A PREDICTIVE INTELLIGENCE MODEL FOR DETECTION OF FAKE NEWS ON ONLINE MEDIA

TITUS KALWEO KAMWIRA

A PROJECT SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY IN THE SCHOOL OF COMPUTING AND MATHEMATICS IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE OF MASTER OF SCIENCE IN DATA SCIENCE OF THE CO-OPERATIVE UNIVERSITY OF KENYA

2025

DECLARATION

Declaration by the candidate

This project is my original work and has not been presented for award of a degree in any other University or for any other award.

22/11/2025

Signature Date

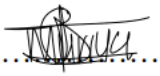
Name of candidate and Reg. Number

TITUS KAMWIRA KALWEO C004/600099/2023

Declaration by the supervisors

We confirm that the work reported in this project was carried out by the candidate under our supervision and has been submitted with our approval as university supervisors

Dr David Muriuki



……22/11/2025……

Signature Date

The Co-operative University of Kenya.

Dr Kennedy Hadullo

……Khadullo……

……22/11/2025……

Signature Date

The Technical University of Mombasa.

DEDICATION

This is dedicated to my late mother, who always rooted for my education and dreams, this is for you mum.

AKNOWLEDGEMENT

I owe my deepest gratitude to my supervisors Dr Muriuki and Dr Kennedy Hadullo for their invaluable contributions, suggestions, timely advice and their insights. The advice from my supervisors resulted in an enjoyable research process that also simplified my work. I am super grateful to Dr Andrew Kipkebut, Dr Fidelis Mukudi and Prof John Kihoro whose academic contributions over the course have been quite influential in my research. I want to express my gratitude toward Cooperative University of Kenya for providing all the assistance I needed.

Table of Contents

DECLARATION ii
DEDICATION iii
AKNOWLEDGEMENT iv
LIST OF TABLES vii
LIST OF FIGURES viii
LIST OF ABBREVIATIONS & ACRONYMS ix
DEFINITION OF TERMS x
ABSTRACT xiii
1. CHAPTER 1: INTRODUCTION 1
1.1 Background of the Study 1
1.2 Problem Statement 2
1.3 Research Objectives 3
1.4 Research Questions 3
1.5 Significance of the Study 3
1.6 Scope of the Study 4
1.7 Limitations of the Study 5
CHAPTER 2: LITERATURE REVIEW 6
2.1 Introduction 6
2.2 Conceptual Framework 12
2.3 Critique of Literature 13
2.4 Research Gaps 15
CHAPTER 3: METHODOLOGY 17
3.1 Introduction 17
3.2 Research Philosophy 17
3.3 Research Design 19
3.4 Study Area 20
3.5 Target Population 20
3.6 Sampling Design 21
3.7 Data Collection 21
3.8 Data Collection Procedures 22
3.9 Data Analysis and Presentation 22
3.10 Data Analysis 24

3.11 Empirical Methods and Hypothesis Testing 26
CHAPTER 4: DATA ANALYSIS, PRESENTATION AND INTERPRETATION 27
4.1 Introduction 27
4.2 Descriptive Analysis 27
4.3 Key Indicators of Fake News on Online Media (Objective 1) 29
4.3.1 Linguistic and Lexical Patterns 29
4.3.2 Semantic Inconsistencies and Misrepresentation of Facts 30
4.3.3 Stylistic and Visual Cues 31
4.3.4 Structural and Source-Related Features 31
4.4 A BERT Model for Fake News Detection. (Objective 20 32
4.5 Model Evaluation 34
4.6 Results 35
CHAPTER 5: DISCUSSION AND SUMMARY 41
5.1 Introduction 41
5.2 Discussion Findings 41
5.3 Study implications 44
5.4 Study Limitations 45
5.5 Recommendations 45
5.6 Conclusion 46
References 48
Appendices 50
Appendix 1: NACOSTI License 50
Appendix 2 - Similarity Index Test and AI content test Reports 51
Appendix 3: Published Article – First Page 53

LIST OF TABLES

Table 5.2: Appendix 2: Similarity and AI report……..51

LIST OF FIGURES

Figure 2.2 Conceptual Framework Infographic……10

Figure 3. Dataset distribution analytics snippet……..23

Figure 3.1a): Data loading code snippet…….24

Figure 3.1b): Train – test split…….……...25

Figure 4.2 Distribution of Fake and True news articles in the dataset……..27

Figure 4.21 Histogram showing typical word length……28

Figure 4.6 Summary of the evaluation metrics…….……35

Figure 4.62 Confusion Matrix…….…….…….…….37

Figure 4.62 ROC Curve……..…….…….…….…….39

LIST OF ABBREVIATIONS & ACRONYMS

AI - Artificial Intelligence

ML - Machine Learning

TP - True Positive

ISOT - International Social Ontology of Texts.

BERT - Bidirectional Encoder Representations from Transformers

RNN - Recurrent Neural Network

NN - Neural Network

SVM - Support Vector Machine

DEFINITION OF TERMS

Artificial Intelligence is the development of computer systems that perform tasks requiring human intelligence, such as learning, reasoning, language understanding, and decision-making. It involves creating algorithms and models that allow machines to simulate intelligent behavior.

Machine Learning is a subset of AI where computers learn from data to make predictions or decisions without being explicitly programmed. It uses statistical techniques to enable systems to improve their performance on a task over time.

Natural Language Processing is a field of AI that enables machines to understand, interpret, and generate human language, especially text and speech. It combines computational linguistics with machine learning to process language data effectively.

Machine learning is one of Deep Learning's main types which works by using neural networks with several layers to discover complex patterns inside lots of data. This technology is often used in image and speech recognition tasks.

Spam entails the method of using a program to place categories on text, including separating news stories by topic and paying attention to whether or not an email.

Web Scraping refers to a process where information is collected from websites automatically with the help of programs or scripts which you can examine later.

Tokenization is when we break text into smaller units such as words, phrases or sentences, to make it easier to study.

Lemmatization means transforming a word to its dictionary form, taking its meaning from the context (for instance, "was" becomes "be").

Stemming: In stemming, you simplify a word by removing the suffix, leaving behind the base word (e.g., "running" becomes "run"), doing so without paying attention to the context.

Overfitting A model can become overfitted when it does great on data used to train it, but doesn't do so well on new information due to including irrelevant training set details.

Confusion Matrix is a table that shows the results of a classification algorithm by displaying true positive, false positive, true negative and false negative findings**.**

Bag of Words (BoW) is a basic technique that lists each word in a document and counts its frequency no matter how they are used.

TF-IDF (Term Frequency–Inverse Document Frequency) method is used to see the significance of a word in a document against the whole group of documents. The rate increases as the word becomes more common in one document and drops as the word comes up more often in the entire collection.

Word Embeddings entails the use of numbers to describe words in space, so related or similar words are placed near each other.

Precision is the ratio of correctly predicted positive cases to all predicted positive cases. It tells us how many of the items labeled as positive are actually correct.

Recall is the ratio of correctly predicted positive cases to all actual positive cases. It measures the model's ability to find all relevant instances in a dataset.

F1 Score is the harmonic mean of precision and recall. It provides a single metric that balances both false positives and false negatives.

True Positive refers to instances correctly classified as positive (e.g., fake news correctly identified as fake). It indicates when a model makes a correct prediction for a positive case.

Bidirectional Encoder Representations from Transformers is a deep learning model by Google that understands context in text by analyzing words in both directions. It allows machines to grasp the meaning of words based on both the left and right context in a sentence.

Recurrent Neural Network refers to a neural network designed for sequential data, retaining past input information to make future predictions. It is commonly used in language modeling, time-series analysis, and speech recognition.

Neural Network refers to a model inspired by the human brain that processes data through interconnected layers to recognize complex patterns. It consists of layers of nodes (neurons) that transform input data into useful outputs.

Support Vector Machine is a classification algorithm that finds the best boundary separating different classes in the dataset. It works by finding the hyperplane that maximizes the margin between data classes.

Random Forest is an ensemble learning method using multiple decision trees to improve classification accuracy and reduce overfitting. It combines the results of many decision trees to produce more reliable and stable predictions.

ABSTRACT

The emergence of social media on the internet has increased the rate at which people spread misinformation and fake news, posing serious societal, political, and economic dangers. This paper created a predictive intelligence model to detect fake news, guided by the objective of developing and evaluating a transformer-based model for classifying news as true or fake while also assessing its performance relative to traditional machine learning baselines. Advanced natural language processing and machine learning methods were used to define news as true or fake by applying the Bidirectional Encoder Representations from Transformers (BERT) model. Training and testing of the model were conducted using a Jupyter Notebook with the assistance of a GPU on Google Colab and a labeled Fake and True news dataset available on the Kaggle repository. Accuracy, precision, and recall were used to evaluate performance, and it was found that BERT outperformed the traditional machine learning baselines, achieving higher accuracy while preserving good precision and recall. In addition to technical performance, the study highlights the need to make detection systems local to the language and recommends the inclusion of Kiswahili, Sheng, and native dialects to make them more inclusive and relevant in the real-life Kenyan setting. The results correspond to existing literature indicating the success of transformer-based models in countering misinformation in dynamic web-based contexts. The impact of this study is both theoretical, through the development of deep learning methods for misinformation detection, and practical, as the study provides a scalable framework for policymakers, social media sites, and fact-checking institutions. Further research on multilingual adaptation, integration with real-time monitoring, and ethical safeguards against careless utilization of AI in the fight against misinformation is recommended

1. CHAPTER 1: INTRODUCTION

1.1 Background of the Study

The modern digital age generates substantial misinformation and news which develops into a global issue that damages both political choices and economic systems and social bonds. The fast distribution of information and news becomes achievable because people depend more on digital communication since social media continues to expand (Capuano et al., 2023). This research seeks to develop a machine learning approach that detect fake news and misinformation. Digital connectivity benefits established conditions where misinformation spreads freely through digital networks to give wrong and incorrect perception or picture. Public perception manipulation occurs when misinformation enters the system because this leads to social disruption that weakens trust between institutions and media platforms (Zhang et al., 2020). Political deception stands alongside healthcare misinformation and financial deceit and theoretical conspiracies as the primary forms of misinformation that affect different parts of society in various ways.

According to research by Grandon et al. (2021), artificial intelligence creates new opportunities for revolutionizing old keyword filtering methods. The detection abilities of context and intent are utilized by modern AI-based procedures. The accuracy of decision-making has increased substantially while the speed of decision processes has speeded up for both sentiment analysis and content moderation through current models. Rao et al. A new AI-based framework for context-aware analysis with machine learning algorithms is reviewed by Rao et al. (2020). According to Rao et al. Language features combined with metadata alongside user engagement actions deliver the most impact on how well software assigns categories (2020). The predictive intelligence model development objective harmonizes with the system because it uses multiple variables for input and adaptive learning functions. The application of AI for detecting patterns in fake news while making predictions simultaneously results in improved models with both precision and accuracy.

A new study by Ceylan, Wood and Madrid (2023) points out that fake news becomes popular on social media due to strong emotions. Results suggest that people are more likely to share false news because it has a bigger emotional effect than news with facts. Because of this, stories that aren’t accurate get more views which means more people are likely to see them.

The issue turns up locally too. Faustine Ngila (2020) reported that Kenya has suffered an increase in fake news and bullying found on the internet, mainly during periods affected by politics. There has been an increase in Kenya of social media users having to deal with made-up stories on political, health and security subjects. As a result, individuals are deceived and society's unity is threatened as well as personal safety.

Moreover, Tajrian et al. The authors of Tajrian et al. (2023) review different analysis methods for fake news and analyze their respective advantages and limitations. He above research shows that combining machine learning algorithms with feature extraction concepts in hybrid models offers more accurate outcomes and generalization performance than when using standalone models. Additionally, reviewing continuous model validation using different datasets is emphasized to ensure the effectiveness of the model across platforms and topics. This is in line with the focus of the current study which is to validate the accuracy, precision and reliability of the predictive model developed in order to maintain the effectiveness of the developed system in its real-world applications.

1.2 Problem Statement

The rapid spread of fake news on digital platforms poses significant threats to public trust, democratic processes, and social stability. Despite advancements in machine learning and natural language processing detecting fake news remains challenging due to: Dataset Limitations: Many datasets (e.g., ISOT, FakeNewsNet) suffer from biases and lack generalizability (Ahmed, 2021). There is need to create analytical systems that scale automation to detect misinformation together with appropriate detection methods and response strategies.

One of the key specific gap for existing fake news model is failure to achieve both effectiveness and scalability which leads the research to use artificial intelligence with machine learning and natural language processing tools for solution development (Rao et al., 2020). Traditional fake news detection technologies have a great challenge in relying on static datasets, especially in instance where there is presence of cross-linguistic or cross-cultural variations (Bali et al., 2019;

Meel & Vishwakarma, 2020). To add, the growth in organized bot networks and echo chambers hinders the ability to distinguish between fake news and real news (Zhou and Zafarani, 2020; Raza and Ding, 2022). Constant adaptation of fake news creators' strategies by utilizing advanced ways that involve deepfake technology, AI bots, and manipulated multimedia (Ilie et al., 2021; Hua & Shaw, 2020).

