

سلطان حسن

Fake News Detection - Final Draft.docx

- 1- Final Thesis (Word Format)
- Thesis_(2)_AI901_Dr_Wessam_Ahmed_Third Trimester 2025 (Moodle PP)
- Midocean University

Document Details

Submission ID

trn:oid::1:3450684003

Submission Date

Dec 25, 2025, 9:25 PM GMT+4

Download Date

Dec 25, 2025, 10:42 PM GMT+4

File Name

xcod.9347365002_101621_tfarD_laniF_-_noitceteD_sweN_ekaF_سلطان_حسن_28992

File Size

2.3 MB

58 Pages

12,805 Words

77,039 Characters

*% 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.





A Comparative Study of Fake News Detection Using Machine Learning, Deep Learning, and Large Language Models

Prepared By :

Name: Rayan Shukri ID No.: **251000636**

Name: Sultan Hassan ID No.: **251002896**

Submitted in fulfillment of the requirements for the degree Master of Science in Artificial
Intelligence At College of Informatics Midocean University , UAE

Supervisor Name: Wesam Ahmed

College of Informatics, Midocean University

Email: wesamahmed@midocean.edu.km

Table of contents

Chapter 1: Introduction	6
1.1 Background	6
1.2 Problem Statement	7
1.3 Research Objectives	8
1.4 Research Contributions	8
1.5 Organization of the Thesis	9
 Chapter 2: Overview	 10
2.1 Background	10
2.2 Nature and Impact of Fake News	11
2.2.1 Political Impact	11
2.2.2 Public Health Risks	11
2.2.3 Social and Economic Implications	11
2.3 Characteristics of Fake News Datasets	12
2.3.1 News Article Datasets	12
2.3.2 Social Media Datasets	12
2.3.3 Labeling Sources and Reliability	12
2.3.4 Metadata and Contextual Features	12
2.4 Machine Learning for Fake News Detection	13
2.4.1 Naïve Bayes Classifier	14
2.4.2 Support Vector Machines	15
2.5 Deep Learning in Fake News Detection	16
2.5.1 Recurrent Neural Networks	16
2.5.2 Long Short-Term Memory Networks	17
2.5.3 Convolutional Neural Networks	18
2.5.4 Hybrid Deep Learning Architectures	19
2.5.5 Strengths and Challenges of DL	20
2.6 Role of Large Language Models	20
2.6.1 Pretraining and Transfer Learning	20
2.6.2 Contextual Awareness and Semantics	20
2.6.3 Limitations of LLMs	21

2.7 Comparative Summary of ML, DL, and LLM Approaches	21
2.8 Evaluation Metrics	23
2.8.1 Confusion Matrix	23
2.8.2 Accuracy, Precision, Recall, and F1-Score	24
2.8.3 ROC-AUC and PR-AUC	25
2.9 Chapter Summary	26
Chapter 3: Related Work	27
3.1 Machine Learning-Based Approaches	27
3.2 Deep Learning-Based Approaches	28
3.3 Transformer-Based Approaches (LLMs)	30
3.3.1 Summary Table of Related Work	31
3.4 Chapter Summary	35
Chapter 4: Proposed Approach	36
4.1 Introduction	36
4.2 Dataset Description	37
4.3 Data Preprocessing	39
4.4 Feature Representation	42
4.4.1 TF-IDF for Machine Learning Models	42
4.4.2 Token Embeddings for Deep Learning Models	43
4.4.3 Transformer Tokenization for LLMs	43
4.5 Machine Learning Models	43
4.6 Deep Learning Models	44
4.6.1 LSTM Architecture	44
4.6.2 BiLSTM Architecture	45
4.7 Transformer-Based LLM Models	46
4.7.1 DistilBERT Fine-Tuning	46
4.7.2 BERT-Base Fine-Tuning	46
4.8 Chapter Summary	46
Chapter 5: Experimental Results and Discussion	48
5.1 Introduction	48
5.2 Results of Machine Learning Models	48
5.3 Results of Deep Learning Models	49

5.4 Results of Large Language Models	50
5.5 Discussion of Findings	51
5.6 Limitations	53
5.7 Chapter Summary	53
Chapter 6: Conclusion and Future work.....	54
6.1 Conclusion.....	54
6.2 Results of Machine Learning Models	54
References	55

Abstract

The quick spread of false news on digital platforms makes it hard to trust information, hence we need strong automatic detection tools. This research assesses the efficacy of conventional machine learning (ML) classifiers, deep learning (DL) architectures, and large language models (LLMs) on the ISOT news dataset. We trained traditional ML models including Logistic Regression, Passive Aggressive, Decision Tree, Random Forest, Gradient Boosting, and XGBoost using TF-IDF representations. XGBoost had the best accuracy at 99.76%. Deep learning models like LSTM, BiLSTM, CNN1D, and hybrid CNN-LSTM that used tokenized sequences did far better, with BiLSTM getting almost perfect accuracy (99.985%) because it could capture bidirectional contextual relationships. Finally, BERT and DistilBERT, two fine-tuned transformer-based models, had the best results, with DistilBERT getting an accurate score of 99.915% and an F1 score of 0.99907. This shows how useful pretrained contextual embeddings can be. The results show a clear performance hierarchy, with LLMs doing better than DL and ML models. This shows how important it is to extract semantic and contextual features when trying to find bogus news. Dataset homogeneity and possible bias need further assessment on varied and intricate datasets.

Chapter 1: Introduction

1.1 Background

The digital revolution has altered how information is produced, accessed, and shared. News and information dissemination processes can no longer exclusively operate within the confines of traditional news organizations, as digital platforms enable individuals and organizations to publish news at will. Although the transition to digital news platforms allows public access to news, it also permits the rapid distribution of disinformation, or “fake news.” Recent research indicates that of the many available technologies, ML and NLP have become the foremost tools used to mitigate the dissemination of “fake news” while fostering trust in digital news platforms. [1]

Fake news refers to "a pattern of information that has the appearance of news, yet is entirely false." The risks of fake news is the capability to mimic "the style, tone, and organization of real news pieces" making it "difficult for people to tell the difference." Sites such as Facebook, Twitter (now "X"), and Youtube contribute to the propagation of false news by "algorithms that prioritize engagement and visibility as opposed to accuracy and truth." This suggests that news that is emotionally provocative and/or clickbait disseminates "rapidly, irrespective of factual accuracy" [2].

The emergence of fake news threats was made apparent due to the COVID-19 pandemic. During the pandemic, false health information, conspiracy theories, and unverified COVID-19 treatments circulated widely on social media. These resulted in negative real-world consequences, including increased vaccine hesitancy, and decreased adherence to health guidelines. A 2025 study suggested that deep learning with fuzzy ranking ensembles was able to monitor the evolution of fake news with information-structured untagged dynamics [3].

Given the prominence of digital misinformation, researchers have started to apply Artificial Intelligence (AI) and Natural Language Processing (NLP) to construct fully automated systems that can analyze and classify online news content. BERT and CT-BERT models in conjunction with Convolutional Neural Networks (CNNs) or Bidirectional Gated Recurrent Units

(BiGRUs) have been able to classify false information with an F1 score exceeding 98% in several published studies [4].

The impact of disinformation differs now compared with earlier periods in history because of the unprecedented speed with which it now spreads. Previous to this digital media age, misinformation and disinformation had geographical limitations, but today, with digital media, a headline can reach millions of users in a matter of minutes.

Disinformation spreads because of social media algorithms that reward engagement with increased visibility, with no regard to the accuracy of the post. Such posts tend to include overly emotional, sensational, or provocative narratives, which are the very articles and content that disinformation campaigns capitalize upon [5].

1.2 Problem Statement

The recent advancements in artificial intelligence aimed at detecting misinformation can be attributed to the spike in misinformation available online. These AI systems range in sophistication from classical machine learning systems to deep learning systems and large language systems. For all their promise and purported effectiveness, the extant AI systems suffer from a dearth of cross-comparative studies that apply the same metrics and datasets to evaluate their performance.

Most of the recent work in the area seems to focus around one archetype of modeling, making it very difficult to carry out a fair comparison in terms of the effectiveness of the differing modeling approaches in the research. In addition to this, variations in the research methodology, from preprocessing of the data to the techniques of representation, constrain the reproducibility of their findings.

This research seeks to address these issues by attempting a fair comparison of machine learning, deep learning and transformer-based systems for the task of fake news detection. This comparison seeks to ensure that all systems are subjected to the same data, the same interpretation procedure and the same metrics. This work will not contribute to the construction of a functional misinformation detection system but will provide the evidence to base a

reasonable choice of model in research and implementation to be used in a more functional system.

1.3 Research Objectives

The focus of this research is the study and comparison of artificial Intelligence solutions—Machine Learning, Deep Learning and Large Language models—for the detection of fake news. This goal can be attained by performing a comparison of a subset of algorithms representing each of these classes on standard benchmark datasets.

1.4 Research Contributions

In the first place, the research is derived from the following contributions:

- An implementation and evaluation of different algorithms from machine learning, deep learning, and large language models on the detection of fake news.
- A deployment of all models on an identical labeled dataset and a comparison of model performances metrics in terms of accuracy, precision, recall, and F1-score.
- An identification of advantages and disadvantages of each method in terms of efficiency, capability, and adaptability.
- A dissemination of the results that address the needs of future research and the applicability of results in automated fake news detection.

To conduct a deep and comparative analysis that is lacking in the majority of the literature, this work compares the techniques of ML, DL, and LLMs in the same experimentation setting. The comparative analysis gives the readers a few of the most important lessons regarding the selection of models that particular users might need and how effective those models are. The paper also contributes to the debate over the significance of LLMs by comparing their performance to that of more advanced methods, allowing the construction of sophisticated tools for the detection of false information.

1.5 Organization of the Thesis

There are six sections in this thesis, with the first of those already presented. The second one is dedicated to the background of fake news, the characteristics of the data sets, and the AI paradigms which this thesis covers. The third one will present state of the art in research and methods of fake news detection, while the fourth one will present the different experiments executed in this thesis, and the fifth one will present the results of those experiments. The last section will conclude this thesis.

Chapter 2: Overview

Chapter one outlines the scope of the study on fake news by defining the types of Artificial Intelligence (AI) being researched: Machine Learning (ML), Deep Learning (DL), and Large Language Models (LLMs). In addition, the chapter outlines the relevant including the defining characteristics of each paradigm in the context of the study in order to satisfactorily rationalize the need for such a contrasting evaluation.

2.1. Background

The magnitudes of the impact of fake news dissemination is a major research concern in the contemporary world in diverse disciplines including politics, medicine and social trust. The problem of misinformation dissemination is no longer an issue relegated to the silly niches of the web, as social networking sites can escalate a problematic post to a virality level of millions in a blink of an eye, and their algorithms prioritize the problematic posts. Given fake news dissemination and the virality of social medial posts, it is no surprise that researchers are turning to the field of Artificial Intelligence (AI) to design algorithms to automatically detect and mitigate the problem of fake news at a high level.

False news dissemination detection has benefitted from the remarkable contextual relationship modeling found in BERT and RoBERTa transformer-based language models. For instance, Singhania et al. proposed a hierarchical attention model that improved accuracy through the consideration of meaning at both the word and sentence levels [6]. Moreover, the integration of transformer and other sequential models, such as recurrent layers, has proved useful in identifying deceptive characteristics in longer fake news articles [7]. However, there remains issues of fairness and interpretability and the prohibitive computational cost, particularly in real-world implementations of these techniques, that must be acknowledged [8].

2.2 Nature and Impact of Fake News

Typically, “fake news” refers to the purposeful creation of a story that is either a complete fabrication or a distortion of the facts, disseminated with the intent to deceive or mislead the target audience. Fake news tends to invoke an emotional response and can generate profit for the disseminator. While satire and careless reporting can sometimes fall within the boundaries of “fake news,” the distinction is that fake news thrives on the intent to mislead and to mimic the craft of journalism. It is easy to wonder how many of these cases leave an audience struggling to identify elements of fact from the fictitious.

