical considerations influencing a typology of fake news', *Journal of Product and Brand Management*, 29(2), pp. 150–158. doi:10.1108/JPBM-12-2018-2149.

GeeksforGeeks (2022) *Removing stop words with NLTK in Python - GeeksforGeeks*. Available at: https://www.geeksforgeeks.org/removing-stop-words-nltk-python/ (Accessed: 27 June 2023).

George, M. (2017) *fake_real_news_dataset*. Available at: https://github.com/joolsa/fake_real_news_dataset/blob/master/fake_or_real_news.csv.zip.

Gilda, S. (2017) 'Notice of Removal: Evaluating machine learning algorithms for fake news detection', *IEEE Student Conference on Research and Development: Inspiring Technology for Humanity, SCOReD 2017 - Proceedings*. Institute of Electrical and Electronics Engineers Inc., pp. 110–115. doi:10.1109/SCORED.2017.8305411.

Gupta, R. (2023) *Accuracy, Precision, Recall, F-1 Score, Confusion Matrix, and AUC-ROC*, *Medium*. Available at: https://medium.com/@riteshgupta.ai/accuracy-precision-recall-f-1-score-confusion-matrix-and-auc-roc-1471e9269b7d.

Horne 2017 Fake News Data (no date). Available at: https://paperswithcode.com/dataset/horne-2017-fake-news-data (Accessed: 3 October 2023).

Javatpoint (2021) ‘Machine Learning Random Forest Algorithm’, *Www.Javatpoint.Com* [Preprint]. Available at: https://www.javatpoint.com/machine-learning-random-forest-algorithm (Accessed: 6 October 2023).

JavatPoint (2018) ‘Logistic Regression in Machine Learning - Javatpoint’, pp. 1–22. Available at: https://www.javatpoint.com/logistic-regression-in-machine-learning (Accessed: 6 October 2023).

Jruvika (2017) *Fake News Detection*, *Kaggle*. Available at: https://www.kaggle.com/datasets/jruvika/fake-news-detection (Accessed: 27 December 2023).

Kaliyar, R.K., Goswami, A. and Narang, P. (2021) ‘FakeBERT: Fake news detection in social media with a BERT-based deep learning approach’, *Multimedia Tools and Applications*, 80(8), pp. 11765–11788. doi:10.1007/s11042-020-10183-2.

Kannan, S. *et al.* (2021) ‘Detection of Fake News related to COVID-19 using Natural Language Processing’, in *2021 Asian Conference on Innovation in Technology, ASIANCON 2021*. Institute of Electrical and Electronics Engineers Inc. doi:10.1109/ASIANCON51346.2021.9545047.

Khan, J.Y. *et al.* (2021) ‘A benchmark study of machine learning models for online fake news detection’, *Machine Learning with Applications*, 4, p. 100032. doi:10.1016/j.mlwa.2021.100032.

Khanam, Z. *et al.* (2021) ‘Fake News Detection Using Machine Learning Approaches’, *IOP Conference Series: Materials Science and Engineering*, 1099(1), p. 012040. doi:10.1088/1757-899x/1099/1/012040.

Kumar, R. (2011) *Research Methodology a step-by-step guide for beginners*. 3rd edn. Available at: http://www.sociology.kpi.ua/wp-content/uploads/2014/06/Ranjit_Kumar-Research_Methodology_A_Step-by-Step_G.pdf.

Lazer, D.M.J. *et al.* (2018) ‘The science of fake news: Addressing fake news requires a multidisciplinary effort’, *Science*, 359(6380), pp. 1094–1096. doi:10.1126/science.aao2998.

Lifferth, W. (2018) *Fake news*. Available at: https://kaggle.com/competitions/fake-news.

Madrid, P. (2023) 'USC study reveals the key reason why fake news spreads on social media', *USC news*, 17 January. Available at: https://news.usc.edu/204782/usc-study-reveals-the-key-reason-why-fake-news-spreads-on-social-media/ (Accessed: 7 June 2023).