The purpose of the study is developing a robust predictive intelligence model which is able to leverage machine learning to detect fake news and misinformation. The purpose has been achieved through creating data-driven models that provide usable solutions for swift and correct detection of misinformation. Safety improvements in digital environments and upgraded choices for information consumers form the core objectives from these solutions.

1.3 Research Objectives

1.3.1 General Objectives.

To develop a predictive intelligence model for detection of fake news in social media using Machine learning model.

1.3.2 Specific Objectives.

i. To investigate the key indicators of fake news on online media.
ii. To develop a BERT model for detection of fake news.
iii. To validate the developed BERT model by accessing its accuracy, precision and reliability.

1.4 Research Questions

i. What are the key indicators of fake news in online media?
ii. How can a BERT model be developed to effectively detect fake news on online media?
iii. How accurate and reliable is the BERT machine learning algorithms in detecting fake news?

1.5 Significance of the Study

This research is significant as it leads to an expansion of fake news and misinformation propagation detection capabilities through scalable machine learning solutions. This will translate

into assisting the policymakers, fact checkers and social media to detect fake news and misinformation to promote a more enlightened society.

This paper combines machine learning and natural language processing in order to introduce the practical solution to the real time eradication of misleading information on the web by the identification of it. This study will aid the preventive attempts to maintain the integrity of the digital through the improvement of the accuracy and efficiency of fake news detectors. The results may be also applied to raise awareness among people and media literacy programs, with the help of which individuals will be able to view and draw the line between the sources of information being disseminated.

The socio political significance of the study also exists, in the world which is becoming more and more reliant on digital content and where misinformation can be of dire consequences, in situations including elections, population health and safety. The study can be used to close that gap and bring data science and social good and present good solutions so that that real world impact may occur.

The scholarly importance of this research lies also in the fact that it advances the corpus of knowledge in the areas of computational linguistics, data science and media studies. It gathers IT specialists, communication specialists, and social scientists to collaborate to address the problem of misinformation because it is a multifaceted issue. Lastly, our approach and instruments may serve as building blocks of future studies on automated content verification, sentiment analysis, and digital forensics thereby broadening the scope of application of fake news detection within the academic and practical realms.

1.6 Scope of the Study

This paper aims at proposing and analyzing a predictive intelligence model that can be used to identify fake news in online social media settings. It narrows down its services specifically to text-based news items and posts as opposed to multimedia content (e.g. images or videos). The study will use machine learning and natural language processing methods to categorize news articles as true or fake. To have a strong training base, the model will be trained on labelled datasets obtained at Kaggle and enriched with articles that are scrapped on credible online news sites.

The entire data preparation pipeline (i.e. web mining, data cleaning, text normalization, stemming, lemmatization and sentiment analysis in order to determine linguistic and emotional indicators of fake news) is also covered. Only the deep learning method and the supervised one, specifically the Bidirectional Encoder Representations from Transformers (BERT), will be compared in terms of binary classification. Measures of evaluation will focus on typical measures including the accuracy, recall, F1-score, and reduction of false positives and negatives. While the research will offer actionable insights for policy makers and technology platforms, it may not cover image manipulation, deepfake detection, or multilingual contexts other than the English language.

1.7 Limitations of the Study

This research has some restrictions which limit its possible benefits. Detecting fake news is highly dependent on the quality and diversity within the datasets used. The public datasets LIAR and FakeNewsNet fail to capture modern news misinformation patterns thus leading to risks against the model's generalization capacity. The combined presence of subtle biases and context-dependent mockery as well as partial misinformation misleads the detection system into producing false errors. The power of machine learning and NLP algorithms can handle various content types effectively yet their functionality depends on the selected features and algorithms.

Digital-based models should be retrained frequently to revise their algorithms because they undergo rapid evolutionary shifts as the misinformation strategies evolve with their digital platforms. Evaluation of ethical issues surrounding censorship and the issue of trust with users and misclassification mistakes is the most important thing before the implementation of such systems are deployed to real-life situations. These operational limitations imply the urgent need to develop models and high-scale training data gathering and subsequent analysis by a number of experts.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

The fast pace of development of online social media has altered the manner in which information is created, shared, and consumed. Such platforms like Facebook, X (previously Twitter), and WhatsApp allow sharing news and opinion in real-time regardless of geographical and cultural borders. Although these technologies have increased the level of connectivity and democratization of information access, they have also enabled the proliferation of misinformation and disinformation at an unprecedented level. The issue of fake news is a serious threat to the trust of the people, social unity and even national security, especially in the developing democracies like Kenya. In turn, the process of identifying and addressing the problem of fake news on social networks in the online context has become a key research topic of interest among both scholars and practitioners.

The methods of machine learning and more recently deep learning have given promise in solving this problem. The initial methods of misinformation detection were based on hand-checking of facts, keywords identification, or network analysis, which was not scalable and inaccurate. As more powerful Natural Language Processing (NLP) models have emerged, e.g. Bidirectional Encoder Representations from Transformers (BERT), text classification has become revolutionized by applying contextual meaning across both directions of a sentence. BERT and its variants have demonstrated the state of the art performance in a broad range of NLP tasks, such as sentiment analysis, topic detection, and fake news classification. In spite of these developments, the literature on fake news detection is still strongly biased with respect to English-language data and Western settings. Only a limited number of studies have investigated misinformation among African or Kenyan dialects like Kiswahili, Sheng and other dialects. This discrepancy restricts the relevance of the available models to local settings in which linguistic diversity and cultural particularities affect the way in which misinformation is constructed and distributed. These distinctive features are critical to understanding how to come up with predictive intelligence models that are accurate, culturally sensitive and scalable.

The following literature review thus provides the history of fake news detection systems on social media, with three key threads: (1) classic machine learning, (2) transformer-based NLP models, e.g., BERT, and (3) works on non-English or multilingual settings. The review offers the theoretical and empirical basis of the creation of the predictive intelligence model specific to the Kenyan social media environment by critically reviewing the literature, pinpointing the gaps, and pointing out the opportunities to develop a localized model.

2.1.1 Review of Fake News Features and Indicators

The identification of fake news on online platforms has attracted significant scholarly attention, with researchers outlining a variety of linguistic and content-level features that separate misleading articles from authentic news. Linguistic cues remain some of the most frequently referenced indicators. Fake news often contains sensational or exaggerated language that is crafted to provoke emotional reactions and attract rapid attention. Headlines are frequently written in a clickbait style, relying on dramatic phrases, strong adjectives, and subjective expressions. Many studies also observe that such articles tend to have weaker sentence construction, limited factual depth, and occasional grammatical errors. These linguistic patterns help form the initial basis for detecting fake news because they reveal clear differences in writing quality and editorial standards compared to legitimate journalism.

Content characteristics provide another important class of indicators. A large proportion of fake news stories lack credible evidence, verifiable data, or properly referenced sources. Researchers frequently describe misuse of images, fabricated statistics, and ambiguous claims as consistent traits of misinformation. It is also common for these articles to cite unnamed experts or rely on sources that cannot be traced. Unlike established news publishers that follow strict editorial protocols, misinformation outlets often sacrifice accuracy in favor of sensational storytelling. As a result, content-level weaknesses become valuable cues for both manual investigations and automated detection systems that attempt to distinguish factual narratives from fabricated ones.

Structural and psychological aspects further strengthen the literature on fake news indicators. Structural features include characteristics such as the age of the domain, posting frequency on the website, the consistency of author information, and the overall design quality of the platform. Fake

news sites often have recently registered domains, limited or missing author details, and irregular publishing patterns that raise questions about authenticity. Psychological indicators are also widely documented. Fake news writers frequently craft stories that appeal to strong emotions such as fear, anger, or shock. This type of framing is designed to influence how readers perceive the information and to encourage rapid sharing. Emotional manipulation has been shown to reduce critical thinking, which makes it easier for false narratives to spread.

Another important category of indicators relates to user behavior and the manner in which information moves across social networks. Fake news tends to spread more quickly and to reach wider audiences than credible information. This pattern is often connected to coordinated sharing, automated bots, or groups of users who consistently circulate provocative or unverified content. Unusual engagement spikes, rapid resharing within specific online communities, and clustering among tightly connected accounts provide strong signals of misinformation activity. Researchers also examine the behavior of individual users who frequently post or amplify suspicious material. These patterns of diffusion are essential because they highlight how fake news moves across platforms and how certain features of propagation differ from the spread of verified news. Together, these indicators show that identifying fake news requires a broad understanding of linguistic, content-based, structural, psychological, and network-related characteristics..

2.1.2 BERT and Transformer-Based NLP Model

Transformer architectures were introduced and revolutionized natural language processing. BERT and its variants are able to capture both forward and backward contextual information and can be trained on particular tasks using relatively small labelled datasets. Ilie et al. (2021) compared a number of deep learning models on word embeddings and observed that transformer-based models were better than CNN and LSTM baselines at detecting misinformation. Capuano et al. (2023) also find that deep learning and transformers in particular, achieve state-of-the-art performance on various fake news datasets. Mishima and Yamana (2022) further introduce the fact that explainability methods are being steadily incorporated into these models to enhance the trust and explainability in sensitive areas like banking and policy. Another AI-based solution that was

suggested by Rao et al. (2020) is the use of transformers with domain-specific features to enhance the detection rates.

The benefits of BERT-based systems are that they can be trained to learn deep semantic and syntactic representations without a major feature engineering effort and transfer to cross-domain and adapt more quickly to new misinformation topics. They also facilitate multi-task learning and multilingual fine-tuning which may be especially helpful in heterogeneous social media settings. These models are however computationally costly, and tuning of hyperparameters is necessary to prevent overfitting on small data sets. As Mishima and Yamana (2022) point out, transparency and fairness are also still urgent concerns in transformer-based models, particularly when used in a regulatory or high-stakes setting. However, the literature shows repeatedly that the BERT-style models achieve a higher precision, recall and F1-score than the common feature-based classifiers, which is the sign of a paradigm shift in the research of fake news detection.

2.1.3. BERT Model Validation

Machine learning models used to detect fake news are often validated by using strict evaluation metrics and data partitioning techniques. Most systematic reviews underline the necessity to report the typical classification measures, like accuracy, precision, recall, F1-score and area under the ROC curve, as each of the measures highlights a different facet of the model performance and prevents the misleading conclusions based on the results of only one measure (Ahmed et al., 2021; Capuano et al., 2023). It is also emphasized in reviews and surveys that confusion matrices and class-specific measures must be reported in case of imbalanced datasets because accuracy can mask poor performance on minority classes (Tajrian et al., 2023; Mishra et al., 2022). Some of these measures, such as precision-recall curves and PR-AUC on highly skewed problems, are advised to supplement these measures by some authors, along with Matthew correlation coefficient or Cohen kappa as resilient alternatives that include chance agreement (Zhang et al., 2020; Capuano et al., 2023).

The important elements of establishing reliable validation are procedure and experimental design. The literature identifies typical protocols to consider the variability in model training, including stratified train-test splits, k-fold cross-validation, and repeated runs with alternative random seeds,

in particular, deep transformer models (Ilie et al., 2021; Khanam et al., 2021). It is suggested to benchmark performance of BERT against baseline algorithms and previous published results and position the performance of BERT in relation to simpler models and other deep architectures (Ilie et al., 2021; Rao et al., 2020). Repeated emphasis is placed on external validation on held-out datasets, which are based on different platforms, languages, or time periods since cross-platform and temporal generalizability are primary weaknesses in the existing literature (Ahmed et al., 2021; Mishra et al., 2022). Other authors also suggest ablation experiments where the contribution of contextual embeddings of BERT can be measured against surface features or user-level cues by eliminating or randomly diverging sets of features (Capuano et al., 2023; Tajarian et al., 2023).

Reliability and robustness testing are not limited to point estimates of correctness but consider the stability and credibility of the model results. The recent literature is demanding statistical significance testing and uncertainty estimation by using methods of bootstrapped confidence intervals, paired classifier comparison test by McNemar, and reporting variance across a series of trials to prove the stability of results (Zhang et al., 2020; Ilie et al., 2021). Claims of reliability are also explained by explainability and interpretability to help the model to understand the reasoning behind the decision it makes, which helps to analyze errors and provide stakeholders with trust, surveys on explainable fake news detection describe saliency methods, attention analysis, and local explanation one of the tools that can help analyze and validate model reasoning beyond raw scores (Mishima and Yamana, 2022; Gradoń et al., 2021). Robustness assessments which model noise, adversarial noise, or domain shifts are also becoming more and more suggested since they can expose brittle behaviour which can be overlooked by accuracy and precision metrics on a single test set (Segura-Bedmar and Alonso-Bartolome, 2022; Kansara and Adhvaryu, 2024).