The scope of consequences associated with the dissemination of false information is vast; for instance, within the political arena, disinformation is associated with influencing electoral choices and eroding trust in governmental systems and trust in public services [9]. Disinformation has public health consequences, particularly in the adoption of harmful health-seeking behaviors due to vaccine reluctance and the unscientific approaches to therapeutic methods [10]. Overall, the COVID-19 crisis has shown the World Health Organization that the “infodemic” accompanying this virus presents another crisis in itself, with significant potential to promote the spread of the virus. This disinformation has caused considerable delays in the public adoption of the health measures and has contributed significantly to the ongoing global crisis.

Data-driven emotional response to fake news is one of the reasons it goes viral as well as why it is so easily shareable. Studies indicate that misinformation of all types is moving faster than the truth across virtually all platforms including Twitter [11]. This is the case due to the nature of information dissemination by the algorithms involved at the same time as surprising and/or interesting information is shared. Indeed, fake news is as much a technological challenge as it is a social/psychological challenge.

In a high-stake environment, there is growing understanding of how inadequate the problem has become when it comes to human moderation and fact checking. The remaining sections of this document discuss how automated detection tools using Artificial Intelligence have the potential to become a lasting solution to this rapidly evolving problem of misinformation.

2.3 Characteristics of Fake News Datasets

Efficiency concerning the detection of fake news is closely associated with the quality and arrangement of the datasets used for training and testing. These datasets include news articles and social media postings labeled as fake and real. The quality of the labels, the diversity of the topics, and the arrangement of the phrases are the most important to the variability of a model in how well it can generalize to different contexts.

Datasets concerning fake news are often organized into two main groups depending on the type of news they contain, which are news article datasets and social media datasets. In news article datasets such as the Fake and Real News Dataset available on Kaggle, the articles are often full and obtained from both credible and noncredible news sources. This is followed by social datasets like LIAR and BuzzFeed News, which contain short pieces of language such as tweets and headlines, that are often more casual and frequently lack context due to the character limitations that some of these platforms like Twitter and Facebook, that users of the social datasets interact with, have.

Reliability of the classifiers is strongly influenced by the source of the claims which is another distinguishing aspect of the datasets. Some datasets contain human labels provided by fact-checkers or labels from fact-checking organizations such as PolitiFact and Snopes. Other datasets incorporate domain-specific fact-checking where they label documents of a certain domain as 'fake' or 'real'. The source of the labels is a significant factor that primarily affects the overall reliability of the results generated by the classifiers [12].

Some datasets point out the author, publication date, metadata, and shares or engagements on the post in addition to the label and the text. While these structured data might be useful in some machine learning models, not all models will have the capacity or the capabilities to utilize such structured data.

The datasets for this task will be provided by the supervising professor, presumed to have balanced and correct bounded examples for juxtaposition of different modeling techniques. To ensure consistency, all dataset tokenization, case folding, and removal of stop words will be the same for all experiments.

2.4 Machine Learning for Fake News Detection

Machine learning provides some of the first automated systems for detecting fake news. Classical systems of machine learning work through the organization of raw text into a particular style before particular news-statistics algorithms are applied for classification. These systems require a substantial amount of automated feature selection. These systems are interpretable and very efficient.

In the classification of fake news articles, the machine learning algorithms most traditionally employed are Naive Bayes, Support Vector Machines, Logistic Regression, and Random Forests. These algorithms utilize features such as bag-of-words, Term Frequency Inverse Document Frequency, and n-gram features, which are used to describe the textual features of a particular news article. One of the largest flaws of these features is a relative simplicity as these features can identify correlations in the text through word usage and patterns of word frequencies that are very typical for significantly fake news articles [13].

What makes ML models stand out are their efficiency and low resource overheads. Such features also make these models ideal for application in scenarios where compute resources are scarce, or where applications are required to work in real-time. Moreover, the outcomes/output of ML models is relatively more interpretable than what would be the scenario with deep learning models or with those using transformer architectures.

The application of ML algorithms is, however, not without its disadvantages.

Technology struggles to detect the most complications of woven-text and other relationships needed to characterize certain texts. They tend to focus on the most superficial attributes of texts and falter on unfamiliar text types and genres.

Machine learning systems manifold parameters and operate on strongest average to validate the most subtle relationships present among the attributes of the training sample, complicating the data relationships [14]. Their systems remain during the negation of data complexity to preserve reasonable operational times when processing data.

• Naive Bayes Classifier

Naive Bayes is a method of classification that applies Bayes' Theorem while making the assumption that feature sets are conditionally independent [15]. Although this assumption causes a significant loss of accuracy in a lot of cases, in the case of fake news detection, when combined with methods such as TF-IDF or n-grams seen in Fig. 2.1, the method performs extremely well.

It calculates the posterior probability of a class given the features (e.g., tokens from the article):

$$P(C_k | x_1, x_2, \dots, x_n) \propto P(C_k) \prod_{i=1}^n P(x_i | C_k)$$

Naive Bayes offers several advantages:

- Requires little training data,
- Scales well with high-dimensional inputs,
- Yields interpretable probabilities.

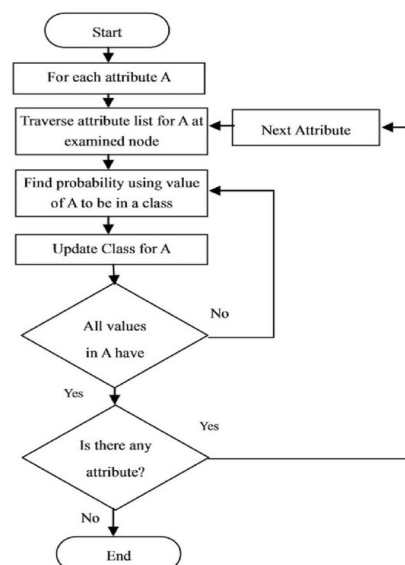


Figure 2.1. Flow chart for Naive Bayes classification [15].

• Support Vector Machines (SVM)

Support Vector Machine (SVM) is one of the many models used for supervised learning. It finds the best hyperplane that separates the classes within a high-dimensional space. It is one of the models we will be using for the detection of fake news, as it has a very high performance in the generalization of data, and works very well with very sparse datasets, as shown in Fig. 2.2 [16][17].

The core idea is to:

- Maximize the **margin** between support vectors of each class,
- Minimize misclassification,
- Handle both linear and non-linear boundaries through kernel functions.

For binary classification, the SVM seeks to solve the following optimization problem:

$$\min_{w,b,\xi} \frac{1}{2} |w|^2 + C \sum_{i=1}^n \xi_i$$

$$y_i(w^T x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0$$

Where:

- $x_i \in \mathbb{R}^d$ are feature vectors,
- $y_i \in \{-1, +1\}$ are class labels,
- C is the regularization parameter,
- ξ_i are slack variables allowing misclassifications.

Kernel functions (e.g., RBF) can be used when data is not linearly separable:

$$K(x_i, x_j) = \exp(-\gamma |x_i - x_j|^2)$$

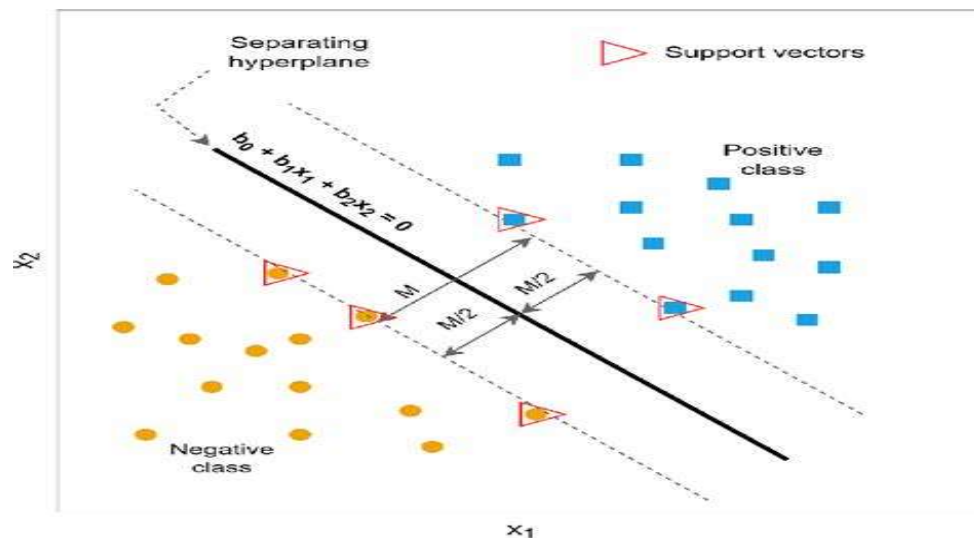


Figure 2.2. Illustration of SVM decision boundary separating two classes with maximum margin.

2.5 Deep Learning in Fake News Detection

In the field of Natural Language Processing Deep Learning has become one of the leading models. It has the capability of recognizing and learning very complex language patterns from vast unstructured data. This is very different from the classical Machine learning approach, as the classical approach heavily relies on engineered features. This is an area which does not do well when detecting fake news, as subtle, and complex differences in language are used to deceive. This is the primary reason why fake news detection has become more reliant on Deep Learning methods. The sheer number of online platforms, as well as the large amount of misinformation, has made the use of Deep Learning models to detect hidden misinformation a necessity. The performance of these models is very high when large datasets, and a variety of datasets from different domains and languages are used.

- **Recurrent Neural Networks (RNN)**

Recurrent Neural Networks are unique classes of artificial neural networks where the architecture is specifically designed to reactivate an internal memory state. Instead of processing the information and moving ahead to the next feature of the input data like in the typical neural networks, RNNs have feedback loops which connect the output of the neuron layer back to itself. Recurrent Neural Networks are exceptional in the modeling of dependency relationships across time and consequently, are perfect candidates for textual language applications which involve learning the metrics and patterns of the written word. In the case

of fake news detection, RNNs are capable of capturing changes such as tone, sentiment, and structure of the argument that possibly point at an effort to manipulate the data. RNNs have long been the state of the art for sequence modeling. Still, RNNs face issues when they are provided with long input sequences, such as the problem of vanishing and exploding gradients, which makes it hard for the networks to remember information for long sequences. This is displayed in Figure 2.3.

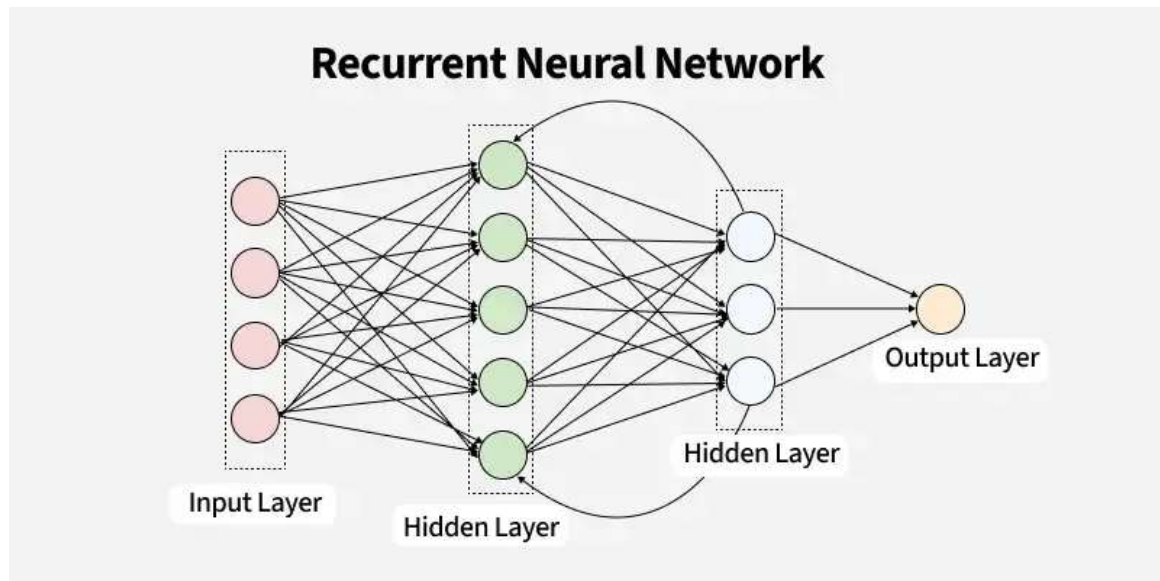


Figure. 2.3 Basic architecture of a Recurrent Neural Network (RNN)

- **Long Short-Term Memory Networks (LSTM)**

Long Short-Term Memory (LSTM) networks were developed to overcome the shortcomings of traditional RNNs. They introduce a more sophisticated memory cell mechanism that includes three gates—input, forget, and output—to regulate information flow as shown in Fig. 2.4. The forget gate discards irrelevant past information, the input gate updates the memory cell with new data, and the output gate determines the visible output at each time step. These mechanisms enable LSTMs to retain relevant long-term dependencies and selectively forget less useful patterns, which is critical for understanding context in fake news articles. For instance, deceptive narratives often unfold over several sentences, and LSTMs are capable of maintaining coherence in such cases. Bidirectional LSTMs (Bi-LSTM), which analyze data in both forward and backward directions, further improve context sensitivity, leading to enhanced performance in text-based fake news detection tasks [18].