Manzoor, S.I., Singla, J. and Nikita (2019) 'Fake news detection using machine learning approaches: A systematic review', in *Proceedings of the International Conference on Trends in Electronics and Informatics, ICOEI 2019*, pp. 230–234. doi:10.1109/ICOEI.2019.8862770.

Menard, S. (2014) 'An Introduction to Logistic Regression Diagnostics', in *Applied Logistic Regression Analysis*, pp. 68–91. doi:10.4135/9781412983433.n4.

Mishra, S., Shukla, P. and Agarwal, R. (2022) 'Analyzing Machine Learning Enabled Fake News Detection Techniques for Diversified Datasets', *Wireless Communications and Mobile Computing*. Hindawi Limited. doi:10.1155/2022/1575365.

Mondal, A. (2022) *LightGBM in Python | Complete guide on how to Use LightGBM in Python*, *Analytica Vidya*. Available at: https://www.analyticsvidhya.com/blog/2021/08/complete-guide-on-how-to-use-lightgbm-in-python/ (Accessed: 23 December 2023).

Monti, F. *et al.* (2019) 'Fake News Detection on Social Media using Geometric Deep Learning'. Available at: https://www.snopes.com/ (Accessed: 6 June 2023).

Morde, V. (2021) *XGBoost Algorithm: Long May She Reign! - Towards Data Science. Medium.* Available at: https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d (Accessed: 29 June 2023).

Mridha, M.F. *et al.* (2021) 'A Comprehensive Review on Fake News Detection With Deep Learning', *IEEE Access*, 9, pp. 156151–156170. doi:10.1109/ACCESS.2021.3129329.

Narwal, B. (2018) 'Fake News in Digital Media', in *Proceedings - IEEE 2018 International Conference on Advances in Computing, Communication Control and Networking, ICACCCN 2018*, pp. 977–981. doi:10.1109/ICACCCN.2018.8748586.

Nasir, J.A., Khan, O.S. and Varlamis, I. (2021) 'Fake news detection: A hybrid CNN-RNN based deep learning approach', *International Journal of Information Management Data*

Insights, 1(1), p. 100007. doi:10.1016/j.jjimei.2020.100007.

Oppermann, A. (2023) *What Is CatBoost?* Available at: https://builtin.com/machine-learning/catboost (Accessed: 23 December 2023).

Pollock, A. and Berge, E. (2018) 'How to do a systematic review', *International Journal of Stroke*, 13(2), pp. 138–156. doi:10.1177/1747493017743796/ASSET/IMAGES/LARGE/10.1177_1747493017743796-FIG6.JPEG.

Sahoo, S.R. and Gupta, B.B. (2021) 'Multiple features based approach for automatic fake news detection on social networks using deep learning', *Applied Soft Computing*, 100, p. 106983. doi:10.1016/j.asoc.2020.106983.

Seddari, N. *et al.* (2022) 'A Hybrid Linguistic and Knowledge-Based Analysis Approach for Fake News Detection on Social Media', *IEEE Access*, 10, pp. 62097–62109. doi:10.1109/ACCESS.2022.3181184.

Shu, K. *et al.* (2017) 'Fake News Detection on Social Media: A Data Mining Perspective'. Available at: http://www.journalism.org/2016/05/26/news-use-across- (Accessed: 6 June 2023).

sirjan13 (2021) *Readability Index in Python(NLP) - GeeksforGeeks*. Available at: https://www.geeksforgeeks.org/readability-index-pythonnlp/ (Accessed: 28 June 2023).

Staples, M. and Niazi, M. (2007) 'Experiences using systematic review guidelines', *Journal of Systems and Software*, 80(9), pp. 1425–1437. doi:10.1016/j.jss.2006.09.046.

Thota, A. *et al.* (2018) *Number 3 Article 10 2018 Part of the Artificial Intelligence and Robotics Commons Recommended Citation Recommended Citation Thota, Aswini; Tilak, Priyanka; Ahluwalia, Simrat; and Lohia*, *SMU Data Science Review*. Available at: https://scholar.smu.edu/datasciencereviewAvailableat:https://scholar.smu.edu/datasciencereview/vol1/iss3/10http://digitalrepository.smu.edu. (Accessed: 6 June 2023).