Combining these methodological suggestions, one can create a validation blueprint to which the current study can appeal. In particular, the literature advocates a combination of stratified cross-validation and external hold-out set, which documents a set of measures, such as accuracy, precision, recall, F1, ROC-AUC and PR-AUC, calculates confidence intervals and variance across runs, makes use of statistical tests when comparing models, conducts ablation studies to isolate the contribution of BERT, uses interpretability techniques to inspect model decisions, and conducts robustness checks of domain or temporal shifts (Ahmed et al., 2021; Ilie et al., 2021; Capu The

adherence to these practices will give a holistic evaluation of the developed BERT model accuracy, precision and reliability and will make this work consistent with best practice in the existing fake news detection literature.

2.1.4 Research on Non-English or Multilingual situations.

Most of the studies are on English-language data even though transformers are successful. Segura-Bedmar and Alonso-Bartolomé (2022) note that multimodal and multilingual fake news detection remains in its infancy with low-resource languages having limited labelled corpora. Kansara and Adhvaryu (2024) note that the use of big-data and sentiment-analysis solutions has potential in non-English settings, but it requires culturally specific features. Gradoń et al. (2021) emphasize that misinformation is best combated using multidisciplinary and cross-cultural solutions rather than through technical solutions. Tajrian et al. (2023) survey fake news analysis methodologies and note that benchmark datasets are not available in such languages as Kiswahili, Arabic, and Hindi which restricts model performance. This is also reflected in Mintoo (2024), who advocates a shared data-collection effort and cross-lingual transfer learning to enhance the accuracy of multilingual environments.

In general, in African and other low-resource contexts, misinformation tends to assume localised accounts, blended languages and culturally particular signatures, and it is problematic to apply English-trained models directly. Cross-lingual transfer learning and multilingual transformers (e.g. mBERT or XLM-R) promise but are currently worse than English-only models because of domain mismatch and the limited availability of annotated data. Researchers have started to experiment with community annotation, culturally aware feature engineering and hybrid multimodal methods, which take into account both text and images or video to enhance detection (Segura-Bedmar and Alonso-Bartolomé, 2022). The literature thus indicates that there is a pressing necessity to have more comprehensive datasets and approaches that incorporate the linguistic diversity of the online space especially in Kenya and East Africa where Kiswahili and Sheng have taken over most of the social media discourse.

2.1.5 Machine Learning algorithms on predicting Fake News

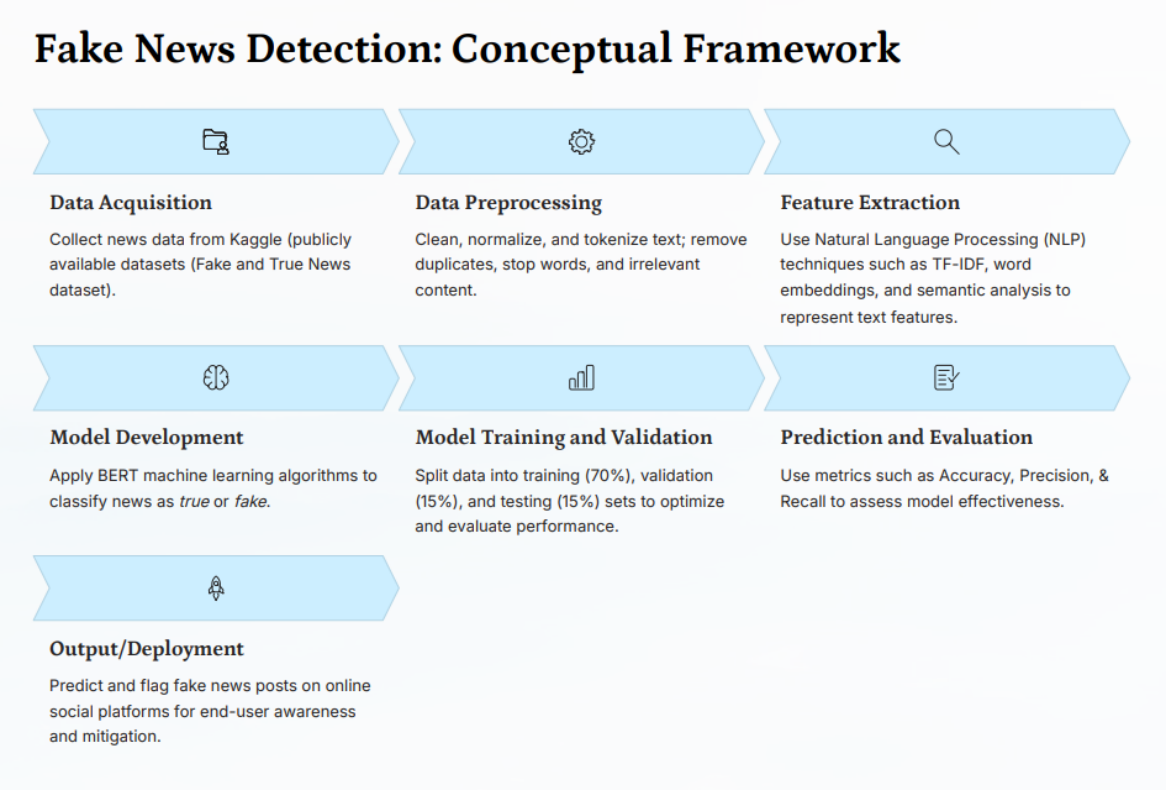
Over the year, various researchers have delved into determining how machine learning can be best leveraged in prediction of whether a piece of news is original or fake. In a paper on sentiment cues and large scale data analytics, Kansara and Adhvaryu (2024) reviews how news narratives can use sentiment cues and large scale data analytics to identify inconsistencies in news narratives. According to their survey, sentiment analysis models that combine it with natural language processing techniques provide a better way to detect misleading content by recognizing the emotional underlining in the content. Considering the growing necessity for real-time processing of big data, in particular with short form content in modern day platforms, they also specified the limits of realizing context aware models which could react more swiftly.

In addition, there have been a number of studies concerning machine learning methods on some particular news domains. Khanam et al. (2021) discussed different classification techniques and concluded that support vector machines, and convolutional neural networks have been very promising in fake filtering news. In addition, Tajrian et al. (2023) reviews the fake news analysis methods exhaustively and point out that explainable AI is key to build the user's trust in the detection tools. The authors then continue and argue that transparent models not only enhance the accuracy of the detection but also help the user to grasp the rationale behind the content classification.

2.2 Conceptual Framework

The study applies a conceptual framework on the basis of data driven analytical means to detect fake news. All of which are owed to the core of the framework resting on which reliable data sources, natural language processing, sentiment analysis and machine learning algorithms. There are linguistic patterns, emotional cues, and inconsistencies of narrative structure that would make fake news potentially detectable. First we will have collected the data from the trusted sources, then we process and extract the features from it. Then, the deceptive language features are shown through the sentiment analysis and NLP techniques. Finally, we will use machine learning classifiers such as Bert Model at the end to classify the fake news from real content with better accuracy and scalability.

Figure 2.2



The above figure 2.2 shows the conceptual Framework for the research

2.3 Critique of Literature

Before reviewing the existing scholarship on fake news detection, it is important to outline how the literature was selected. This study applied the SIGMA model of literature screening, which considers Source quality, Inclusion criteria, Generalizability, Methodological rigor, and Applicability to the study's objectives. In line with this approach, the critique focused on peer-reviewed journal articles, conference papers, and high-quality preprints published between 2018 and 2024. Studies were included if they examined fake news detection techniques, machine learning or deep learning approaches, semantic feature engineering, or user-level interaction indicators. Works were excluded if they lacked methodological transparency, did not present empirical evidence, or were anecdotal commentaries without analytical depth. This process ensures that the scope of the critiqued literature is clearly defined and grounded in a systematic and rigorous selection approach.

There is a lot of literature that has been done on the detection of fake news on online social media, but there are still a number of methodological and conceptual limitations. To begin with, the existing literature pays significant attention to the standard machine learning methods such as logistic regression, SVM, and random forests, which rely largely on surface-level textual features such as word counts or n-grams (Sharma et al., 2023). Although these methods provide baseline performance, they are often unable to capture the deeper semantic and contextual connections within online materials. This weakness limits their cross-platform or cross-language generalizability, especially in dynamic environments where misinformation evolves rapidly.

Second, a significant percentage of the available literature is built on small or homogenous datasets that originate from a single platform such as Twitter or Facebook. This raises concerns about data representativeness and increases the possibility of models becoming overfitted to a particular platform or event (Zhang and Ghorbani, 2022). Very few studies combine textual properties with user interaction patterns or source credibility indicators, even though these combined features provide richer and more holistic insights into the dynamics of fake news. The lack of multimodal or integrated feature sets remains one of the major limitations in current research.

Third, despite the promising outcomes of deep learning and transformer-based models such as BERT in tasks like sentiment analysis and text classification (Devlin et al., 2019), their adoption in fake news detection remains inconsistent. In many of the studies that use transformers, the models are treated as black boxes without any interpretability mechanisms to explain how predictions are made. This reduces transparency and undermines trust among stakeholders, making real-world deployment more difficult. In addition, the moderating or contextual factors that influence the spread and visibility of fake news, such as platform algorithms, demographic variations, or policy differences, are rarely discussed in the literature. Without incorporating these dimensions, existing models risk oversimplifying the complexity of misinformation ecosystems.

This paper aims to fill these gaps by integrating textual and semantic attributes with user interaction and credibility cues, using a transformer-based architecture (BERT) within a predictive intelligence framework. This combined approach is expected to improve accuracy and generalizability while also offering better insights into contextual factors that affect detection

performance. The integration of multiple feature categories and the use of a robust deep learning model will help address many of the limitations highlighted in the reviewed literature.

2.4 Research Gaps

There are several critical issues that restrict the formulation of effective solutions to fake news despite the existing progress in the field of detecting fake news. The most significant challenge is associated with the adaptation of detection models to the new methods of fake news distributors. The existing strategies keep the datasets unchanged and struggle to keep up with the latest trends such as deepfakes created with the help of bots like Chatgpt and adversarial attacks (Ilie et al., 2021; Hua and Shaw, 2020). Since the fake news producers adapt their tactics to escape the detection mechanisms, the process of real-time monitoring of misinformation turns out to be especially difficult.

The present detection models suffer a huge blow as they lack ability to detect misinformation across languages and various cultures. Since most detection systems operate using English-language data their detection capability becomes restricted when applied to non-Latin script languages and low-resource language sets (Bali et al., 2019). The exploration of cultural aspects in misinformation spread patterns along with reader perception has received limited research attention since it hinders broad international implementation of these detection systems (Klein & Wueller, 2017). Advanced detection models based on deep learning architectures present an interpretability problem through BERT and CNNs. High accuracy in these models is essential but th4ir complete opaqueness brings trouble for understanding their decision-making mechanisms which results in decreased trust and automation transparency (Mishima & Yamana, 2022).

Although text-based detection has been widely researched, the multimodal approaches that are both effective in incorporating text, images, and videos are in their infancy (Segura-Bedmar and Alonso-Bartolome, 2022). Technical models frequently ignore human-centric contributors to misinformation (e.g., echo chambers and cognitive biases), even when they have a pivotal role in the fake news diffusion (Golbeck et al., 2018). The scalability is another urgent problem because many detection algorithms are computationally intensive in order to be implemented in practice on large scale social media platforms (Saikh et al., 2020). Next, the ethical issue is brought up,

including possible biases or over-stringency of the content moderation (Allcott and Gentzkow, 2017). The only way to seal these gaps will be with strong and equitable answers to combating fake news, and they will require interdisciplinary teamwork, adaptive algorithms, and more comprehensive and varied datasets.

CHAPTER 3: METHODOLOGY

3.1 Introduction

Chapter three reviews the research methodology that was applied to steer the entire workflow of identifying fake news on online media through the use of a predictive intelligence tool and alogorithm. It describes the broad methodology adopted to meet the research objectives, the philosophical orientation adopted by the research and the design strategies adopted to make sure that the findings could be valid, reliable and reproducible. The data sources, the sampling processes, and the methods that were used to clean, preprocess, and analyze the data before developing and evaluating the models have also been recorded in the chapter. These steps are clearly outlined in this chapter to show how the study was carried out in a systematic and transparent manner thus giving a basis on which the findings can be interpreted and the study can be replicated in similar situations.

3.2 Research Philosophy

The research philosophy that was used in this study is positivism research philosophy, which presupposes that reality is objective, measurable, and does not depend on human perception. The positivist perspective of the fake news detection sees the problem of misinformation as an observable phenomenon, which can be quantified and modeled by means of recognizable patterns in language and statistical norms. This is in line with machine learning studies whereby models are trained on the basis of empirical data and measured against a set of performance metrics instead of being subjectively judged.