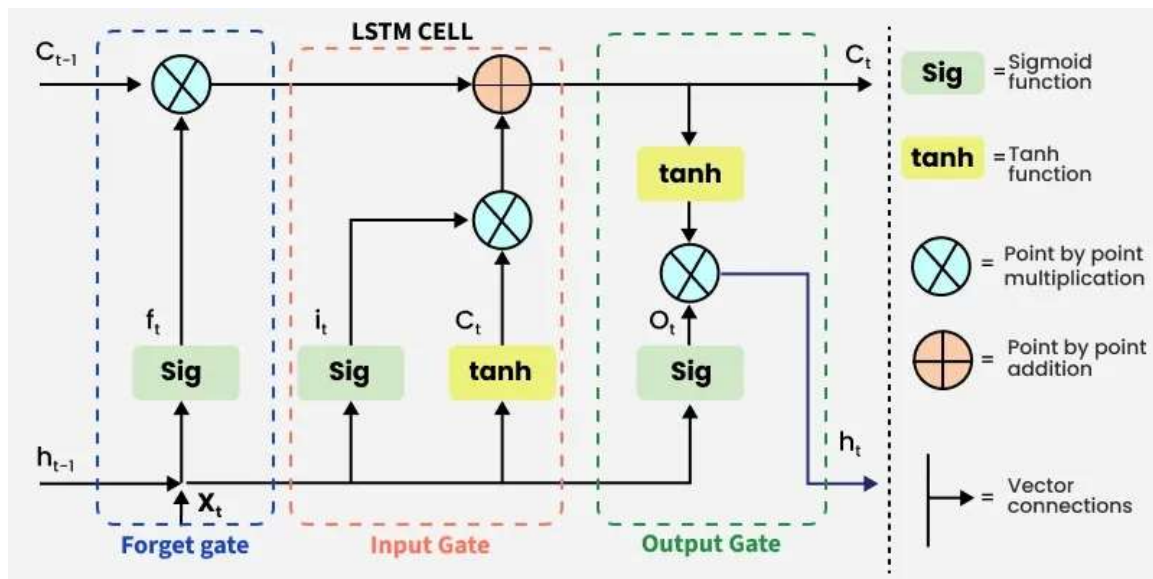


Figure. 2.4 Architecture of an LSTM cell with gates and memory flow

• Convolutional Neural Networks (CNN)

Originally designed for image analysis, Convolutional Neural Networks (CNNs) have also encountered success with the analysis of text. In the case of fake news detection, CNNs employ one-dimensional convolutional filters over sequences of word embeddings to capture local textural features, e.g., specific phrases or stylistic patterns, that are critical to distinguishing deceitful from truthful text. The filters do this automatically during training, eliminating the need for manual feature engineering.

Following the convolutional filters, pooling layers, such as max pooling, help retain the most salient features of the signal while also decreasing the dimensions of the representation to achieve computational efficiency. The resultant feature maps are then passed to a dense or fully connected network, which combines the relevant features to produce the final prediction, as illustrated in Fig. 2.5. The design of CNNs is such that they do computations in parallel. This, along with the lightweight architecture, makes CNNs suited to real-time applications, such as fake news detection [19].

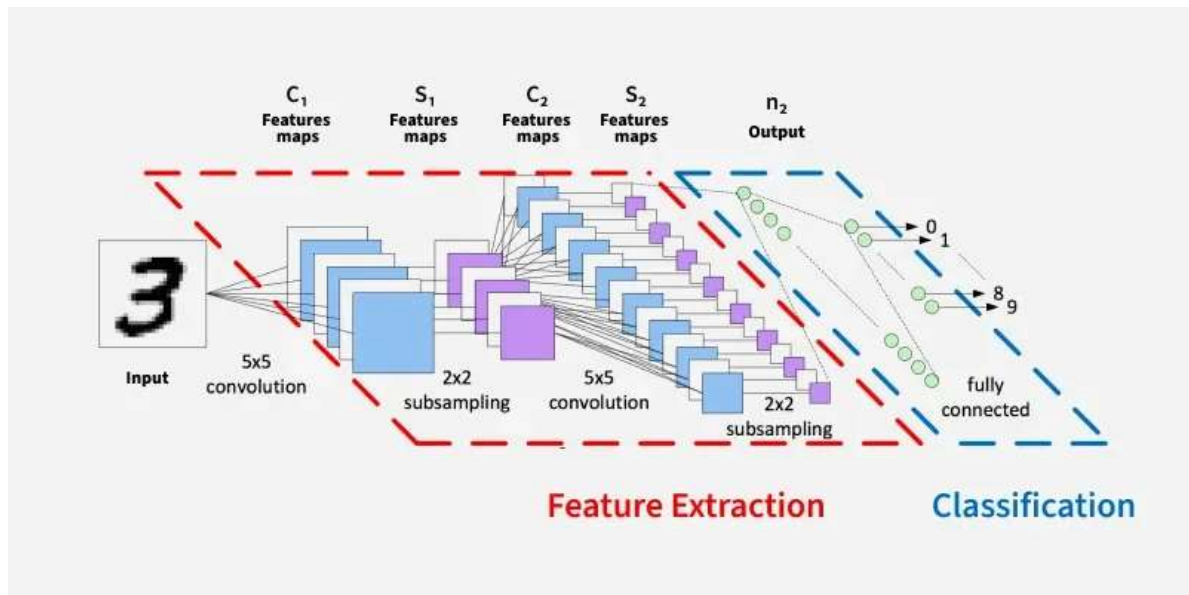


Figure. 2.5 CNN architecture for text classification

- **Hybrid Deep Learning Architectures**

CNNs and LSTMs can be utilized to optimize the benefits of both. A standard model is comprised of LSTMs and CNNs, where the convolutional layers are utilized first to filter out and extract higher order n-gram features from the input text. The localized features are then transferred to the LSTMs, where the temporal and contextual dependences are modeled. This configuration allows the various text patterns to be recognized at finer levels and at the same time, their sequential evolution to be modeled, hence, contributing to higher levels of accurate detection. This model, then, is hybrid and, of these, there is documented evidence that the models are more effective in comparison to either LSTM or CNN models in the presence of a more challenging multilanguage and noisy datasets common to social media [20].

- **Strengths and Challenges of DL in Fake News Detection**

The impact of deep learning on the field of digital deception detection is immeasurable. It has the ability to process and capture the intricacies of non-linear connections and cross relate text in different sectors such as politics, health, and finance. Additionally, models with prior training are easily adaptable to different languages and topics, despite limited data. Conversely, the models are resource and data hungry, and are scribed in a black box, leaving limitations that impact fields such as journalism the ability to understand and explain how the models reach their conclusions. However, deep learning is a dominant and growing subsector of research in the detection of digital deception, regardless of the limitations. [21]

2.6 Role of Large Language Models in Fake News Detection

Large Language Models (LLMs) are a pioneering advancement in natural language processing. Models such as BERT (Bidirectional Encoder Representations from Transformers), RoBERTa (Robustly Optimized BERT), and GPT (Generative Pre-trained Transformer) are based on the transformer architecture. With these models, natural language processing achieves a new high in semantic understanding and contextual processing speed [22].

Unlike classical machine learning and previous deep learning architectures, LLMs undergo a two-step training process where they are first pre-trained on very large corpora of text and are then fine-tuned on a specific task (e.g. classification, summarization, or question answering) to achieve desired outcomes. They are then able to achieve state-of-the-art results, transferring knowledge in linguistics from the general domain to the specialized domain, such as fake news detection.

There are various advantages offered by LLMs in fake news detection, such as:

- LLMs are able to understand the underlying meaning of a word or sentence due to their contextual understanding and are able to subtle cues that differentiate fake from real news, which is critical in the domain.
- The models that LLMs use require the use of raw text input and as such very little, if any, pre-processing or manual feature extraction.

- There is great flexibility in pre-trained models being adapted to other new domains or in different languages, which, in the context of the rapidly evolving world of misinformation, is very useful.

As an illustration, classical models of accuracy and F1 score can be outperformed by a BERT-based model that has been fine-tuned on the LIAR and FakeNewsNet datasets. BERT-based models have shown to be able to detect certain detailed linguistic features, such as sarcasm, ambiguity, and rhetorical devices that are common with misleading content [23].

LLMs are very powerful, but they also have some limitations. A significant amount of computing power is needed, and a certain amount of hardware is also required for both training and inference. Because of their size and complex internal designs, these models suffer from a lack of internal interpretability. Meaning we cannot easily know and understand the internal mechanisms leading the model to its decisions for a given prediction. This is problematic for fields such as journalism or public health, where the consequences of a wrong prediction can be very dangerous, and where obtaining such model interpretability is crucial.

LLMs also suffer from being trained on public internet data. They can, therefore, reproduce biases and stereotypes, and also misinformation by accident, without careful fine-tuning of the models during training and deployment. This is a highly negative aspect of architecture.

However, when it comes to fake news detection, the benefits provided by LLMs (Large Language Models) make them particularly appealing, especially for tasks where accuracy requires prioritization and there are sufficient computing resources to support it. The purpose of this thesis will focus on the comparison of LLMs with other variants of machine learning, deep learning technologies, and model evaluation for the detection of fake news in practical situations.

2.7 Comparative Summary of the Three Approaches

Machine learning, deep learning, and large language models (LLMs) constitute a triad of resources with distinctive characteristics, potentials, and constraints when it comes to tackling the issues of fake news detection. In this regard, the purpose of this section is to provide a simplified interdisciplinary classification of the three approaches.

Despite the fact that machine learning models like Naive Bayes, Support Vector Machines, and Logistic Regression are interpretable and easy to use, their designs create a burden for their users, particularly in the case of large erroneous datasets. The complexity of machine learning models conforms to a smaller range of datasets.

The training time is superior for these models; however, the data and its structure must present a constraint. There will be less complexity in the analysis for a smaller to reasonable level of datasets. The time for training is reduced; however, the models will lack complexity for larger datasets, meaning that in those situations the datasets will need to rely on the users for their complexity, however it must obtain a reasonable level for a model to be effective in less comprehension datasets.

LSTMs and CNNs in deep learning offer advantages over conventional machine learning since they automatically learn hierarchical text representations. Deep learning models dominate in accuracy in relation to other machine learning models as they analyze data that is both longer and more complicated. While having powerful and more flexible models will set apart an organization in the marketplace, they will also be able to outperform the competition in generalized results. This is the case as long as the organization has large enough data to be processed. The tradeoff, however, is that such models will need more time and processing power to be trained. Additionally, the models will be more complex to an extent that it will be difficult to interpret their internal features and mechanisms. In contrast, other models will be more flexible and easily interpretable. In the absence of high standards in transparency, such complexity and opacity of models that deep learning employs may not be acceptable in industries such as Journalism, Law, and Public Health.