Umer, M. *et al.* (2020) 'Fake news stance detection using deep learning architecture (CNN-LSTM)', *IEEE Access*, 8, pp. 156695–156706. doi:10.1109/ACCESS.2020.3019735.

Unisa (2007) 'Policy on research ethics', *Research Department, University of South Africa*, 1, pp. 1–26. Available at:

https://www.unisa.ac.za/static/corporate_web/Content/Colleges/CGS/documents/Policy-on-Research-Ethics-rev-appr-Council-20.09.2013.pdf (Accessed: 20 November 2022).

Varma, R. *et al.* (2021) 'A systematic survey on deep learning and machine learning approaches of fake news detection in the pre- and post-COVID-19 pandemic', *International Journal of Intelligent Computing and Cybernetics*, pp. 617–646. doi:10.1108/IJICC-04-2021-0069.

Wani, A. *et al.* (2021) 'Evaluating Deep Learning Approaches for Covid19 Fake News Detection', in *Communications in Computer and Information Science*. Springer Science and Business Media Deutschland GmbH, pp. 153–163. doi:10.1007/978-3-030-73696-5_15.

Waweru Muigai, J.W. (2019) 'Understanding Fake News', *International Journal of Scientific and Research Publications (IJSRP)*, 9(1), p. p8505. doi:10.29322/ijsrp.9.01.2019.p8505.

Wood, T. (2019) 'F-Score Definition | DeepAI', *DeepAi.org* [Preprint]. Available at: https://deepai.org/machine-learning-glossary-and-terms/f-score (Accessed: 6 October 2023).

Yash, R. (2022) *Python | Stemming words with NLTK*, *GeeksforGeeks*. Available at: https://www.geeksforgeeks.org/python-stemming-words-with-nltk/ (Accessed: 27 June 2023).