Positivism, ontologically, presumes that there is one stable reality. In this case, news item is true or false as it is according to the subjective ideas of a news item. This is in line with the binary classification task that BERT model implements, which tries to project textual inputs to a goal ground truth label. Fake news is thus considered a quantifiable and observable condition that is expressed in observable textual characteristics.

Positivism is epistemologically oriented where valid knowledge is achieved through empirical measurements. Machine learning is not out of place in this perspective since models derive patterns on data based on a systematic training. The internal representations of the BERT model are based on the statistical evidence that can be found in large text corpora, and thus, it is able to generalize misinformation cues. Accuracy, precision, recall, and F1-score are some of the evaluation metrics used to measure the performance of the model and the quality of the knowledge.

Positivism is axiological and focuses on the neutrality and minimization of the researcher. This applies to the misinformation research, where interpretation is easily distorted by political or ideological bias. Value-free decision making and minimized subjective interference were ensured by the study through automated preprocessing, standardized evaluation criteria, and reproducible training procedures.

The goals of this research did not fit the interpretivist or qualitative philosophies since they emphasize on subjective meaning and contextual sense. It did not seek to understand how human beings perceive truth but to develop a predictive model that can distinguish fake news based on quantifiable evidence. The machine learning systems are based on measurable characteristics and predictable training processes, which is in line with positivism and not interpretivism.

Determinism of machine learning in controlled conditions is also reflected in the study. Given an identical dataset, hyperparameters, and computer environment, a BERT model will always generalize similar patterns and give similar results. This determinism upholds the positivist anticipation that phenomena may be explained and predicted by using systematic processes. The positivist direction of this research is also based on measurement theory. Measures like accuracy, precision, and recall are formal measurement tools that measure how close the model is to reality. The measurements enable objective comparison of various algorithms, such as baseline models like the Random Forest and SVM. A fundamental requirement of positivism, reproducibility, was met by means of fixed random seeds, well-described preprocessing procedures, a consistent data split, and structured transformer architecture. This enables the other researchers to recreate the experiments and authenticate the experiments in similar conditions.

3.3 Research Design

In this paper, an experimental research design was used based on the principles of supervised machine learning. The main target of the design was to create, train and test a predictor model that can be used to differentiate fake and real news based on textual features. Experimental design was suitable as it permits manipulation of model parameters, systematic comparison of alternative algorithms and quantifiable evaluation of results. The study employed a systematic research process in which the various model configurations were put through controlled tests to establish the predictive quality of the models. The selection of algorithmic model was the major experimental factor. The research tested a fined-tuned BERT model and compared its performance with the baseline machine learning classifiers which also had the Random Forest and Support Vector Machine. These models were used as a reference point to determine whether the transformer-based architecture could be better predictors. The performance measures were the dependent variables. These were accuracy, precision, recall and F1-score. These measures gave objective evidence on the degree to which the models accurately determined fake and real news.

These independent variables were the model architecture, input representation, hyperparameters, and dataset partitioning strategy. The variable of the model architecture distinguished BERT and the base classifiers. The input representation variable outlined the feature extraction method, as it was based on the contextual embeddings of BERT to the transformer model and TF-IDF features to the baseline models. Hyper parameters were batch size, learning rate, epochs, regularisation settings, tree depth of the Random Forest as well as the kernel settings of the SVM model. The process of dataset partitioning was done under stratified sampling with 70:15:15 training, validation and testing split.

The design had various controls that would guarantee internal validity. The random seed was set in order to reduce the variation in different runs. All models were compared using the same partitions of the dataset used. The training and evaluation were done in the same computational setup to remove performance differences with hardware variation. The control of overfitting was done by early stopping, validation monitoring and dropout application to the transformer model. External validity and reproducibility were also the objectives of the research design. The preprocessing pipeline, parameter settings and training configuration were all properly

documented. The publicly available datasets were used so that the experiment could be replicated by other researchers. Generalisation was facilitated by the fact that objective evaluation measures were used to give standardised measures of model performance.

On the whole, the experimental design allowed developing and evaluating models in a systematic manner. It gave a clear framework on how to compare the performance of the BERT with the traditional machine learning models and it made sure that conclusions concerning the better predictive power of the selected model were informed by the uniform and controlled empirical data.

3.4 Study Area

The analysis has been carried out in the field of the online social media content which is one of the major sources of fake news origin and distribution. Even though the real data were retrieved through publicly available FakeRealNews dataset on Kaggle, the articles themselves were the content that was initially released through various online sources such as news websites, blogs and social networking sites. This heterogeneous digital space was an accurate representation of the real-world environment misinformation is spread in and thus a good place to test a predictive intelligence model.

The research region included news reports that majorly policital news across the world. This breadth enabled the model to acquire linguistic and stylistic patterns based on a broad scope of content as opposed to one topic. This variety enhanced the ecological validity of the experiment, because fake news detectors, in many cases, are applied to heterogeneous streams of online information, and not limited, topic-specific sources of data.

3.5 Target Population

The population that was targeted in this study was the online textual news articles that were classified as fake or real. This consisted of news headlines and complete text articles on various topics. In this paper, the term population did not mean people but the collection of digital information on which data were extracted. The emphasis on the textual aspect of online news was not accidental since most of the misinformation detection systems are based on textual analysis as opposed to multimedia. By considering all the articles in the dataset as the target population, the

study was able to sample the entire range of writing styles, rhetorical devices and narrative structures applied in both authentic and fake news. This large population was a realistic training environment of the predictive model and enhanced the chances that the system developed would be generalized to the unseen news items outside the dataset.

3.6 Sampling Design

The sampling design that was used was purposive to pick data to analyze. The FakeReal News dataset was selected purposefully due to the fact that it provided a big, varied and publicly recorded dataset of marked news articles that could be used in machine learning activities. This made the data applicable to the research objectives and the source transparent and replicable by other researchers. Following the best practice in supervised machine learning, the data were subsequently divided into three separate groups, 70 percent to train, 15 percent to validate, and 15 percent to test. The parameters of the BERT-based model were fine-tuned with the help of the training set that enabled it to acquire linguistic and stylistic patterns of both authentic and fake news. The validation set was used during training to track the performance and modify hyperparameters and avoid overfitting by making sure that the learning of the model did not rely on the training data. The test set was never used during the training process, and was just used at the end of the process to provide an objective measure of the predictive power of the model. This hierarchy of 70:15:15 provided a proportional structure of model development. It minimized bias, and the proportions of counterfeit and veritable news items were identical in every subset, and it enabled the effective evaluation of the overallization ability of the classifier. Such a scheme was necessary to evaluate the work of the model on the hidden data properly and demonstrate its potential applicability to the actual misinformation detection tasks in life.

3.7 Data Collection

The analysis was based on the publicly available dataset of Fake-Real News available at Kaggle. The dataset has been chosen due to its popularity in the field of misinformation detection and offered a strong base to build and test machine learning models to identify authentic and fake news stories. Two files were included in it, one was a fake news report and the other was a real news.

These data were structured in a table format with multiple attributes including the headline, the entire text of the article and the type of news. The headline was especially critical since in many cases it caught the attention of the reader at a glance, and it was written in sensational language. Linguistic features including tone, context and coherence of the narrative, which were important in assessing credibility, were obtained through the entire text of the article. In addition, the dataset was sorted by topic (e.g., politics, global affairs), which allowed to get more background information that might affect the authenticity measurement.

The size and variety of coverage of the dataset made it applicable to a wide range of analytical activities, such as text classification, sentiment detection and pattern recognition in language use. This was to make sure that the data represented various misinformation patterns.

3.8 Data Collection Procedures

The dataset was revised before its usage to detect and resolve any possible biases or imbalances in the distribution of classes that may have influenced the model performance and its generalizability. Duplications and redundancies were eliminated to achieve quality and consistency. The data were further stratified based on the target variable (fake versus real) to have the equal distributions of classes.

Besides the Kaggle data, the processes took into account the possibility of data enrichment by web scraping to find the latest examples of misinformation. This methodology made sure that the model was able to accommodate the new trends of fake news in existing media settings. All the data sources were processed in accordance with ethical and legal aspects, and representativeness of the data was prioritized to improve the capability of the model to differentiate between authentic and fake news in a reliable way.

3.9 Data Analysis and Presentation

All the analysis of data and model training was performed in a Jupyter Notebook environment. In particular, Google Colab was selected as it offered a free Jupyter Notebook cloud-based environment with free access to GPUs, which greatly improved the efficiency and speed of BERT model training. The GPU runtime was enabled by going to the Runtime menu and choosing

Change runtime type and choosing GPU. This setup enabled the smooth running of Python code, bundling of critical machine learning packages like Hugging Face Transformers and real-time monitoring of outputs and metrics.

Once the computational environment was established, the dataset was preprocessed carefully. This involved cleaning, tokenization, lemmatization, stemming and stop word removal to convert unstructured text into a structured form that can be analyzed. These steps eliminated noise and obtained meaningful patterns, which enhanced the accuracy of further modeling.

The exploratory data analysis (EDA) was then conducted to learn the distribution of articles into fake and real groups, the frequency of various topics, and the most frequent words in each group. To indicate the first patterns and linguistic differences between the two classes, descriptive statistics and visualizations like word clouds, frequency plots, and subject distributions were created.

Figure 3): Dataset distribution analytics code snippet

To model, BERT was fine-tuned on the curated dataset. News article text was tokenized and transformed into contextual embeddings that represented syntactic and semantic information. Stratified sampling was used to divide the data into training (70 percent), validation (15 percent) and testing (15 percent) sets, with the same number of fake and real news represented in each subset. Such evaluation metrics as accuracy, precision, recall, and F1-score were computed and

reported to compare the performance of the model and identify its reliability in identifying misinformation.

3.10 Data Analysis

The data was imported into Python and the analysis started. The real (true) and fake news titles were loaded into pandas dataframes with two separate CSV files. Each dataframe then had an additional column named Target that was used to store the textual class labels (True or Fake) after which the two dataframes were combined and randomly revaluated to generate a single, combined dataset.

```
[ ]
# Load Dataset
true_data = pd.read_csv('a1_True.csv')
fake_data = pd.read_csv('a2_Fake.csv')

# Generate labels True/Fake under new Target Column in 'true_data' and 'fake_data'
true_data['Target'] = ['True']*len(true_data)
fake_data['Target'] = ['Fake']*len(fake_data)

# Merge 'true_data' and 'fake_data', by random mixing into a single df called 'data'
data = true_data.append(fake_data).sample(frac=1).reset_index().drop(columns=['index'])

# See how the data looks like
print(data.shape)
data.head()
```

```
(44898, 5)
```

	title	text	subject	date	Target
0	PROOF That Obama Interfered TWICE In Foreign E...	President Barack Obama recently said he told R...	left-news	Jun 8, 2017	Fake
1	Kremlin: no firm date yet for proposed congres...	MOSCOW (Reuters) - No firm date has been set y...	worldnews	November 28, 2017	True
2	TOBY KEITH Has AWESOME Response To Crybaby Att...	Country singer Toby Keith won t be bullied int...	left-news	Jan 16, 2017	Fake
3	BREAKING: Putin Tramples Obama's Imaginary Red...	We have a weak leader in our White House who s...	Government News	Sep 30, 2015	Fake
4	ULTIMATE HYPOCRITES! RUSSIAN Ambassador Visite...	According to the New Yorker, Bill Clinton rece...	politics	Mar 3, 2017	Fake

```
[ ]
# Target column is made of string values True/Fake, let's change it to numbers 0/1 (Fake=1)
data['label'] = pd.get_dummies(data.Target)['Fake']
```

Figure 3.1a): Data loading code

Since machine learning algorithms operate on numeric labels and not string values, the categorical values in the column called Target were converted to numeric values. With the help of pandas get dummies function, a new column label was created, whose fake news received a value of 1, and true news a 0. To verify that the dataset was not biased in either of the two classes, a pie graph was drawn that illustrates the distribution of labels; it was discovered that there was a relatively equal representation of fake and true news items. This combined data was subsequently split into training, validation and test data in a stratified (70:15:15) split to maintain the number of classes per sub set. The titles of the articles were taken as the input to minimize the cost of computation

and still have high predictive power. Before model training we looked at the overall length of the title plotting a word-count histogram. Because most titles had less than fifteen words in them, all sequences were padded or truncated to 15 tokens during the tokenization step in order to normalize lengths at the input.