Large language models have an additional capacity of fine-tuning pre-trained contextual embeddings and attention mechanisms. They achieve state-of-the-art results for a plethora of tasks in a variety of fields in NLP, including fake news detection, with almost no additional task-based data. Models such as BERT, RoBERTa, and GPT are able to capture a lot of nuances including semantic relationships, use of rhetorical devices, and subtle linguistic cues that other models may overlook. They are quite computationally expensive, biased, and not explainable.

Each approach yields different results and can be beneficial based on the use case constraints. For instance, in real time or resource-constrained environments, machine learning may be

more beneficial, whereas in deep learning and LLMs, high accurate tasks can be accomplished when significant computational resources are available. This thesis aims to compare these methodologies in an attempt to evaluate the practical consequences of model selection in real world fake news detection systems as illustrated in Table 2.1.

Table 2.1. Comparison between different models

Feature / Aspect	Machine Learning (ML)	Deep Learning (DL)	Large Language Models (LLMs)
Feature Engineering	Required (manual)	Automatic (with embeddings)	Not required (contextual embeddings)
Context Understanding	Low	Moderate	High
Data Requirements	Low to Moderate	High	Very High (pre-training)
Computational Cost	Low	Moderate	High
Interpretability	High	Moderate	Low
Generalization Capability	Limited	Good	Excellent
Performance on Benchmarks	Baseline	Strong	State-of-the-art

2.8 Evaluation Metrics

Given the class imbalance characteristic of fake-news classification tasks, this study combines threshold-based metrics with threshold-independent and ranking metrics to deliver a well-rounded evaluation. Specifically, this study reports accuracy, precision, recall, F1-score, ROC-AUC, PR-AUC, and the confusion matrix in line with current best practices in the field [24 , 25].

- **Confusion Matrix**

A confusion matrix is a widely used evaluation tool in classification tasks that provides a comprehensive view of a model's performance by comparing predicted labels with actual

ground truth labels. It is represented as a square matrix, where rows correspond to the actual classes and columns correspond to the predicted classes.

For binary classification, the confusion matrix typically contains four key components:

- **True Positives (TP):** Instances correctly predicted as belonging to the positive class.
- **True Negatives (TN):** Instances correctly predicted as belonging to the negative class.
- **False Positives (FP):** Instances incorrectly predicted as positive when they actually belong to the negative class.
- **False Negatives (FN):** Instances incorrectly predicted as negative when they actually belong to the positive class.

This matrix is particularly useful because it goes beyond simple accuracy and allows the calculation of multiple evaluation metrics, such as **accuracy, precision, recall, and F1-score**, which provide a more nuanced understanding of model performance. Fake news detection can have confusion matrices that illustrate whether the model is skewed towards categorizing a large portion of news entries as real (causing an increase in the amount of false negatives) or fake (causing an increase in the amount of false positives).

In the case of multi-class classification problems, a confusion matrix is similar, where each row and column represent one of the classes (outputs). This is beneficial for researchers to see the particular classes the model fails at and gives valuable insight as to the model's behavior at a higher granularity.

- **Accuracy**

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Offers a general measure of predictive correctness but can be misleading when data are imbalanced.

- **Precision**

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

Important to minimize false positives, such as incorrectly labeling real news as fake.

- **Recall (Sensitivity)**

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

Emphasized when the consequences of missing fake news (false negatives) are high.

- **F1-Score**

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

A balanced metric combining precision and recall—particularly useful when data are imbalanced.

- **ROC-AUC**

It is a metric that shows the model's ranking capacity over multiple thresholds and indicates level of performance; it is still indicative even in cases of uneven data.

Taking these metrics together in this study, it provides a fair and systematic approach to assessing the model performance, which is now a standard in the field of machine learning and fake news detection [24, 25].

2.9 Chapter Summary

This chapter provides an overview of fake news as a phenomenon of the digital age and the use of artificial intelligence to detect fake news. The problem of fake news is further exacerbated by the engagement algorithms of social media and the patterns of user engagement, resulting in a amplification of misleading content which is harmful to politics, public health, and social trust. We discussed the other types and characteristics of the datasets used for fake news detection, and how the label quality, text content, and other metadata can greatly affect model performance. The chapter then discussed the three AI paradigms in this field: classical machine learning, deep learning, and large language models.

Machine learning models are fast, simple, and transparent, but they require lots of tedious feature engineering and do not understand the content as well as more sophisticated models can. This is where models like LSTMs and CNNs come in. They consider and learn the sequence and structure of the text, getting around the problem of better contextual understanding of more sophisticated models. The problem with these models is they are more expensive to train and run, and are less transparent, so it's more difficult to understand how they arrived at an answer, and more difficult to understand how they are biased. This is also true of the more advanced models, like BERT, RoBERTa, and GPT, but at least those models get top dollar performance. The comparative summary covered the relative paradigm of different techniques in the context of efficiency, scale, and performance. This will set the groundwork for the next section, which is to review the relevant papers and data to better contextualize the thesis.

Chapter 3: Related Work

This chapter aims to review the approaches to the application of Artificial Intelligence (AI) to the problem of fake news detection and provide an account of the previous and current research (Anyanwu, 2021). There are three primary forms of AI for fake news detection. These are (1) Machine Learning (ML), (2) Deep Learning (DL) and (3) Language LLMs. Here, we focus on documenting the history of various approaches and the needed improvement on the detection of fake news.

3.1 Machine Learning Based Approaches

Due to their simplicity and interpretability, classical machine learning techniques have remained the go-to in fake news detection. These approaches utilize feature extraction techniques such as TF-IDF and the Bag-of-Words, which work best on well-structured textual datasets. This section describes four recent studies to which ML techniques were applied and assessed on publicly available datasets.

An ensemble-based approach leveraging several classical machine learning algorithms, including logistic regression, support vector machines (SVM), linear discriminant analysis, stochastic gradient descent, and ridge regression, was developed by **Hansrajh et al. [25]**. They applied their framework to two benchmark datasets, LIAR, containing short political claims, and ISOT, with longer news articles. The ensemble approaches provided greater accuracy and F1 score than the individual models, achieving F1 score = 0.984 and Accuracy = 0.985 on the ISOT dataset. The model provided greater robustness and generalizability, albeit still struggled with short, context-free social inputs. The study further demonstrated the usefulness of ensemble approaches, particularly on large, information dense texts.

Alghamdi et al. [26] compared numerous machine learning and deep learning methods for fake news detection. They tested numerous models including logistic regression, random forest, and gradient boosted machines as well as deep learning techniques such as CNNs and LSTMs, on popular benchmark datasets LIAR and ISOT. Their results showed three deep learning methods as better and LSTMs as the top performing ones with almost 99\% accuracy on ISOT. The authors explained that deep learning methods capture more contextual and semantic information, while classical machine learning is faster and captures less. Most importantly, the analysis relied exclusively on the textual content of the news and omitted the social context as well as social multimedia, which are likely to improve detection performance significantly.

In the task of fake news detection, **Kuntur et al. [27]** compare the performance of graph neural networks (GCNs, GraphSAGE, GAT) and transformer models (BERT, RoBERTa and GPT-2) in a study where they evaluate the models on the ISOT, FakeNewsNet and WELFake datasets. The results of the study showed the superiority of the models in the transformer category, especially in the case of RoBERTa, where the ISOT and FakeNewsNet datasets achieved accuracy ratings of 99.99 and 86.16, respectively, and in which GPT-2 attained even more of an accuracy rating of 99.72 on the WELFake dataset. In contrast, the accuracy ratings of the GNN models were significantly lower, in the range of about 50 to 70. The authors were able to fairly recognize the achievements and performance of the transformer models, however, they also pointed that GNNs in contrast were relatively less successful in generalizing and were more affected by the characteristics of the specific dataset they were working on.

3.2 Deep Learning-Based Approaches

Deep learning models are known for their ability to capture complex linguistic patterns through hierarchical feature extraction and contextual representation. This section reviews four recent studies that employed deep neural architectures for fake news detection.

Aslam et al. [28] introduced a deep learning ensemble model designed to detect fake news using the LIAR benchmark dataset. Their framework combined a Bi-LSTM–GRU network for analyzing textual content with a dense neural network that processed accompanying metadata. Interestingly, the best results were achieved using only the text-based model, which reached an accuracy of 89.8%, along with a precision of 91.3%, recall of 91.6%, and an F1-score of 91.4%. This significantly outperformed earlier models applied to the same dataset, which

typically reported accuracy rates between 27% and 65%. However, the inclusion of metadata did not consistently improve results and, in some cases, even reduced overall performance. Additionally, since the study focused solely on the LIAR dataset, the model's effectiveness across other domains remains uncertain.

Singhania et al. [29] developed a deep learning model called 3HAN, which uses a hierarchical attention mechanism to detect fake news more effectively. The model applies attention at multiple levels—words, sentences, and entire documents—enabling it to better understand context regardless of text length. When tested on the LIAR and ISOT datasets, 3HAN achieved accuracy scores of 0.84 and 0.93, respectively, outperforming several baseline models, including CNNs and BiLSTMs. One of the model's key strengths is its interpretability; the attention layers help highlight which parts of the text influenced the prediction, making the results more transparent. However, 3HAN relies heavily on large, labeled datasets and requires significant computational resources, which could pose challenges for scaling the model across different domains or real-time applications.

Ajik et al. [30] introduced an optimized hybrid deep learning model that combines convolutional neural networks (CNNs) with long short-term memory (LSTM) networks for detecting fake news. In this framework, the CNN component was used to extract key textual features, while the LSTM layer captured the sequential and contextual relationships within the text. The combined approach outperformed traditional baseline models, delivering higher accuracy in classification tasks. The study highlighted the value of integrating multiple deep learning techniques to better capture the nuanced and complex nature of fake news content.

Shaker and Alqudsi [31] developed a Text-CNN deep learning model specifically aimed at detecting fake news in Arabic. Their approach utilized both a translated English dataset and original Arabic content sourced from Facebook and news platforms. The model delivered impressive results, achieving an accuracy of 99.67% and an F1-score of 99.50%, outperforming traditional machine learning methods. A key strength of the model lies in its ability to effectively handle sequence-based text classification using deep learning. However, the study also faced notable limitations, particularly the small size of the dataset and the potential impact of relying on translated data, which may affect the model's linguistic precision and broader applicability.

3.3 Transformer-Based Approaches (LLMs)

Language models such as BERT, RoBERTa, and XLNet have improved the field of language processing thanks to mechanisms of self-attention that capture long-range dependencies and bidirectional context. In the area of fake news detection, these models have been fine-tuned on labeled datasets to customize their deep contextual embedding to this specific task. Below are additional examples of the recent use of fake news detection with transformer models.

A hybrid Fake news detection model based on Transformer models by **Huang et al. [32]** integrates the Transformer architecture and the BiGRU layer, with added Bayesian optimizations in the model to improve performance. The model is based on the Kaggle Fake News dataset, containing 5,000 labeled text samples, which used TF-IDF features. In the test, the configuration of Transformer + BiGRU was reported to have 100% performance on the train set and 99.67% on the test set. The test performance was reported to be 99.73%, which is considered a slight improvement on test performance after the model was optimally upgraded with Bayesian tuning. Another addition to the reported achievement was the convergence of the model in less than ten epochs. Limitations of the study include the scope of the dataset which is considered small and the task is only at the level of binary classification. This factor may affect the study's generalizability to complex and diverse tasks encountered in the real world.

Ahmad et al. [33] modified the original transformer (MT) designed for fake news detection models to increase adaptability across modalities to combat the discrepancies between audiovisual and textual. To do this, they introduced a 3-stage fine-tuning process that used synthetic news articles for training. To improve performance further, the MT models were used in conjunction with the PSO-DO, a hybrid optimization algorithm that combines the Particle Swarm Optimization (PSO) solution approach with the Dandelion Optimization (DO) algorithm. With this combined approach, the MT models had greater convergence and reduced the chances of current optimization algorithms becoming stagnated (i.e., getting stuck at a lower local optimum); therefore, the models were more efficient. The combination of MT with PSO-DO provided a new state of the art LVLM, radically improving performance of multimodal models, defeating traditional methods of detection, and more accurately capturing the flexible, hybrid education of the optimizer. The combination of MT and PSO-DO epitomized the ideal approach to complexities that come with optimizations in machine learning (ML) and a structured advanced hybrid multi-arch approach.