7 Appendices

7.1 Appendix A: Mendeley Screenshot

Authors	Title	Year	Published In	Added
	Horne 2017 Fake News Data			10/3/23
Abdulrahman, Awf; Baykara, Muhammet	Fake News Detection Using Machine Learning and Deep Learning Algorithms	2020	3rd International Conference on Ad...	6/6/23
Addison Howard devrishi, Phil Culliton Yufeng Guo	Natural Language Processing with Disaster Tweets	2019	Kaggle	12/23/23
Ahmad, Iftikhar; Yousaf, Muhammad; Yousaf, Suhail; Ah...	Fake News Detection Using Machine Learning Ensemble Methods	2020	Complexity	6/6/23
Ahmed, Hadeer; Traore, Issa; Saad, Sherif	Detecting opinion spams and fake news using text classification	2018	Journal of Security and Privacy	12/27/23
Al Asaad, Bashar; Erascu, Madalina	A tool for fake news detection	2018	Proceedings - 2018 20th International ...	6/6/23
AI, Maverick	Light Gradient Boosting Machine. Light GBM (Light Gradient Boosting... | by AI Maverick | Medium			12/23/23
Al-Ahmad, Bilal; Al-Zoubi, Ala' M.; Khurma, Ruba Abu; Aljarah, Ibra...	An evolutionary fake news detection method for covid-19 pandemic information	2021	Symmetry	6/6/23
Arora, Suvrat	Sentiment Analysis Using Python	2022	Analytics Vidhya	6/28/23
Baarir, Nihel Fatima; Djeffal, Abdelhamid	Fake News detection Using Machine Learning	2021	2020 2nd International Work...	6/6/23
Belova, Gabriela; Georgieva, Gergana	Fake News as a Threat to National Security	2018	International conference KNOW...	6/7/23
D'Ulizia, Arianna; Caschera, Maria Chiara; Ferri, Fernando; Grifoni,...	Fake news detection: A survey of evaluation datasets	2021	PeerJ Computer Science	6/6/23
Dame Adjin-Tettey, Theodora	Combating fake news, disinformation, and misinformation: Experimental evidence for media literacy education	2022	Cogent Arts & Humanities	6/7/23
Dubey, Praveen	An introduction to Bag of Words and how to code it in Python for NLP	2018	FreeCodeCamp	6/28/23
Erika D	Accuracy, Recall & Precision. When it comes to evaluating how well a... | by Erika D | Medium	2019	Medium	6/30/23
Feed, Implemented	Feed Forward Neural Network Definition | DeepAI			12/23/23
Ferreira, Caitlin Candice; Robertson, Jeandri; Kirsten, Mar...	The truth (as I see it): philosophical considerations influencing a typology of fake news	2020	Journal of Product and Brand Manage...	6/6/23
GeeksforGeeks	Removing stop words with NLTK in Python - GeeksforGeeks	2022		6/27/23
George, McIntire	fake_real_news_dataset	2017		12/27/23
Gilda, Shlok	Notice of Removal: Evaluating machine learning algorithms for fake news detection	2017	IEEE Student Conference on Re...	6/6/23
Gupta, Ritesh	Accuracy, Precision, Recall, F-1 Score, Confusion Matrix, and AUC-ROC	2023	Medium	12/29/23
Javatpoint	Machine Learning Random Forest Algorithm	2021	Www.Javatpoint.C...	10/6/23
JavatPoint	Logistic Regression in Machine Learning - Javatpoint	2018		10/6/23
Jruvika	Fake News Detection	2017	Kaggle	12/27/23
Kaliyar, Rohit Kumar; Goswami, Anurag; Narang, Pratik	FakeBERT: Fake news detection in social media with a BERT-based deep learning approach	2021	Multimedia Tools and Applications	6/6/23
Kannan, Shreya; Saravanan, Shreyaa; Chandirasekeran, Pree...	Detection of Fake News related to COVID-19 using Natural Language Processing	2021	2021 Asian Conference on Inn...	6/8/23
Khan, Junaed Younus; Khondaker, Md. Tawkat Islam; Afroz, Sadia;...	A benchmark study of machine learning models for online fake news detection	2021	Machine Learning with Applications	6/6/23
Khanam, Z; Alwasel, B N; Sirafi, H; Rashid, M	Fake News Detection Using Machine Learning Approaches	2021	IOP Conference Series: Materials ...	6/5/23
Kumar, Ranjit	Research Methodology a step-by-step guide for beginners	2011		11/20/22
Lazer, David M.J.; Baum, Matthew A; Benkler, Yochai; Berinsky, Ad...	The science of fake news: Addressing fake news requires a multidisciplinary effort	2018	Science	6/7/23