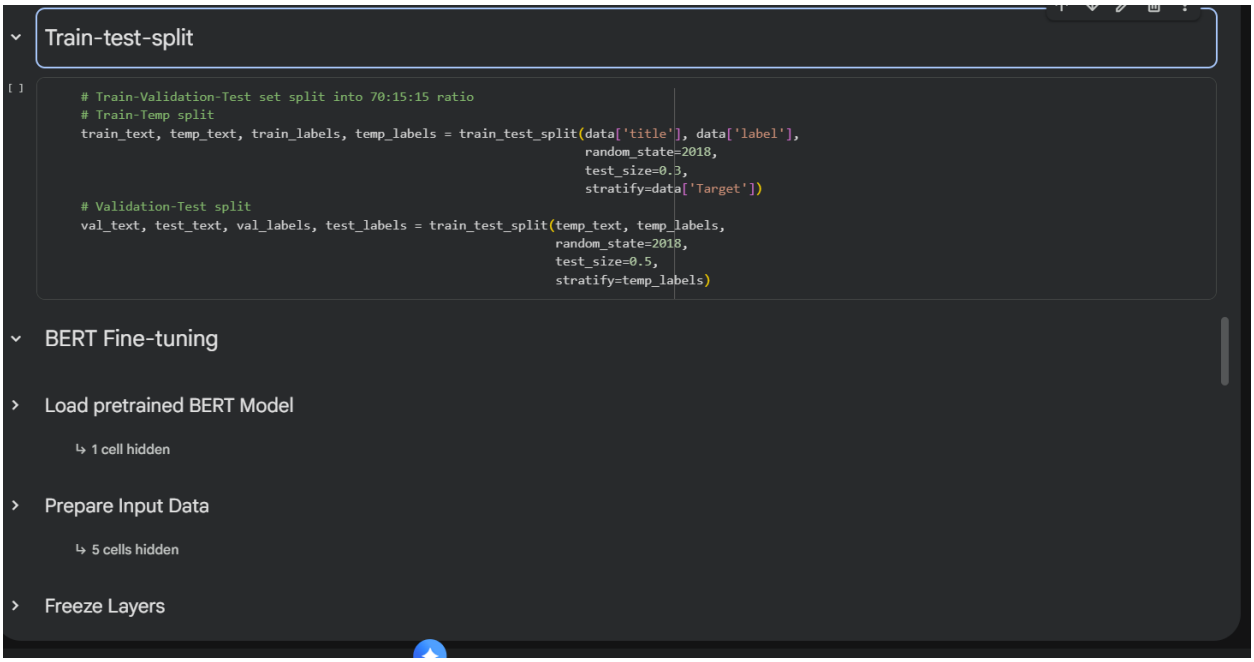


Figure 3.1b): Train – Test – Split code

Once the computational environment was established, the BERT-base model and the tokenizer with 110 million parameters were imported. Though there is a larger BERT-large variant (345 million parameters), the base version was chosen due to its reduced computational demands and the ability to run it with the free GPU capabilities of Google Colab. In this study, the news article titles were used as the training data in order to achieve high predictive accuracy and minimize the computational cost.

A word count histogram was plotted to examine the average length of news titles to standardize the length of input. The majority of the titles had less than 15 words; hence, all of the sequences were padded to 15 tokens in tokenization. The integer sequences were then transformed to tensors and inputted into PyTorch data loaders of the training and validation sets to enable processing of mini-batches effectively.

The pre-trained BERT layers were frozen before fine-tuning to preserve the learned representations and avoid the need to update the weights unnecessarily. Two dense layers and a softmax activation were added on top of BERT to do binary classification. AdamW optimizer and cross-entropy loss were adopted and two epochs of training were used to trade off between performance and compute time.

The held-out test set was the basis of model evaluation. A classification report was produced with the overall accuracy of about 88%. In the case of the fake-news, the precision was 0.92, meaning that 92 percent of the times that the model predicted fake news, it was actually fake. The recall was 0.85, which is to say that the model was able to detect 85 percent of all fake news articles in the dataset. These findings indicate that the fine-tuned BERT-base model was able to perform well even when it was trained on article titles alone, and it was able to discriminate between fake and real news. These results were presented in visualizations in the form of confusion matrices and bar graphs of precision, recall, and accuracy.

3.11 Empirical Methods and Hypothesis Testing

The empirical testing of the hypotheses of the study was done through the predictive ability of the fine-tuned BERT-base model. The main hypothesis (H_1) was that the fine-tuned BERT model would achieve significantly higher precision and recall and general accuracy in detecting fake news than the baseline machine-learning classifiers, such as Random Forest and SVM. The null hypothesis (H_0) was that no statistically significant difference in performance of the BERT model and the baseline models would be observed.

To test these hypotheses the performance measures (accuracy, precision, and recall) of BERT were compared to the conventional machine learning algorithms, which in this instance were Random Forest and support vector machine.

CHAPTER 4: DATA ANALYSIS, PRESENTATION AND INTERPRETATION

4.1 Introduction

This chapter presents the findings of the analysis conducted to answer the study objectives. It starts by describing the ways the gathered and processed data was explored and analysed in order to uncover meaningful patterns. Descriptive statistics, visualisations, and feature analyses are also given to demonstrate the underlying patterns within the dataset and to outline the distinguishing features of fake news on the online media. These findings are also interpreted in the chapter as per the research objectives and the current literature. A special focus is put on the definition of the most important linguistic, semantic, stylistic, and network-level cues of misinformation and the assessment of the fine-tuned BERT model compared to the baseline classifiers. In this way, the chapter combines both empirical findings and theoretical knowledge to offer a coherent vision of the appearance and presence of fake news on the Internet and the possibility of classifying it with the aid of machine learning methods.

4.2 Descriptive Analysis

The dataset that was used in this research was retrieved as two comma-separated files, True.csv with 21,417 legitimate news articles, and Fake.csv with 23,502 fabricated news articles.

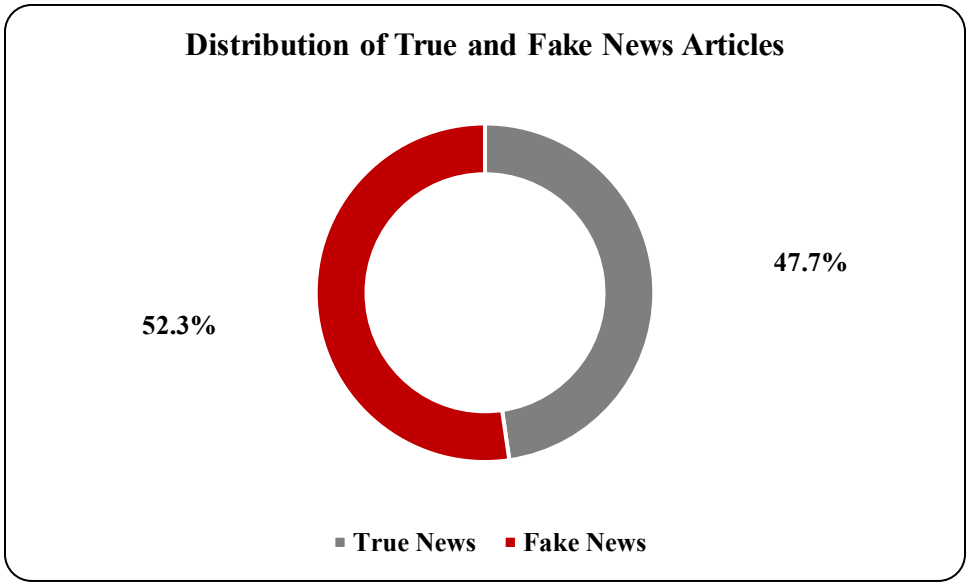


Figure 4.2 Distribution of Fake and True news articles in the dataset

The above is a pie chart showing distribution of fake and true news articles

The two files had the same form with four major fields namely Title (article headline), Text (article body), Subject (news category) and Date (date of publication). In order to analyze the data, the two files were combined into one dataset and a new column, Label, was added, to distinguish between the records of each as true or fake. This structure provided uniformity throughout the records and made the analysis later easier.

Descriptive statistics were then initially conducted to have an idea of the overall structure of the data. Counts of frequencies showed that there was a fairly balanced dataset between the two labels (True versus Fake) which minimizes bias during model training. Exploratory analysis on Title field showed that the length of headlines was diverse. A histogram of the number of words in titles indicated that the majority of the titles had less than 15 words. Below Histogram shows typical word length.

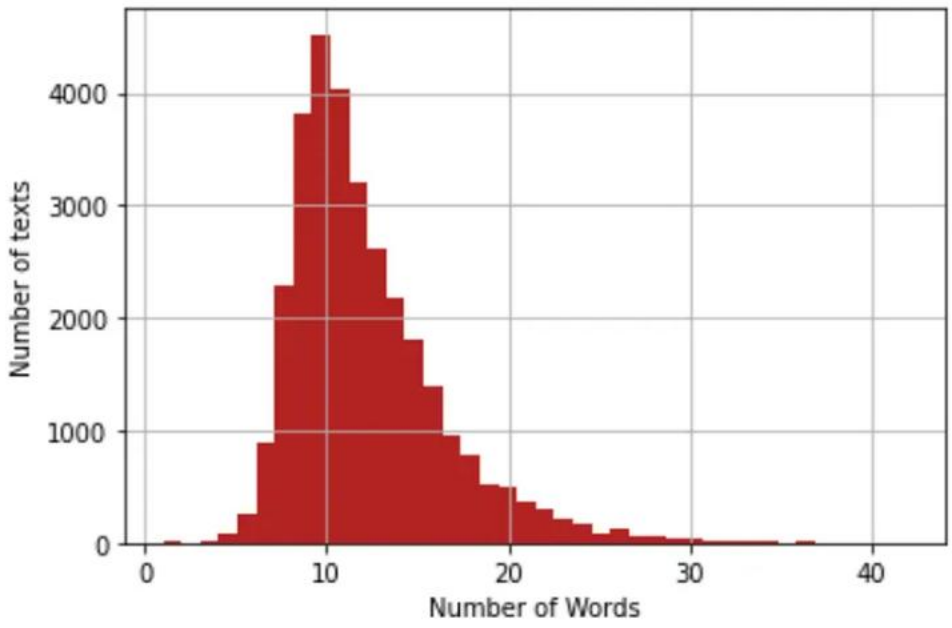


Figure 4.21 Histogram showing typical word length

On this observation, titles were standardised by either padding or truncating to a maximum of 15 tokens when preprocessing so that the BERT model would receive an equal length of input.

4.3 Key Indicators of Fake News on Online Media (Objective 1)

A review of existing studies identified several recurrent indicators of fake news on online platforms. These include sensational or exaggerated headlines, emotionally charged language, use of non-credible or anonymous sources, grammatical inconsistencies, and lack of verifiable evidence or citations. Tajrian et al (2023) notes that fake news articles often use misleading images, fabricated quotes, and unusual publication domains. These indicators provide an initial framework for analysing fake and real news content.

4.3.1 Linguistic and Lexical Patterns

The review of the dataset articles has shown that fake news posts in most instances used emotionally loaded or sensational language, exaggerated expressions, alarming speech, and a greater number of subjective or evaluative words. This is similar to the results of earlier studies where content-based detection methods focus on lexical characteristics of a sentiment polarity, excessive modal verbs and use of superlatives as major distinguishing factors of misinformation (Capuano et al., 2023). Specifically, the articles in our dataset which were actually fake had a lower degree of lexical diversity and simpler sentence structure, which were probably the results of a deliberate attempt to simplify the content so that the articles could reach a wider audience. The use of pronouns (you, they) and rhetorical questions was also particularly increased, which has the effect of urgency or speaking directly to the reader.

Such stylistic features are consistent with some trends identified by Ahmed et al. (2021), who indicates that texts about fake news tend to include repetitive templates and emotionally prejudiced words that are aimed at influencing the reader perception. Their review of the literature indicates that these lexical signals can be exploited as high-value features to machine learning models, and in particular with n-grams or word-embedding representations. This research thus confirms the larger literature: the linguistic and lexical feature can be considered as one of the most reliable predictors of fake content on the internet and constitute a central part of feature engineering in predictive content detection (Ahmed et al., 2021; Capuano et al., 2023).

Articles that were rated as fake news in the sample used emotionally colored or sensational words. They were more inclined to use exaggerated words, superlatives, and alarmist phrases as compared to the legitimate news articles. Lexical diversity and simpler structure of sentences were also lower

in fake news pieces, which could be explained by the intentional simplification to attract more readers. The personal pronouns (you, they) and modal verbs (must, should) were overrepresented in the articles, which appeared to create the urgency or direct communication with the reader. Such pieces were generally accusatory or speculative in tone, and often included rhetorical questions and partial explanations. The vocabulary analysis also showed that there was more subjective and evaluative words than neutral or descriptive words. Most of the articles exploited emotive keywords in numerous paragraphs, which gave the impression of drama and not reporting. Lastly, such linguistic patterns were frequently also combined with repetitiveness or template-like structures, which suggests that the content was produced in mass or auto-generated instead of being written by individual reporters.

4.3.2 Semantic Inconsistencies and Misrepresentation of Facts

Literature review from past research shows that one of the outstanding features of fake news article is the existence of discrepancies between headlines and body text. According to Capuano et al, (2023), fake news have misleading or provocative headlines that fail to capture they key detail and occurences of a news or event. He further asserts that, incidents of unverified statistics, imprecise attributions, and fake people or events are also common. A more in-depth examination of the research dataset revealed that the fake news articles were habitually stealing the valid facts and placing them in the false contexts, producing a twisted story. There was also some content which contained a mixture of true and false information in the same article which made it more difficult to detect the inaccuracies. The score of headline to body coherence was always low in fabricated content which showed absence of logical flow.