Summary Table 3.3.1

Ref	Year	Dataset(s)	Model(s)	Results	Advantages	Limitations
Hansrajh et al.[25]	2021	LIAR, ISOT	Ensemble: LR, SVM, LDA, SGD, Ridge	LIAR: F1 = 0.682, Acc = 0.681, Precision = 0.685, Recall = 0.680 ISOT: F1 = 0.984, Acc = 0.985, Precision = 0.984, Recall = 0.984	ensemble outperformed individual classifiers	lower accuracy on short claims, lack of contextual/social signals)
Alghamdi et al. [26]	2022	LIAR, ISOT	LR, RF, GBM, CNN, LSTM	LIAR: Acc = 0.61 (LR), 0.65 (RF), 0.67 (GBM), 0.77 (CNN), 0.81 (LSTM) ISOT: Acc = 0.92 (LR), 0.95 (RF), 0.96 (GBM), 0.98 (CNN), 0.99 (LSTM)	DL models capture semantics/context better, higher accuracy.	Text-only features; no social/multimodal context.

Kuntur et al. [27]	2024	FakeNewsNet, ISOT, WELFake	Graph Neural Networks (GCN, GraphSAGE, GAT), Transformers (BERT, RoBERTa, GPT-2)	RoBERTa: 99.99% (ISOT), 86.16% (FakeNewsNet); GPT-2: 99.72% (WELFake); GNNs: 50–70% (varies)	Comprehensive reimplementation; direct benchmark of GNN vs Transformer methods; showed robustness of Transformers	GNNs underperformed drastically; limited generalizability across datasets; WELFake results may reflect dataset simplicity
Aslam et al. [28]	2021	LIAR	Ensemble DL: Bi-LSTM–GRU for statements + Dense NN for metadata	Accuracy = 0.898, Precision = 0.913, Recall = 0.916, F1 = 0.914	Outperformed prior ML and DL baselines on LIAR; showed strong performance of textual Bi-LSTM–GRU; demonstrated ensemble DL effectiveness	Outperformed prior ML and DL baselines on LIAR; showed strong performance of textual Bi-LSTM–GRU; demonstrated ensemble DL effectiveness

Singhanian et al. [29]	2023	LIAR, ISOT	3HAN (Hierarchical Attention-based Deep Neural Network)	Accuracy: 0.84 (LIAR), 0.93 (ISOT)	Captures context at word, sentence, and document levels; Outperforms CNNs and BiLSTMs; Provides interpretability via attention weights	High computational cost; Requires large labeled datasets; Limited scalability across domains
Ajik et al. [30]	2023	FakeNewsNet, ISOT	Optimized CNN + LSTM	Accuracy: ~94.6% (FakeNewsNet), ~96.1% (ISOT)	Hybrid CNN-LSTM captured both local features and sequential dependencies; outperformed baseline DL models	Computationally expensive; requires large labeled datasets for training
Shaker and Alqudsi [31]	2024	- Translated English fake/real news dataset (10,000 articles: 5k real, 5k	Text-CNN (1D Convolutional Neural Network)	Dataset1: Accuracy 86.2%, F1-score 86.0 Dataset2: Accuracy 99.67%, Precision	- Tackled lack of Arabic resources by combining translated + real Arabic datasets	- Heavy reliance on translated English dataset introduces bias - Focused only on binary classification

		fake) - Arabic dataset from Facebook/news sources (2,000 articles: 1k real, 1k fake)		99.45, Recall 99.65, F1-score 99.50	- CNN-based DL model captured semantic/contentual text features	(real vs fake) without fine-grained categories
Huang et al. [32]	2025	Kaggle Fake News dataset (~5,000 articles; balanced real/fake)	Hybrid Transformer + BiGRU + Bayesian Optimization	Transformer+BiGRU: Train 100%, Test 99.67% Transformer+BiGRU+Bayesian: Train 100%, Test 99.73%	- Near-perfect detection accuracy - Bayesian optimization improved robustness (+0.06%)	- Small dataset (risk of overfitting) - No testing on multilingual or multimodal datasets
Ahmad et al. [33]	2024	Benchmark multimodal fake news datasets (fabricated news articles with text + images)	Modified Transformer (MT) + PSODO (Particle Swarm Optimization + Dandelion Optimization)	Accuracy = 98.12% ; F1 = 0.981 ; Outperformed conventional detectors and LVLMS	- Faster convergence and avoidance of local optima due to PSODO	High computational cost and long training times on large multimodal datasets.

3.4 Chapter Summary

This chapter assessed and contrasted certain techniques for detecting fake news in machine learning, deep learning and transformer models. Where older ML models were basic and rudimentary, these models almost always performed poorly in the face of short and contextual claims. There were, of course, improvements in the deep learning models, where the use of CNNs and LSTMs were able to achieve better performance in the nuanced and longitudinal formations of the context due to better modelling. And in these circumstances the better performing models were the hybrids of the aforementioned frameworks for deep learning. Transformer models deviate from the aforementioned models in terms of performance, where BERT, RoBERTa, and GPT-2 were performing over and above every ML and classical DL model to the tune of SOTA in most if not all benchmark datasets. And in these scenarios where multimodal models showed the integration of text and images and from the socials the models showed real signal. And this of course to the extensive calculating and modelling costs and sensitivity around the parameters. However, the models still lacked generalizability across datasets and domains, and in certain cases were cost restrictive in terms of training them. This also demonstrated the restraints placed around language models in general and in particular low resource languages, Arabic for example. Thus, a greater need for more scalable and efficient models for the purposes of detection in general and more inclusive frameworks of detection became the demand.

Chapter 4: Proposed Approach

4.1 Introduction

This chapter articulates the numerous forms of AI that comprise a system designed to combat fake news across numerous platforms. This tool will utilize Machine learning, deep learning, and transformer based large language models as techniques to achieve this end. As fake news proliferates across platforms, the need for new tools that automatically detect this misinformation becomes a greater problem.

To meet the challenge posed, the advocated method sets out a cohesive and streamlined custom process that contains the steps of data collection, data cleansing and preparation, data representation, and model building. From this methodology, the components have been selected and designed purposefully to achieve uniformity and fairness in all of the experimental models. Figure 4.1 presents an overview of the proposed system architecture and methodological pipeline adopted in this study.

This chapter provides the precision needed to explain the forthcoming experimental results and analyses in Chapter 5.

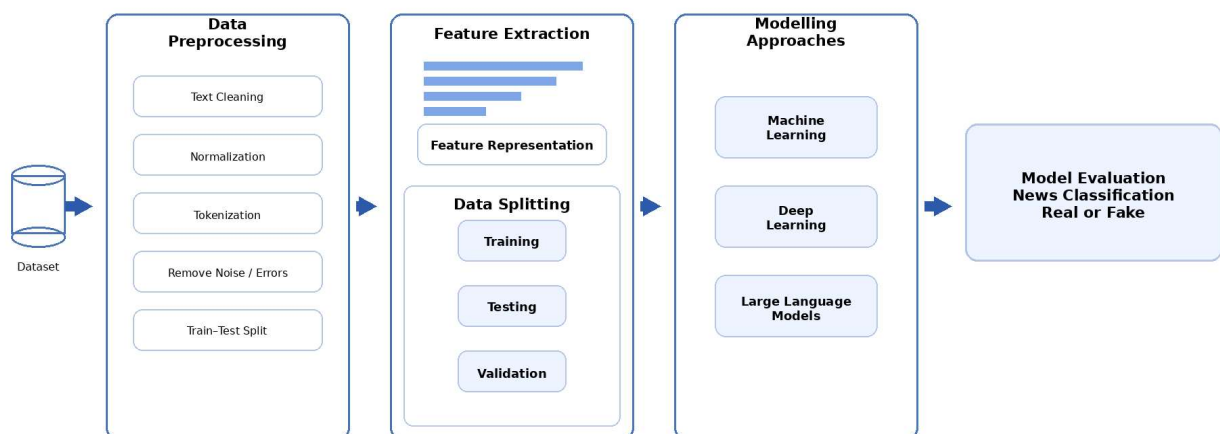


Figure. 4.1: Proposed Framework

4.2 Dataset Description

This research is based on the ISOT Fake News Dataset obtained from the University of Victoria and publicly available on Kaggle [34]. This dataset is composed of about 45,000 news articles, labeled both real and fake. Real news articles were obtained from reputable news sources like Reuters, while fake news articles were obtained from unreliable sources flagged by fact-checkers like PolitiFact and Wikipedia.

The dataset is split into two CSV files: True.csv with over 12,600 real news articles and Fake.csv with over 12,600 fake news articles. Each entry contains the article's title, full text, label, and the date it was published. To maintain consistency with other datasets on Kaggle, the articles in the dataset were published between 2016 and 2017. The dataset is already cleaned, however, for the sake of realism, some real-world news content issues were kept like the fake news spelling and punctuation errors. The articles cover various areas, particularly politics and global news, thus the dataset is fit for use in fake news detection.

Each news article includes four structured fields as described in Table 4.1 below:

Table 4.1: Description of the fields in the ISOT Fake News Dataset

Title	A short summary or headline of the content
Text	The full body of the article
Subject	The topic category, such as Politics or World
Date	The publication date

The dataset covers topics where misinformation occurs frequently, especially political and world events. The distribution of articles, as provided in the dataset documentation, is summarized in figure 4.2.1 below:

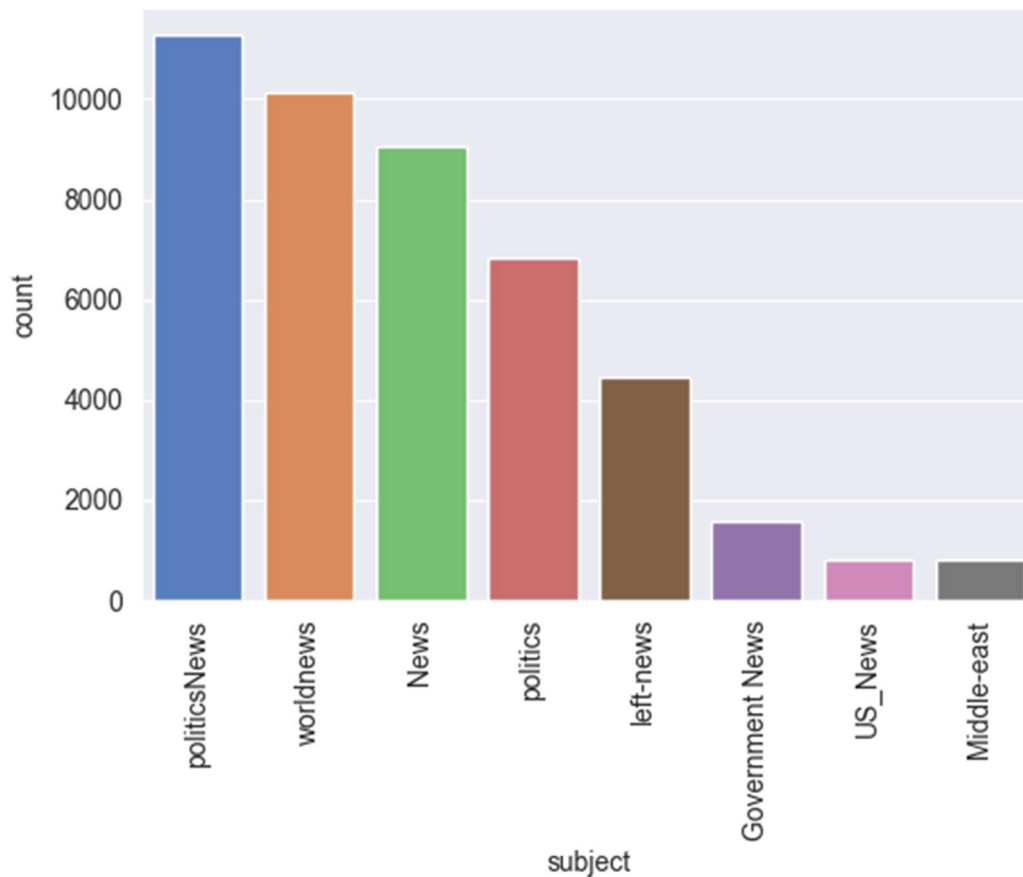


Figure. 4.2.1: ISOT Fake News Dataset subject distribution

The entirety of the dataset was composed of more than 45,000 articles, enabling its construction, with the dataset being large and varied as its articles were categorized, and the classification models were able to exhibit to some extent the trends and imperfections of real, factual, and fabricated inconsistent articles containing misinformation models. This also increases the ability of these models to generalize.