☆		Lifferth, William	Fake news	2018		12/27/23
☆		Madrid, Pamela	USC study reveals the key reason why fake news spreads on social media	2023	USC news	6/7/23
☆		Manzoor, Syed Ishfaq; Singla, Jimmy; Nikita	Fake news detection using machine learning approaches: A systematic review	2019	Proceedings of the International Conf...	6/6/23
☆		Menard, Scott	An Introduction to Logistic Regression Diagnostics	2014	Applied Logistic Regression Analysis	10/6/23
☆		Mishra, Shubha; Shukla, Piyush; Agarwal, Ratish	Analyzing Machine Learning Enabled Fake News Detection Techniques for Diversified Datasets	2022	Wireless Communications a...	6/5/23
☆		Mondal, Arnab	LightGBM in Python | Complete guide on how to Use LightGBM in Python	2022	Analytica Vidya	12/23/23
☆		Monti, Federico; Frasca, Fabrizio; Eynard, Davide; Mannion, Damo...	Fake News Detection on Social Media using Geometric Deep Learning	2019		6/6/23
☆		Morde, Vishal	XGBoost Algorithm: Long May She Reign! - Towards Data Science. Medium.	2021		6/29/23
☆		Mridha, M. F.; Keya, Ashfia Jannat; Hamid, Md Abdul; Mono...	A Comprehensive Review on Fake News Detection With Deep Learning	2021	IEEE Access	6/5/23
☆		Narwal, Bhawna	Fake News in Digital Media	2018	Proceedings - IEEE 2018 International...	6/7/23
☆		Nasir, Jamal Abdul; Khan, Osama Subhani; Varlamis, Iraklis	Fake news detection: A hybrid CNN-RNN based deep learning approach	2021	International Journal of Information Ma...	6/6/23
☆		Oppermann, Artem	What Is CatBoost?	2023		12/27/23
☆		Pollock, Alex; Berge, Eivind	How to do a systematic review	2018	International Journal of Stroke	7/4/23
☆		Sahoo, Somya Ranjan; Gupta, B B	Multiple features based approach for automatic fake news detection on social networks using deep learning	2021	Applied Soft Computing	6/6/23
☆		Seddari, Noureddine; Derhab, Abdelouahid; Belaoued, Mohame...	A Hybrid Linguistic and Knowledge-Based Analysis Approach for Fake News Detection on Social Media	2022	IEEE Access	6/6/23
☆		Shu, Kai; Sliva, Amy; Wang, Suhang; Tang, Jiliang; Liu, Huan	Fake News Detection on Social Media: A Data Mining Perspective	2017		6/6/23
☆		sirjan13	Readability Index in Python(NLP) - GeeksforGeeks	2021		6/28/23
☆		Staples, Mark; Niazi, Mahmood	Experiences using systematic review guidelines	2007	Journal of Systems and Software	7/4/23
☆		Thota, Aswini; Tilak, Priyanka; Ahluwalia, Simrat; Lohia, Nibrat;...	Number 3 Article 10 2018 Part of the Artificial Intelligence and Robotics Commons Recommended Citation Recommended Citatio...	2018	SMU Data Science Review	6/6/23
☆		Umer, Muhammad; Imtiaz, Zainab; Ullah, Saleem; Mehmoo...	Fake news stance detection using deep learning architecture (CNN-LSTM)	2020	IEEE Access	6/8/23
☆		Unisa	Policy on research ethics	2007	Research Department, Unive...	11/20/22
☆		Varma, Rajshree; Verma, Yugandhara; Vijayvargiya, Priya;...	A systematic survey on deep learning and machine learning approaches of fake news detection in the pre- and post-COVID-1...	2021	International Journal of Intelligent Com...	7/6/23
☆		Wani, Apurva; Joshi, Isha; Khandve, Snehal; Wagh, Vedang...	Evaluating Deep Learning Approaches for Covid19 Fake News Detection	2021	Communications in Computer and Info...	6/5/23
☆		Waweru Muigai, Jane Wambui	Understanding Fake News	2019	International Journal of Scientific and R...	6/7/23
☆		Wood, Thomas	F-Score Definition | DeepAI	2019	DeepAi.org	10/6/23
☆		Yash, R	Python | Stemming words with NLTK	2022	GeeksforGeeks	6/27/23

7.2 Appendix B: Plagiarism Pledge

Plagiarism Pledge

1. I have read Unisa's plagiarism policy.
2. I understand Unisa's plagiarism policy.
3. I agree to abide by Unisa's plagiarism policy.
4. I have read the direct copying, plagiarism, and "patch-writing" document.
5. I understand what direct copying, plagiarism, and "patch-writing" is.
6. I undertake to avoid copying directly, plagiarism and patch writing.
7. All academic work, written or otherwise, that I submit is expected to be the result of my own skill and labour.
8. I understand that, if I am guilty of the infringement of breach of copyright/plagiarism or unethical practice, I will be subject to the applicable disciplinary code as determined by Unisa.
9. I understand that it is my responsibility to use Turnitin (or similar plagiarism tool) to check submitted research for direct copying or plagiarism.
10. The supervisor has the right and responsibility to return any work to be revised if plagiarism is detected.

Student's name: Brandon Ashley

Student number: 63017822

Student Signature

Date: 01/02/2024