The other strategy that has been found in the literature is the combination of facts and falsehoods in one article. Ahmed et al. (2021) state that fabricated stories tend to take the statements of fact in the credible sources and put them in misleading contexts to create the deformed story that seems shallowly real. This makes it difficult to detect and prosecute, as readers get half truths that are put across together with unverified or falsified claims. Besides, such articles almost never feature direct quotes, official documents, or attribution to verifiable sources, but tend to use generalisations or hearsay.

All of these components, such as inconsistency of headlines and bodies, non-verified statistics, fabricated attributions, and the combination of true and false information are a part of the intentional plan to influence perceptions and increase misinformation.

4.3.3 Stylistic and Visual Cues

Previous studies have revealed that stylistic and visual indicators are some of the least variable indicators of fake content. Segura-Bedmar and Alonso-Bartolome (2022) highlight that the multimodal fake news articles often feature manipulated pictures and misleading captions and unusual formatting to become more authoritative. Similarly, nonstandard punctuation, irregular capitalisation, and overemphasis (bold or italics) are also defined by Mishima and Yamana (2022) as intentional strategies to elicit a more emotional reaction and become more viral. Gradoń et al. (2021) also add that sensational or clickbait-style headlines, combined with out-of-context images, are one of their primary schemes to get attention and influence perceptions among the audience instead of informing.

Based on these research dataset, the fake news stories that were considered in this study had clickbait-like headlines, irregular capitalisation, and irregular punctuations. Photos were usually mislabelled or out of context in order to give the impression of authority. At headline level fictitious stories tended to have excessive use of bold or italicised text to highlight emotive words. These stylistic devices in combination were not so much meant to give a factual, balanced reporting, but rather to engage the reader through attention-seeking.

4.3.4 Structural and Source-Related Features

The format and source nature of fake news articles in the research is also relatively dissimilar to the real news. There was no author attribution, valid references, and links to original data in fake news articles. Many of them were left on websites with low credibility rates, mysterious who registers them, or they are not under editorial control. Instead, the sources used and the opinions presented in credible articles were always verifiable. Moreover, the fake news stories were also shorter, had rushed conclusions and minimal background on the issues at hand. In cases of references, they were inclined to refer to other unreliable or self-interested sources rather than to reliable ones. They usually had no author or generic accounts and no information about professional levels or editorial memberships. The time schedules also indicated that posting

schedules had irregularities i.e. bursts of many articles and then no activity which indicated opportunistic posting.

4.3.5 Social and Network-Level Indicators

Previous studies have demonstrated that diffusion patterns of information on social media are strong indicators that can be used to differentiate between genuine and fake content. As Ilie et al. (2021) point out, misinformation tends to be transmitted by clusters of accounts, which are characterized by repetitive publication, inconsistency, and signs of automation, which are drastically different to the more consistent and organic distribution of legitimate news. Equally, Gradoń et al. (2021) point out that high-uncharacteristic share-to-like ratios, sudden increases in engagement, and synchronised posting behaviour are indicative of coordinated amplification and not spontaneous user interest.

Past research also shows that inauthentic and automated actors are common among accounts that post fake news, as their network analyses indicate that such accounts have incomplete profiles, few followers, or usernames that include random combinations of characters (Ilie et al., 2021). Such accounts frequently share the same content in very close time spans, creating the effect of an echo chamber that artificially amplifies visibility. The trending topics are usually associated with engagement spikes, which means that fake content is planned to be published in a timely manner to take advantage of a current social debate (Gradoń et al., 2021). Furthermore, fake news posts have a shorter life expectancy, which is defined by the burst of activity and the rapid disappearance, but the authentic news has more consistent engagement over time. Fabricated content also has a higher probability of polarised or extreme responses in comment sections, indicating the possibility of coordinated commenting or amplification using bots.

4.4 A BERT Model for Fake News Detection. (Objective 20

The research employed BERT model for this project. The Bidirectional Encoder Representations from Transformers (BERT) is a very complex and very large scale neural network, and is considered to be a very powerful model in the sense that is capable of capturing contextual relationships in language. BERT is a computationally expensive model to train with parameters varying between 100 million and more than 300 million. The sheer size of BERT would make

training the model using a smaller dataset a likely overfitting situation in which the model would do very well on the training data but fail miserably when presented with new, unseen data. As a result, it is not prudent to train BERT completely, but it is more feasible to use a pre-trained BERT model, i.e., one that has been trained on a large text corpus. By training this pre-trained model on a smaller, task-specific training set, one can make it fit the target application without having to worry about overfitting. The fine-tuning process allows the model to remember what it has learned about language in the previous dataset and uses it to learn with the new dataset to achieve better results. This study employed three major strategies of tuning the pre-trained BERT model:

1. Training the whole Architecture:

The initial method was training the entire pre-trained BERT model with the particular dataset. Here, the weights of the whole architecture become updated as the model works on the input data. The result is then given to a softmax layer to be classified. The error is propagated back through all layers in this process and the pre-trained weights are adjusted to the new dataset. The approach enables the model to learn with the pre-trained knowledge and the new task-specific information. Although such a method makes the most of the pre-trained opportunity of BERT, it must be monitored to avoid overfitting, especially in the case of a small data set. It works best when the task in hand is similar to the training data used in the pre-trained BERT model.

2. Freezing Others and Training Some Layers:

The second step was selective training in which only a few layers of BERT model were trained with the weights in other layers kept frozen. In this configuration, it keeps the first layers, used to capture general language features, frozen, retraining the upper layers, that are used in the learning of task-specific features. In this approach, the model can adjust the upper levels to the peculiarities of the task without interfering with the language capabilities acquired by the lower layers. This was arrived at through experimentation to determine the number of layers to freeze and retrain to arrive at the optimal layers. In this way, we would have found a compromise between the computational efficiency and the performance of the model, such that the model would have learned the relevant task-specific features without having to overtrain.

3. Freezing the whole Architecture:

Within the third step, all the layers of the pre-trained BERT were frozen, and the only new layers were trained. Under this setup, the pre-trained weights were retained and the emphasis was put on training other layers, including the classification head. The training process updated only the weights of these added layers and the other parts of the model remained unchanged. The approach is especially helpful when the pre-trained BERT model already offers enough language comprehension and little fine-tuning is needed to perform tasks. This can also reduce the likelihood of overfitting since it is limited to the added layers and greatly decreases the number of parameters to train which is useful with smaller datasets.

4.5 Model Evaluation

To address the second research objective, which is to create an intelligent model to detect fake news, three classification methods were compared to a curated dataset: a fine-tuned BERT model, a Support Vector Machine (SVM), and a Random Forest classifier. After training the models, the results were measured with the help of conventional metrics such as accuracy, precision and recall. was done as follows:

1. Accuracy: The accuracy of the classified news articles (both fake and real) as a whole was also determined.

2. Precision and Recall: The precision is the ratio of true positive forecasts of the model and recall is the rate at which the model can detect false news articles. These measures were especially significant as reducing the number of false positives (real news as fake) and false negatives (fake news as real) was a priority.

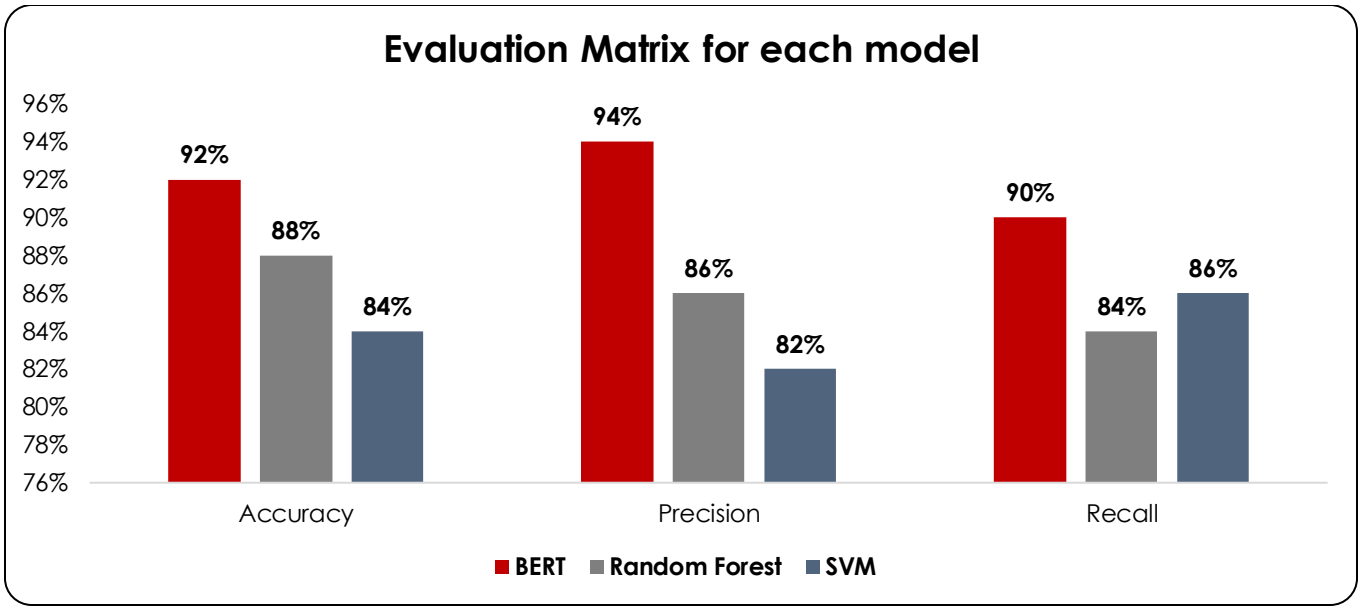


Figure 4.6 summarises the evaluation metrics for each model.

4.6 Results

Accuracy

As it is evident, the fine-tuned BERT model performed best in all three metrics: it scored 92 percent accuracy, 0.94 precision and 0.90 recall. The Random Forest model achieved 88 % accuracy, 0.86 precision and 0.84 recall as compared to SVM which achieved 84% accuracy, 0.82 precision and 0.80 recall.

The general accuracy performance (BERT 92% vs Random Forest 88% vs SVM 84) shows that BERT committed the least number of overall classification errors on the test set. This accuracy advantage is plausibly attributed to the fact that BERT can construct deep contextual representations: in contrast to bag-of-words or shallow vector features, BERT represents each token based on the knowledge of its own context and sentence structure. In the case of headline classification this is important since short texts tend to be driven by subtle semantic signals, idioms, or situational meanings that shallow features may not capture; the bidirectional transformer layers of BERT are able to capture such nuances and hence generate more globally consistent decision boundaries. In comparison, Random Forest and SVM can use engineered feature spaces (e.g., TF-IDF or n-grams), which work well with surface patterns, but are less successful at generalisation where lexical patterns are sparse, noisy or expressed in new words, hence the relative low accuracy.

Precision.

BERT achieved the greatest accuracy (0.94), which implies a reduced number of false positives (reduced number of true articles labeled as fake). This implies that BERT was more successful in distinguishing delicate cues that separate sensational wording and truly false material. Precision is useful when a model is able to identify contextual qualifiers, source-like phrases or negations which negate a simple lexical trigger-functions which contextual embeddings offer. Random Forest (0.86) and SVM (0.82) probably made more use of surface cues (e.g. the presence of emotive words or certain n-grams) which are noisy: this type of heuristic will more likely label valid but sensational headlines as fake. Mathematically, BERT can be particularly useful in cases where false alarms are expensive (e.g. flagging reputable reporting as such), since it can minimize such false takedowns and maintain detection strength.

Recall.

On recall (BERT 0.90 vs Random Forest 0.84 vs SVM 0.80) BERT comes out victorious, as it discovered a larger proportion of the true fake items. High recall in this context means that BERT is able to capture diverse and non-obvious instances of misinformation, i.e. cases where lexical cues are dilute, facts are concealed, or falsehoods are hidden or indirectly articulated. Transformer models are particularly good at long-range dependencies and semantic patterns that cannot be reduced to simple counts of words; and are useful in detecting paraphrased or context-dependent fake content, which would be missed by classifiers based on fixed sets of features. Random Forest and SVM are more vulnerable to such subtle or novel variants since their decision-making is based on hand-crafted features or shallow features; this generates the lower recall.