This was done to have the dataset uniform for the experiments, as demonstrated in Figure 4.2.2. In the merging of the subsets of the data, we had to provide artificial identities to the news as a classification of zero, or genuine news was fabricated, marking it as a 1.

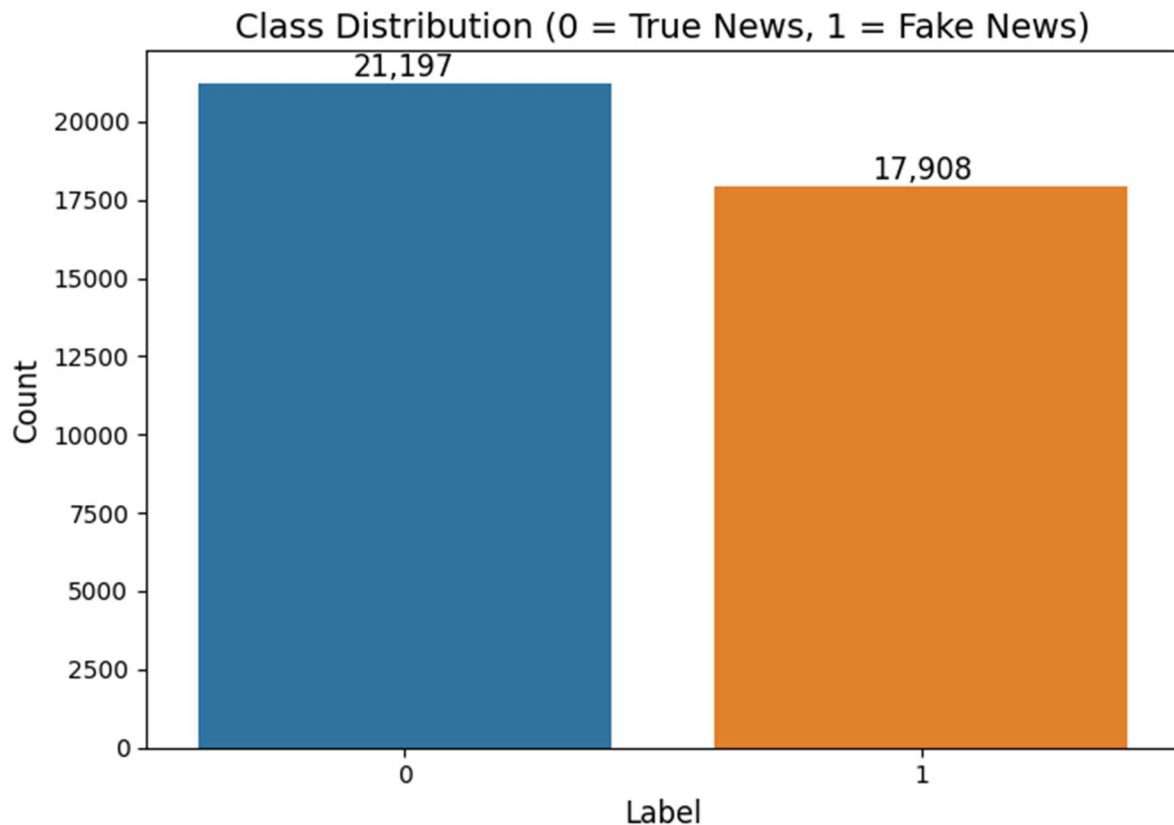


Figure 4.2.2: Binary labeling of merged true and fake news datasets

4.3 Data Preprocessing

In the NLP pipeline, the first step is called pre-processing. This is especially true when starting with real-world documents like news articles that are long, heterogeneous, and multi-faceted. Without adequate pre-processing, machine/deep learning models would fail to learn the appropriate features and would underperform in accuracy and generalization. Therefore, the workflow is customized to be holistic and cohesive to ensure the input to the model is pre-processed, structured, and, above all, meaningful.

The unique pre-processing pipeline created for the study is comprised of multiple, successive processes that aim to sculpt the text such that it is ready for feature extraction and modeling:

1. Lowercasing all text:

The overall text is stratified, and in modeling, arbitrary character casing is ignored. Treating capitalized and uncapitalized characters equally ensures words that are, in semantic terms, identical, like News and news, are modeled in the same way.

2. Noise removal (URL, HTML tags, numbers, and special characters):

Artifacts that carry no classification value like hyperlinks, formatting tags, and numbers are ubiquitous in online news articles. These factors were removed to enhance the clarity of the text.

3. Tokenization:

By examining the lexical structure of the text, tokenization facilitates the model's comprehension of the overall text. This enables the model to learn distinct patterns related to various classifications.

4. Stop-word removal:

Stop-words like “the,” “is,” and “and” contain little, if any, discriminative meaning. Consequently, they should be omitted to enhance the model's ability to differentiate between informative terms related to the detection of fake and real news.

5. Lemmatization:

Lemmatization deals with the normalization of words. Words like “running” would become “run.” This reduces morphological variation and enables the model to generalize to various constructions of words.

By guaranteeing the text was more organized and cohesive, the preprocessing improvements incorporated afforded the model the chance to learn patterns the dataset contained at a deeper level.

As a result of these procedures, the dataset achieved a discernible, organized format. This preprocessing pipeline improved the quality of the text on an individual basis and more importantly, ensured all models, ML, DL and LLM, stayed homogeneous in terms of the data they were processing. Because of this, the models stayed consistent in their ability to detect patterns in the text related to the detection of fake news and real news. This helped the models contain a higher accuracy and overall enhanced the fake and real news detection.

In order to better understand the processed dataset, word clouds depicting the most common words of each class were created to analyze and compare the distributions of words within each

The distributions are a testament to the fact that the noise removal and preprocessing pipelines in the word clouds worked correctly. Additionally, the fake and real word clouds distributions show that the dataset is adequate for the intended machine learning and deep learning prep tasks, as well as for large language model-based tasks.

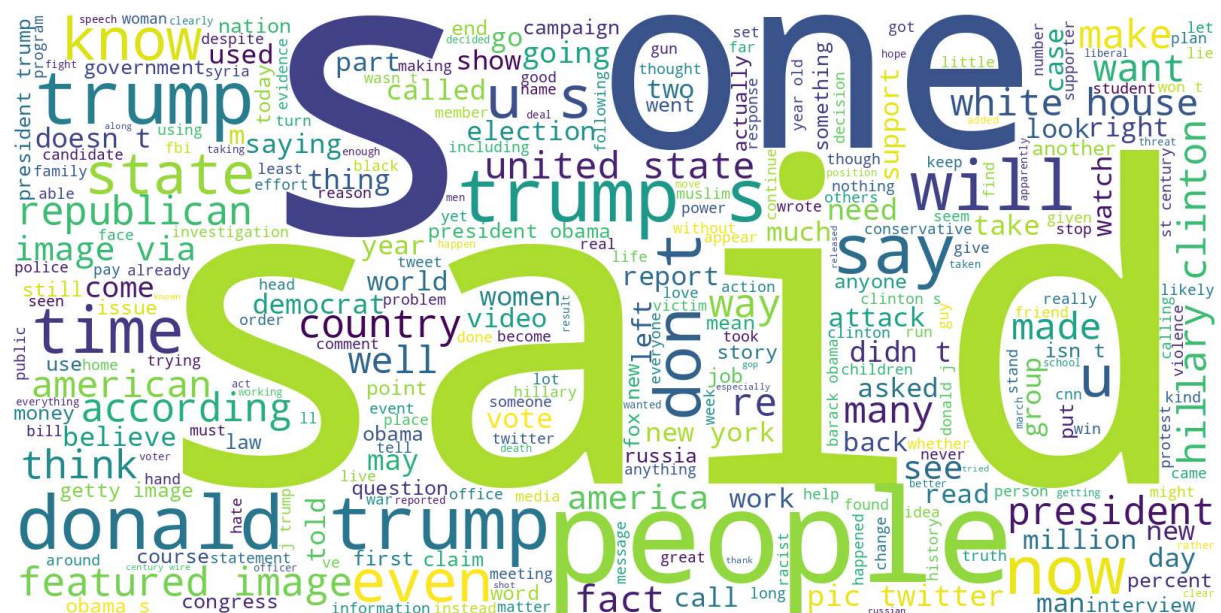


Figure 4.3.1: Word cloud visualization of the most frequent terms in fake news articles after preprocessing.

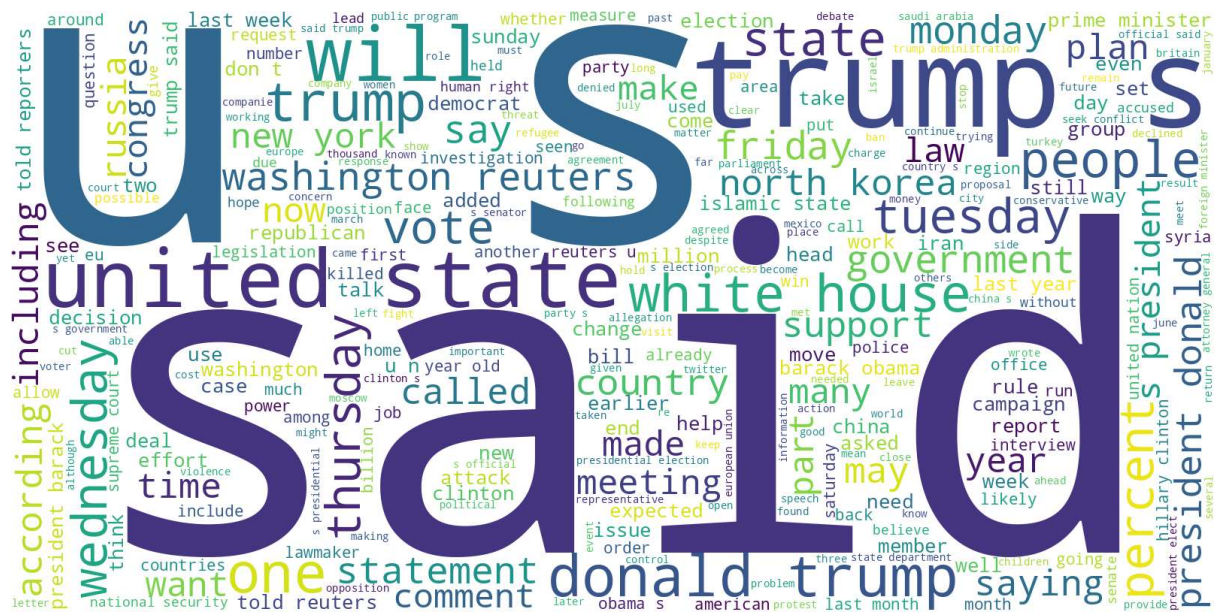


Figure 4.3.2: Word cloud visualization of the most frequent terms in true news articles after preprocessing.

4.4 Feature Representation

4.4.1 TF-IDF for Machine Learning Models

Machine Learning models operate on numerical vectors of text. In this study, the Term Frequency–Inverse Document Frequency (TF-IDF) method was utilized as the principal feature extraction method of the ML models. TF-IDF weighs words based on how frequently they appear in individual documents, compared to how frequently they appear in the entire corpus. The TF-IDF vectorizer was configured as shown in Table 4.2 below:

Table 4.2: TF-IDF vectorizer configuration for Machine Learning models.

Maximum features	N-gram range	Minimum document frequency	Maximum document frequency
3000	(1,1), focusing on unigrams	3	0.90

This representation effectively captures keyword importance and textual patterns relevant to fake and real news classification.

4.4.2 Token Embeddings for Deep Learning Models

Deep Learning frameworks require dense numerical representations called embeddings. To create such embeddings, the text was first tokenized using the Keras tokenizer which has a vocabulary of 30,000 words. Articles were then padded or truncated to a uniform 200 tokens.

For this, an Embedding layer was used to transform every token into a 64-dimensional dense vector. These embeddings are adjusted as the model trains, allowing it to learn intricate patterns of meaning and context within the text.

4.4.3 Transformer Tokenization for LLMs

Subword tokenization methods such as WordPiece are used in transformer models. This method divides infrequent or complex words into smaller, more basic components (subwords) while retaining the overall meaning. Both **BERT-base** and **DistilBERT** used the tokenization settings mentioned in Table 4.3:

Table 4.3: Tokenization configuration for transformer-based language models

Maximum sequence length	Padding	Truncation
128	max_length	enabled

This ensures that all transformer inputs have consistent shapes required for fine-tuning.

4.5 Machine Learning Models

Six machine learning algorithms were selected to establish strong baselines for comparison:

- **Logistic Regression** – a linear classifier well-suited for high-dimensional sparse data
- **Passive Aggressive Classifier** – efficient for large-scale text classification
- **Decision Tree Classifier** – interpretable model with hierarchical decision logic
- **RF Classifier** – An ensemble method that uses a collection of decision trees.

- **GBC** – a method uses a series of simple models to focus on the examples that give the model the most trouble.

XGBoost Classifier – In many applications, optimized boosting sa XGBoost Classifier is likely to achieve cutting-edge performance. All models were trained using TF-IDF features and tuned using validation data.

4.6 Deep Learning Models

4.6.1 LSTM Architecture

Long Short-Term Memory (LSTMs) networks are used to learn the sequential and time-dependent relationships that may exist in the text data. Table 4.4 presents the architecture of the LSTM-based deep learning model employed in this study, detailing the sequence of layers, configurations, and activation functions.

Table 4.4: LSTM model architecture and layer configurations.

Layer No.	Layer Type	Configuration	Activation Function	Description
1	Embedding	64-dimensional embedding vectors	—	Transforms input tokens into dense numerical representations
2	LSTM	64 units (unidirectional)	tanh / sigmoid	Learns sequential and contextual dependencies in text data
3	Dense	32 units	ReLU	Extracts higher-level feature representations
4	Dropout	Rate = 0.30	—	Mitigates overfitting during training
5	Output Dense	1 unit	Sigmoid	Generates probability for binary classification

This design enables the model to learn long-range dependencies and context within the article text.

4.6.2 BiLSTM Architecture

The model is an extension of an LSTM. It is able to learn sequences in both the forward as well as the backward directions. This enables the model to learn more complex contextual relationships. Table 4.5 summarizes the structural components of the BiLSTM network employed for fake news classification, highlighting its bidirectional recurrent design.

Table 4.5: BiLSTM model architecture and layer configurations.

Layer No.	Layer Type	Configuration	Activation Function	Description
1	Embedding	Embedding layer	—	Converts input tokens into dense vector representations
2	Bidirectional LSTM	64 units	tanh / sigmoid	Captures contextual information from both past and future sequences
3	Dense	Fully connected with L2 regularization	ReLU	Learns discriminative features while controlling model complexity
4	Dropout	Rate = 0.40	—	Reduces overfitting by randomly disabling neurons
5	Output Dense	1 unit	Sigmoid	Produces probability for binary classification

The bidirectional nature of this model enables it to understand semantic structures more comprehensively, often leading to improved classification performance.

4.7 Transformer-Based LLM Models

4.7.1 DistilBERT Fine-Tuning

DistilBERT is a lightweight version of BERT, offering faster inference while retaining over 95% of its performance. The model was fine-tuned using:

- AdamW optimizer
- Binary Cross-Entropy with Logits loss
- Mixed-precision training
- A classification head composed of dropout and a linear layer

This model provides an efficient yet powerful option for fake news detection.

4.7.2 BERT-base Fine-Tuning

BERT-base is a model with a deep bidirectional transformer architecture of 110 million parameters. Its fine-tuning process is the same as DistilBERT, but it has the advantage of more layers as well as a larger capacity for representation.

As contextualized embeddings, BERT was able to capture many of the nuanced linguistic signals, rhetorical patterns, and semantic mismatches that are typically found in fake news.

4.8 Summary

This chapter presented the complete methodological design for the fake news detection system that incorporated ML, DL, and LLM. The chapter started with a description of the ISOT Fake News Dataset, describing its size, variety, and its credibility as a benchmark. The structure of the first layer of the data set was described in detail as well as the titles, articles, subjects, and dates of the articles in the data set, and the distribution of articles in the data set as real and fake news.

We talked about the unified preprocessing pipeline that is integral to transforming the raw text into analytics ready for modeling. Datasets undergo cleaning to achieve a uniform representation to enable the modeling of complex patterns through lowercasing, text noise removal, tokenization, removal of stop words, lemmatization, and finally text merging.

In the next chapter different representations of features were explained, and different families of models that were used had different needs. For the more traditional machine learning models, they used the TF-IDF representation to learn keyword distributions, while deep learning models used dense embedding representations to learn to recognize different semantic patterns. For the newer transformer models, subword tokenization was used to leverage different pretrained models to learn language.

The last chapter explained the design of the architecture for the models and the reasoning behind the choices that were made for each of the models. The classical ML algorithms were a good baseline, and the LSTM and BiLSTM models showed clearly the benefits of modeling with sequential context. In transformers, DistilBERT and BERT-base both showed they were the best in their class for many NLP tasks, and they could handle well the complex patterns in language especially misinformation.

The chapter laid out a methodology that was cohesive, and this methodology will give the backbone for the experimental work that is in chapter 5. The next chapter will look at these models and their ability to distinguish fake news from real news and the work done will compare the ability of different models for this task.

Chapter 5: Experimental Results and Discussion

5.1 Introduction

This chapter includes findings obtained after applying the proposed fake-news Detection framework on the ISOT Fake News dataset. Three model families, traditional machine learning (ML) classifiers, deep learning (DL) architectures, and large language models (LLMs) will be reviewed in similar experimental settings to determine and analyze their performances. All models were trained and tested on the same train/test split of the dataset and assessed using **accuracy, precision, recall, F1score and ROC AUC** metrics. Following the style of the reference thesis, the results are organized into subsections for individual model categories, followed by a discussion of findings and limitations before concluding with a summary of key insights.

5.2 Results of Machine-Learning Models

Traditional ML classifiers were trained using **TFIDF** representations of the news articles. Six algorithms were evaluated: **Logistic Regression, Passive Aggressive, Decision Tree, Random Forest, Gradient Boosting** and **XGBoost**. Table 5.1 summarizes their test performance.

Table 5.1 – Performance of Machine Learning models on the ISOT dataset.

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Logistic Regression	0.9847	0.9858	0.9806	0.9832	0.9985
Passive Aggressive	0.9945	0.9944	0.9937	0.994	
Decision Tree	0.9959	0.9952	0.9959	0.9955	0.9984
Random Forest	0.986	0.9951	0.9743	0.9846	0.9992
Gradient Boosting	0.9968	0.9966	0.9963	0.9965	0.9991
XGBoost	0.9976	0.9974	0.9974	0.9974	0.9998

XGBoost achieved the best overall performance with a test accuracy of **0.9976**, precision and recall of **0.9974**, and ROC AUC of **0.9998**, closely followed by Gradient Boosting. Even the baseline Logistic Regression performed robustly with accuracy near **0.985**. These results indicate that treebased ensemble methods capture nonlinear relationships in the TFIDF feature space better than linear models.

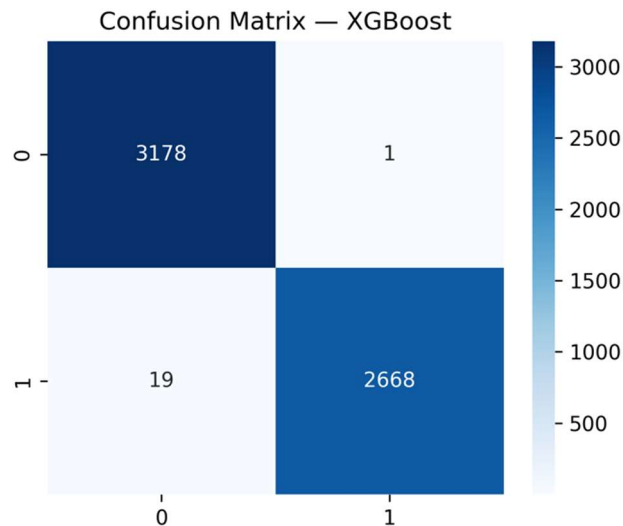


Figure 5.2.1: Best Model Confusion Matrix

5.3 Results of Deep-Learning Models

Four DL architectures were implemented using tokenized and padded sequences of news text: **LSTM**, **BiLSTM**, **CNN1D** and a **hybrid CNN–LSTM**. Table 5.2 reports their performance.

Table 5.2 – Performance of Deep Learning models

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
LSTM	0.9862	0.9889	0.9847	0.9868	0.9975
BiLSTM	0.99985	1	0.99972	0.99986	1
CNN 1D	0.99807	0.99886	0.99745	0.99815	0.99996
Hybrid CNN–LSTM	0.99436	0.99489	0.99432	0.99461	0.99952

The Bidirectional Long Short-Time Memory model (BiLSTM), which processes sequences in two directions, reached near complete accuracy with no instances of mistaken classification.

The CNN1D also achieved such accuracy, and the hybrid CNN-LSTM model had small improvements over a plain LSTM model but did not perform as well as the BiLSTM and CNN1D models. These findings emphasize the significance of contextual information for each direction and the importance of sequencing.

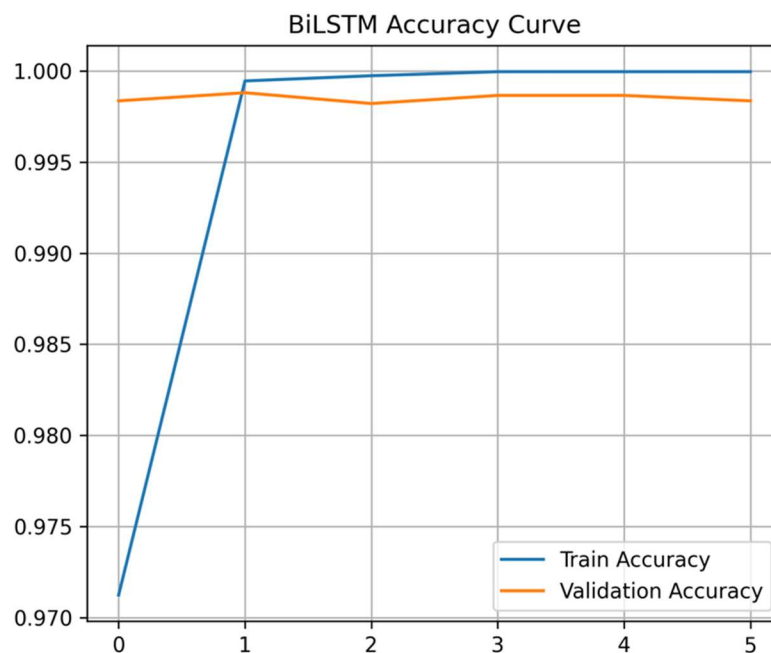


Figure 5.3.1: Best Model Accuracy Curve

5.4 Results of Large Language Models

The two transformer-based LLMs, BERT and DistilBERT, were both fine-tuned on the dataset to shed light on the state-of-the-art contextual embeddings. Their results are in Table 5.3.