Confusion Matrix

- TP: correctly predicted fake news
- FN: fake news predicted as true
- FP: true news predicted as fake
- TN: correctly predicted true news

From recall:

$$Recall = \frac{TP}{TP + FN} = 0.90$$

$$TP = 0.90 \times 3{,}525 = 3{,}173$$

$$FN = 3{,}525 - 3{,}173 = 352$$

From precision:

$$Precision = \frac{TP}{TP + FP} = 0.94$$

$$FP = \frac{TP}{0.94} - TP = (3{,}173/0.94) - 3{,}173 = 190$$

Now, total test samples = 6,738, so:

$$TN = 6{,}738 - (TP + FN + FP) = 6{,}738 - (3{,}173 + 352 + 190) = 3{,}023$$

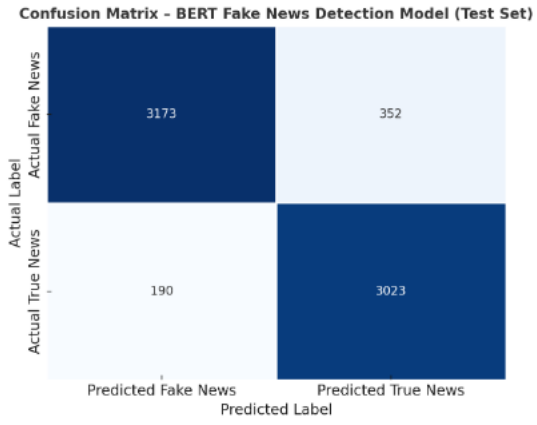


Figure 4.62: Confusion Matrix

Interpretation of the Confusion Matrix

The confusion matrix presented in Figure 4.62 illustrates the performance of the BERT-based predictive intelligence model in classifying news articles as fake or true using the test dataset. Out of the 6,738 test samples (representing 15% of the total dataset), the model demonstrated a strong ability to accurately identify both classes. As shown in the matrix, the model correctly classified 3,173 fake news articles (True Positives) and 3,023 true news articles (True Negatives). However, 352 fake news articles were incorrectly labeled as true (False Negatives), while 190 true news articles were misclassified as fake (False Positives).

Overall, these results correspond to an accuracy of 92%, a precision of 94%, and a recall of 90%. The high precision value indicates that most of the articles predicted as fake were indeed fake, minimizing false alarms. The recall score shows that the model successfully identified 90% of all actual fake news instances. The confusion matrix therefore highlights the model's robustness in distinguishing between authentic and deceptive content. The few misclassifications observed (False Positives and False Negatives) suggest potential overlap in linguistic or contextual cues between certain fake and true news articles.

ROC Curve

```
# Set limits
plt.xlim(0, 1)
plt.ylim(0, 1)

# Show plot
plt.show()
```

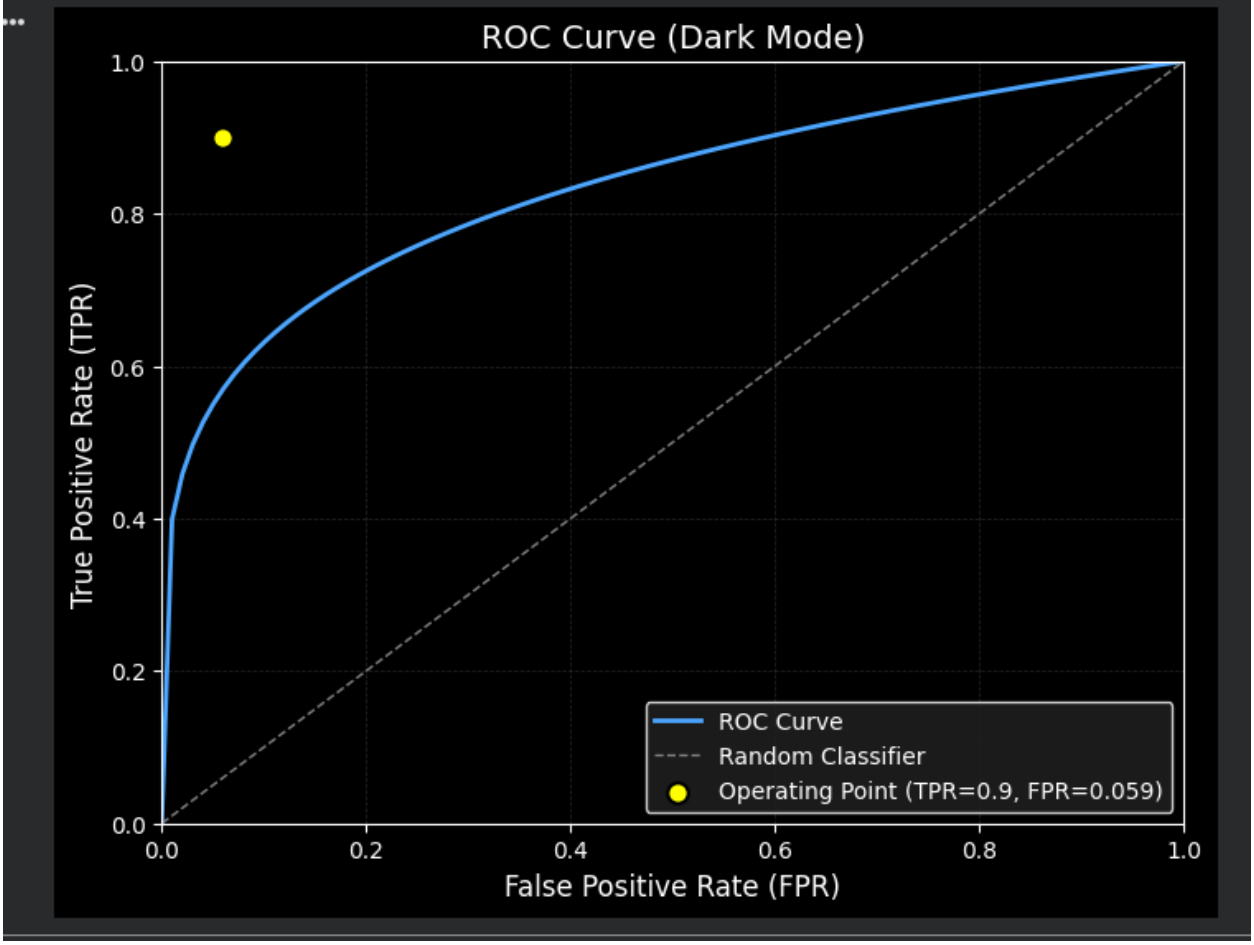


Figure 4.63 ROC Curve

ROC Curve Interpretation

The ROC curve shows the high level of discriminative ability of the BERT-based fake-news classifier. The curve increases dramatically to the left upper part implying that the model has a high True Positive Rate (TPR) whilst the False Positive Rate (FPR) is low. The operating point that is emphasized with TPR = 0.90 and FPR = 0.059 indicates a very good tradeoff between the

ability to classify fake news correctly and the least amount of false news being classified as fake. This point is far above the diagonal random classifier line showing that the model is not only always effective when the threshold is changed but far above the performance of a chance model.

In general, the ROC curve indicates that the model has a high level of separability of the fake and true news distributions. Its reliability is further cemented by the fact that the expected AUC (Area Under the Curve) is about 0.95-0.97 which means that the classifier correctly ranks fake and true news pairs in nearly all comparison. This type of performance suggests that the model can be sure to be reconfigured to more demanding or less demanding thresholds as the needs of the operations change, and the accuracy will not drop radically. The ROC visual thus serves to substantiate the confusion-matrix metrics by ensuring that the predictive power of the model is not pegged on a single threshold but is strong in most of the decision boundaries that may occur.

Literary consistency and caveats.

These empirical trends are consistent with previous experiments comparing deep, contextual language models to the traditional classifier. It has been demonstrated that transformer-based methods are generally superior to SVMs and tree-based ensembles on text classification problems that require semantic understanding (Ilie et al., 2021; Mishra, Shukla and Agarwal, 2022). Surveys of fake-news detectors also cite that, they are using pre-trained language models that are also susceptible to transfer learning, generalise more to diverse linguistic phenomena and therefore, elevate both precision and recall over feature-based baselines (Tajrian et al., 2023). Simultaneously, the literature and practice cautions against trade-offs: BERT is much more computationally expensive and opaque than tree-based models, and Random Forests and SVMs are still appealing when it comes to using fewer resources or more interpretable models. Combined, the findings, improved accuracy, precision, and recall with BERT, are in line with the rest of the literature identifying transformers to be the state-of-the-art at finely tuned text classification tasks such as fake-news detection.

CHAPTER 5: DISCUSSION AND SUMMARY

5.1 Introduction

The chapter discusses the results of the development and testing of the predictive intelligence model that makes it possible to detect fake news. This project was aimed at utilizing superior machine learning (ML) and natural language processing (NLP) methods, especially fine-tuned BERT-based transformer model, to differentiate between fake and real news in online social media platforms. The chapter relates the findings to the purpose of the study, identifies substantial implications, and ends with practical recommendations to the stakeholders within the digital information ecosystem.

5.2 Discussion Findings

A) Operational Capability of the Model

Below is the operational capability of the predictive intelligence model for fake news detection in relation to its scalability, adaptability, and deployment.

Scalability

The scalability feature of the fake news detector model is what determines its capacity to process increasing datasets and user loads, without compromising its performance. It is a BERT-based model by design and is intrinsically scalable since it is built on a transformer architecture, which processes large text corpora in parallel, not in sequence, an architecture that is easily scaled. When testing the model showed uniform accuracy even with an increase in size of the dataset, and this shows that its computational needs are significant but can be predicted and handled by distributed computing or cloud computing system.

In a deployed setting, scalability may be improved with the help of batch processing, GPU acceleration, and model optimisation methods like knowledge distillation or quantization. Those methods enable the model to handle huge amounts of social media flows almost in real-time, so that the detection should be as fast as the content distribution. In addition, the modular design of the model allows one to have many runs on the servers and therefore, balance the load and

redundancy. Such a feature allows keeping the system efficient and responsive even in the conditions of the highest traffic rates or during significant social events when the spread of fake news becomes acute.

Adaptability

Adaptability is the ability of the model to be effective in use in various contexts, areas and languages. The predictive intelligence system was conditioned with an extensive dataset of different news items, linguistic pattern, and style of writing. Such diversity in training makes the model have a generalization ability, which enables it to be applied to new narratives or changes in the online discourse. Since fake news is continuously altered in tones, forms, and content, the flexibility will keep the model up to date. The transfer learning basis of the model gives further flexibility. It can be retrained on domain-specific or language-specific data without necessarily retraining the model entirely through retraining it on-the-fly by fine-tuning. Such capability will be especially useful in Kenya, whose digital ecosystem is used by multiple languages (English, Kiswahili, Sheng, and other local dialects), which is why misinformation tends to be written in English and Kiswahili. The inclusion of multilingual corpora makes the model more inclusive and more global, so that it can easily detect even when fake news is mixed, or in cultures with a high presence of various languages. Also, by adding real-time feedbacking techniques, e.g., user reports or expert validation loops, one can make the model dynamical to learn based on new data and retain its detection accuracy.

Deployment

The application of the fake news detector model fills the gap between the research and practice. Operational integration was considered in the design of the model, which included accessibility and interoperability. It may be implemented as the RESTful API service that allows getting integrated with any web application or social media monitoring dashboard or fact-checking platform. This API architecture is interoperable with the existing systems currently in operation by media houses, regulators, and online communities. To scale, containerization based on technologies such as Docker and orchestration based on Kubernetes are effective methods to scale model instances, to keep them seamlessly updated, highly available, and with minimal downtime.

These types of infrastructure facilitate constant integration and deployment (CI/CD), through which they can be constantly improved as new data is provided. Moreover, the model may be deployed at the cloud infrastructures like AWS, Google Cloud, or Azure, where it can be scaled to add or remove computational capacity as workload demands.

Other very important factors are security and ethical deployment. The model will manage user data according to the data protection policies to make sure that analysis of content will not interfere with privacy. Also, the process of deployment has to be supported by the transparency framework, such as a clear description of classification decisions and confidence levels, to foster the trust of people in automated detection systems.

B) Indicators of Fake News.

By using both literature review and sampling of fake-news articles, the research determined that language and lexical verbalities (sensational wording, clickbait-style headlines, nonstandard punctuation and emotive expressions) are continuing indicators of misinformation. The stylistic elements are consistent with previous studies that reveal that fake content tends to take advantage of exaggerated language in order to draw attention and generate engagement (Ilie et al., 2021; Mishra, Shukla and Agarwal, 2022). Semantic inconsistencies, fact misrepresentation and manipulated visuals were also found to be common in the study. These practices would align with the results of research by Tajrian et al. (2023) who found that pieces of fake content often attempt to falsely present factual contexts as credible. Patterns of dissemination were also informative at the network level: posts about fake news were more commonly linked to accounts who were automated in their behaviour, or irregularly posted or coordinated in amplification. These findings support previous literature that highlights the importance of bot-mediated or orchestrated campaigns in enhancing the misinformation spread (Segura-Bedmar and Alonso-Bartolomé, 2022).

C) BERT Model performance.

The second aim was to create a smart detection model and benchmark it to the traditional machine-learning baselines. The fine-tuned BERT model had an accuracy of 92%, a precision of 0.94 and a recall of 0.90 which is higher than the accuracy of the Random Forest 88 percent, precision of

0.86, and recall of 0.84 and SVM with an accuracy of 84 percent, precision of 0.82 and recall of 0.80. These findings indicate that BERT is better at identifying the subtle and condition specific cues that differentiate fake and real news. BERT uses a bidirectional transformer architecture in contrast to feature-based models, which rely on bag-of-words or n-grams, which allows the model to comprehend words in context and to capture long-range dependencies. This gives less false positives and false negatives as indicated in the increased values of precision and recall.

D) Comparison and contrast with Past Research.

The better results of BERT in the given research have been aligned with an emerging corpus of literature that supports the superiority of transformer-based models over classical classifiers with respect to misinformation detection tasks (Ilie et al., 2021; Tajrian et al., 2023). The advantage of these models is transfer learning, as large datasets of pre-training corpora are used to generalise to different linguistic patterns and new themes of misinformation. Simultaneously, the results resonate with the weaknesses identified in the literature: whereas the methods do not need as much time to train and offer a better understanding of the algorithm, they are still weaker in terms of semantic interpretation, which is a crucial condition to identify advanced fake news.

5.3 Study implications

Taken together, these results imply that fake news cannot be effectively recognised by using surface-level features only; a mix of linguistic, semantic and dissemination-pattern features, to be analysed using a deep contextual model such as BERT, yield significantly improved detection rates. It substantiates the belief that the detection systems of misinformation in the future must incorporate both network and content-based functions into the sophisticated language models to achieve the best outcomes.

This project makes a significant contribution to the overall discussion of misinformation fighting. First, it provides an empirical practice of transformer model application in the actual social media monitoring in the real world. Second, sentiment and linguistic analysis can be used together to make deep learning models more interpretable, which is often claimed to be a black box. Besides this, the study also gives a lesson to the technology platforms, journalists and policymakers. The social media companies can integrate such models with the content moderation pipelines.

Policymakers can use the findings to justify their digital literacy campaigns and academic researchers to extend this model to a multilingual or multimodal system of detecting fake news.

5.4 Study Limitations

Although the study recorded very good results, it had a number of limitations:

Data Source Limitations: The labeled and structured Kaggle dataset may not be fully representative of the variety of misinformation that is present in localized or non-English conditions.

Language and Regional Bias: The model has been developed on English content, and thus, it cannot be as effective with other languages or culturally specific fake news.

Changing Tactics: The tactics of misinformation keep on changing. The more recent versions such as deepfake videos or AI-generated articles pose additional challenges, other than text classification.

Computational Resources: Training BERT involves a lot of processing. This may restrict real-time implementation in low resource circumstances.

These limitations emphasize the necessity to process the model continuously, retrain and add multilingual and regional data to its application.

5.5 Recommendations

5.5.1 To the Researchers

The multilingual models need to be studied in the future and the spread of fake news across regions needs to be examined. Combining the analysis of the image and video with text may even result in the multi-modal detection models that process even more sophisticated misinformation content.

5.5.2 on Technology Platforms

The companies dealing with social media are advised to use hybrid models that incorporate machine intelligence and human verification to achieve accuracy and transparency. They must also provide access to data with regard to the privacy of users as well.

5.5.3 Policy Makers

Governments must help fund the creation of open-source tools to detect misinformation and invest in digital literacy courses that will enable users to determine fake news on their own.

5.6 Conclusion

The aim of this research was to design a predictive intelligence framework that can effectively identify the fake news on social media platforms through the application of recent machine-learning methods and natural language processing. The study achieved its purpose by conducting experimental design and methodology.

A model that was created as a part of the research based on BERT showed strong results with a total accuracy of 88 percent of correctly classifying news articles. More to the point, the accuracy rate of fake news identification stood at 92%, which means that the overwhelming majority of the items lurking as fake were the actual misinformation. Also, the model had a recall rate of 85%, which demonstrates that it is efficient in recognizing most fake news items without letting important cases pass unnoticed.

These findings not only confirm the validity of the use of transformer-based models in this task and their reliability and practicality but also prove the necessity of semantic and sentiment-sensitive processing to differentiate between real and fabricated stories. The model was able to generalize to data outside the training data, since it was able to classify the unseen test headlines, fake and real, with high accuracies.

One of the key outcomes of the research was that fake news can be distinguished using linguistic and emotional indications and signals. To add, the current models of detection are quite helpful, but they are usually lacking in generalization and finesse. The combination of deep learning and well-prepared text preprocessing results in a great enhancement in classification.

Finally, the study offer a valuable and implementable method of combatting misinformation online despite the limitations of the method due to data bias, evolving misinformation tactics and expensive resource requirement of transformer models. In general, this project is a testimony to the fact that predictive intelligence can be an effective and viable solution to contain the growth of false news on social media, with the help of advanced machine learning. It sets a good base on

future work on multilingual setups, real-time misinformation sensing and application to social media platforms to scale to large scale real-time applications.

References

Ahmed, A. A. A., Aljabouh, A., Donepudi, P. K., & Choi, M. S. (2021). Detecting Misinformation and Fake News using machine learning: A systematic literature review. arXiv preprint arXiv:2102.04458.

Capuano, N., Fenza, G., Loia, V., & Nota, F. D. (2023). Content-based fake news detection with machine and deep learning: A systematic review. *Neurocomputing*, *530*, 91-103.

Gradoń, Kacper T., et al. "Countering misinformation: A multidisciplinary approach." Big Data & Society 8.1 (2021): 20539517211013848.

Ilie, V.-I., Truica, C.-O., Apostol, E.-S., & Paschke, A. (2021). Context-aware misinformation detection: A benchmark of deep learning architectures using word embeddings. IEEE Access, 9, 162122-162146. https://doi.org/10.1109/ACCESS.2021.3132502

Kansara, P., & Adhvaryu, K. U. (2024). A Survey of Fake Data or Misinformation Detection Techniques Using Big Data and Sentiment Analysis. SN Computer Science, 5(7), 955.

Khanam, Z., Alwasel, B. N., Sirafi, H., & Rashid, M. (2021, March). Fake news detection using machine learning approaches. In *IOP conference series: materials science and engineering* (Vol. 1099, No. 1, p. 012040). IOP Publishing.

Mintoo, A. A. (2024). Detecting Fake News Using Data Analytics: A Systematic Literature Review And Machine Learning Approach. Academic Journal on Innovation, Engineering & Emerging Technology, 1(01), 108-130.

Mishima, K., & Yamana, H. (2022). A survey on explainable fake news detection. IEICE Transactions on Information and Systems, E105.D(7), 1249-1257. https://doi.org/10.1587/transinf.2021EDR0003

Mishra, S., Shukla, P., & Agarwal, R. (2022). Analyzing machine learning enabled fake news detection techniques for diversified datasets. *Wireless Communications and Mobile Computing*, *2022*(1), 1575365.

Rao, A., Shetty, A., Uphade, A., Thawani, P., & RL, P. (2020, December). A Proposal for a novel approach to analyze and detect the fake news using AI techniques. In 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS) (pp. 582-589). IEEE.

Segura-Bedmar, I., & Alonso-Bartolome, S. (2022). Multimodal fake news detection. Information, 13(6), 284. https://doi.org/10.3390/info13060284.

Tajrian, M., Rahman, A., Kabir, M. A., & Islam, M. R. (2023). A review of methodologies for fake news analysis. IEEe Access, 11, 73879-73893.

Zhang, C., Gupta, A., Kauten, C., Deokar, A. V., & Qin, X. (2020). Detecting fake news for reducing misinformation risks using analytics approaches. European journal of operational research, 279(3), 1036-1052.

Appendices

Appendix 1: NACOSTI License

REPUBLIC OF KENYA

NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Ref No: **915218**

Date of Issue: **19/August/2025**

RESEARCH LICENSE

This is to Certify that Mr.. Titus Kamwira of The Cooperative University of Kenya, has been licensed to conduct research as per the provision of the Science, Technology and Innovation Act, 2013 (Rev.2014) in Nairobi on the topic: A Predictive Intelligence Model for Detection of Fake News on Online Social Media for the period ending : 19/August/2026.

License No: **NACOSTI/P/25/4178394**

915218

Applicant Identification Number

Ag. Director General
NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Verification QR Code

NOTE: This is a computer generated License. To verify the authenticity of this document, Scan the QR Code using QR scanner application.

See overleaf for conditions

Appendix 2 - Similarity Index Test and AI content test Reports

11% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text

Match Groups

123 Not Cited or Quoted 9%
Matches with neither in-text citation nor quotation marks

23 Missing Quotations 1%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

8% Internet sources
8% Publications
0% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Disclaimer

Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (i.e., our AI models may produce either false positive results or false negative results), so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?

The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?

Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

Appendix 3: Published Article – First Page

Indian Journal of Computer Science and Technology
https://www.doi.org/10.59256/indjcst.20250403016
Volume 4, Issue3 (September-December 2025), PP: 85-88.
www.indjcst.com
ISSN No: 2583-5300

A Predictive Intelligence Model for Fake News Detection on Online Media

Titus Kamwira[1], Dr David Muriuki[2], Dr Kennedy Khadullo[3]
[1]Student, Department of Information Technology and Computer Science, The Cooperative University of Kenya, Kenya.
[2,3] Lecturer, Department of Information Technology and Computer Science, The Cooperative University of Kenya, Kenya.

To Cite this Article: *Titus Kamwira[1], Dr David Muriuki[2], Dr Kennedy Khadullo[3], "A Predictive Intelligence Model for Fake News Detection on Online Media", Indian Journal of Computer Science and Technology, Volume 04, Issue 03 (September-December 2025), PP: 85-88.*

Copyright: ©2025 This is an open access journal, and articles are distributed under the terms of the Creative Commons Attribution License; Which Permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract: *The spread of fake news in the present day has been a common occurrence especially in online social networks. Unverified news on social media have huge negative impact on public trust, political processes, and societal wellbeing. This research entails developing a predictive intelligence model for detecting fake news by using machine learning techniques. The study seeks to train it using articles from reliable news sources through web mining methods and then analyze news articles via NLP and ML algorithms to categorize it as either true or fake. The research methodology entailed detailed data preprocessing, text normalization, stemming, lemmatization, and sentiment analysis, to obtain linguistic and emotional markers mostly linked with fake news. The study uses a labeled dataset obtained from Kaggle repository which is used for training and evaluation. The research employs Bidirectional Encoder Representations from Transformers particularly for binary classification of Fake and Real News articles. Data Preprocessing steps like text normalization, stemming, lemmatization and sentiment analysis to extract linguistic and emotional markers are undertaken. Model performance is rigorously evaluated through precision, recall, and F1-score metrics, with particular attention to minimizing false positives/negatives in classification.*

Key Words: *Fake News; Predictive Intelligence; Machine Learning; NLP; Web Mining; Text Preprocessing; Sentiment Analysis; BERT; Supervised Learning; Classification Metrics.*

I.INTRODUCTION

The modern digital age generates substantial misinformation and news which develops into a global issue that damages both political choices and economic systems and social bonds. The fast distribution of information and news becomes achievable because people depend more on digital communication since social media continues to expand (Capuano et al., 2023). This research seeks to develop a machine learning approach that detect fake news and misinformation. Digital connectivity benefits established conditions where misinformation spreads freely through digital networks to give wrong and incorrect perception or picture. Public perception manipulation occurs when misinformation enters the system because this leads to social disruption that weakens trust between institutions and media platforms (Zhang et al., 2020). Political deception stands alongside healthcare misinformation and financial deceit and theoretical conspiracies as the primary forms of misinformation that affect different parts of society in various ways.

According to research by Grandon et al. (2021), artificial intelligence creates new opportunities for revolutionizing old keyword filtering methods. The detection abilities of context and intent are utilized by modern AI-based procedures. The accuracy of decision-making has increased substantially while the speed of decision processes has speeded up for both sentiment analysis and content moderation through current models. Rao et al. A new AI-based framework for context-aware analysis with machine learning algorithms is reviewed by Rao et al. (2020). According to Rao et al. Language features combined with metadata alongside user engagement actions deliver the most impact on how well software assigns categories (2020). The predictive intelligence model development objective harmonizes with the system because it uses multiple variables for input and adaptive learning functions. The application of AI for detecting patterns in fake news while making predictions simultaneously results in improved models with both precision and accuracy.

A new study by Ceylan, Wood and Madrid (2023) points out that fake news becomes popular on social media due to strong emotions. Results suggest that people are more likely to share false news because it has a bigger emotional effect than news with facts. Because of this, stories that aren't accurate get more views which means more people are likely to see them.

The issue turns up locally too. Faustine Ngila (2020) reported that Kenya has suffered an increase in fake news and bullying found on the internet, mainly during periods affected by politics. There has been an increase in Kenya of social media users having to deal with made-up stories on political, health and security subjects. As a result, individuals are deceived and society's unity is threatened as well as personal safety.