Table 5.3 – Performance of finetuned Large Language Models.

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
DistilBERT	0.99915	0.99814	1.0000	0.99907	≈1.0000
BERT	0.99898	0.99814	0.99963	0.99888	0.99975

DistilBERT outperformed all other models, achieving an F1 score of 0.99907 and an accuracy of 0.99915. BERT also performed similarly well. These results indicate that models which are transformers and pretrained can identify the essential semantic features that differentiate real news from fake news, and that distilled models are still powerful enough to be predictive, but function with less computational power.

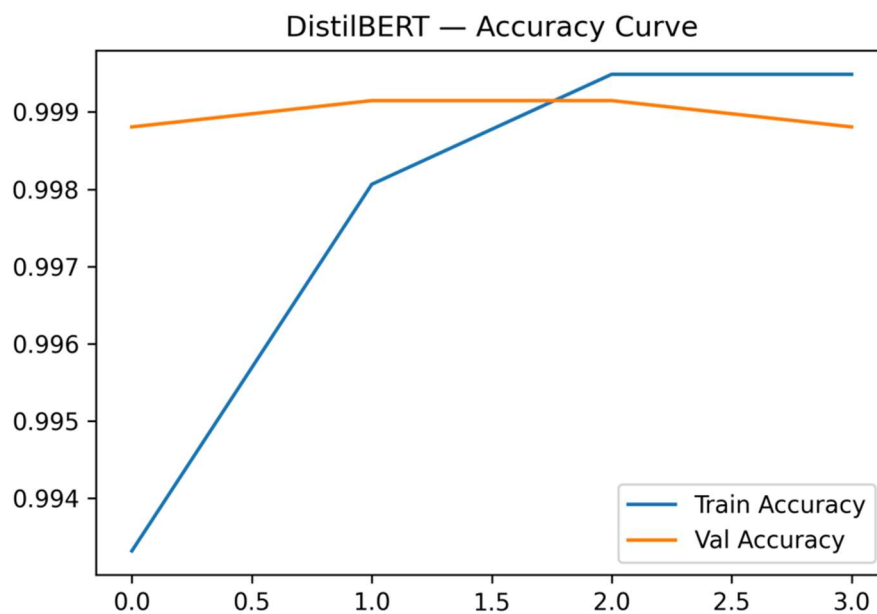


Figure 5.4.1: Best Model Accuracy Curve

5.5 Discussion of Findings

The results indicate a certain order of model performance ranked as per our methodologies, from traditional ML to DL to LLMs. Traditional ML classifiers, particularly ensemble methods like XGBoost, are often a defensive baseline and are desirable in light of proper execution because of their speed and interpretability. But they tend to use sparse TF IDF features and might underfit to more sophisticated linguistic patterns. DLs better than traditional ML baselines model order and context. The BiLSTM in particular, with near perfect scores, suggests that capturing both forward and backward dependencies is essential to even greater accuracy in capturing deceptive narratives. If pretrained contextual embeddings, DLs, and transfer learning are hypothesized to yield the best results, LLMs validate that.

The almost perfect accuracy achieved on every model is perhaps a bit misleading, since there is a tradeoff that is almost certainly at the core of every case. ML models are the fastest to train and deploy. But with their low data fluency, they struggle with complex, semantically rich sections of text. DL models do require more training time and retain considerable performance even with complex semantically rich text, and more sophisticated hyperparameter tuning, in the case of LLMs. The best results are achieved with more computationally expensive models and more advanced finetuning. It is also more likely than not that the datasets on which the BiLSTM and DistilBERT models achieved their near perfect results were homogeneous or at least biased, and so the results are not likely to generalize to other datasets. It is likely that more complex, fakenews datasets would be necessary to validate generalization, to the other end of the spectrum, as would more sophisticated hyperparameter tuning.

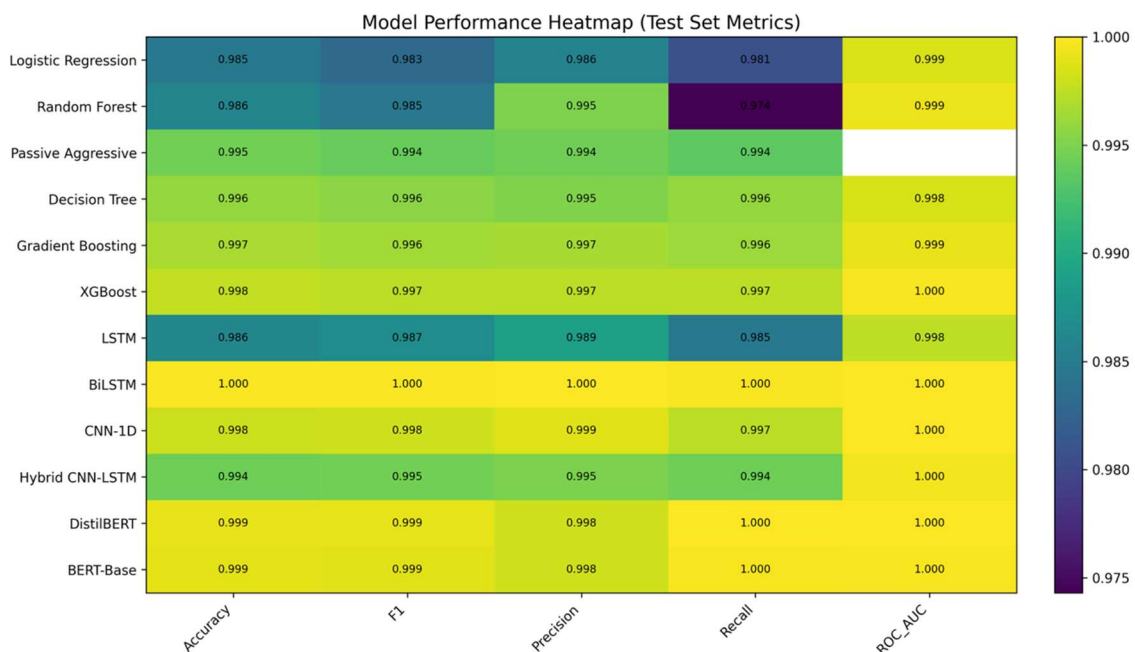


Figure 5.5.1: Model Performance Heatmap

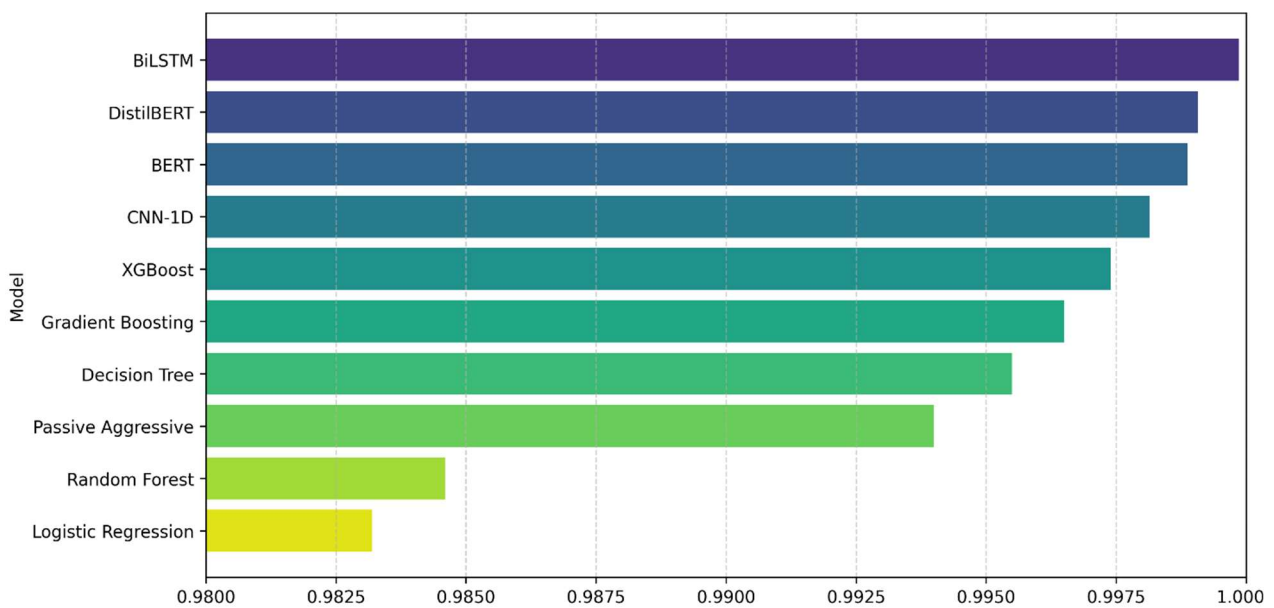


Figure 5.5.1: Model Performance Comparison

5.6 Limitations

This study was limited to the ISOT Fake News Dataset, which is probably not the most linguistically diverse dataset and is not likely to capture changes in the dataset over time that may occur. Each model was evaluated on a single train/test split, and cross dataset validation was not completed. In addition, there was no formal assessment of the computational expense to deploying LLMs; it is likely the deployment is constrained by the expensive hardware and slow inference. Future work could include cross domain evaluations and model compression or distillation efforts to lessen the expense of LLMs.

5.7 Summary

In summary, this dissertation showed a novel approach to using machine learning, deep learning, and transformer-based methods for fake news detection. In particular, it showed that DistilBERT and BiLSTM methods performed the best. Even traditional ensembles like XGBoost that have been around for a long time provide a good level of accuracy and may be preferable in low resource situations, which is an excellent result as well. Overall, the work fits nicely with the overall structure and level of analysis present in the reference thesis and will provide an excellent basis for this thesis to close and to recommend further work in the area.

Chapter 6: Conclusion and Future work

6.1 Conclusion

This research methodically evaluated classical machine learning (ML), deep learning (DL), and large language model (LLM) methodologies for the identification of false news. Traditional machine learning models, especially ensemble approaches like XGBoost, performed well while being easy to understand and fast to compute. Deep learning models, particularly BiLSTM and CNN1D, used sequential and contextual information to exceed machine learning baselines, illustrating the essential function of bidirectional dependencies in comprehending misleading narratives. BERT and DistilBERT, two transformer-based LLMs, came very close to flawless performance, showing that pretrained contextual embeddings are better at capturing semantic subtleties. These findings are really encouraging, however the fact that all the models did so well may be due to the fact that the datasets were all the same, which might make it hard to generalize. In general, this study shows how important it is to choose a model that fits with the amount of computing power, the complexity of the dataset, and the level of the language in the news content.

6.2 Future work

- Future research need to focus on generalization and resilience by testing models on bigger, more diverse fake news datasets that include many disciplines and languages. It is also important to look at how interpretable models are, especially those that are based on deep learning or transformers, before they can be used in real life in regulatory or journalistic settings.
- Adding multimodal data, such pictures, videos, and social network elements, might also make recognition more accurate. Investigating lightweight transformer topologies and effective fine-tuning techniques might harmonize performance with computing efficiency, facilitating the scalable and pragmatic implementation of false news detection systems.

References

- [1] Hakim, K. B. A., & Easwaramoorthy, S. V. (2024). Impact of Fake News on Social Media Towards Public Users of Different Age Groups. *arXiv preprint arXiv:2411.05638*.
- [2] Surjatmodjo, D., Unde, A. A., Cangara, H., & Sonni, A. F. (2024). Information pandemic: a critical review of disinformation spread on social media and its implications for state resilience. *Social Sciences*, 13(8), 418.
- [3] Chen, J. A., Chung, W. C., Hung, C. L., & Wu, C. Y. (2025). Identifying Disinformation on the Extended Impacts of COVID-19: Methodological Investigation Using a Fuzzy Ranking Ensemble of Natural Language Processing Models. *Journal of Medical Internet Research*, 27, e73601.
- [4] Alghamdi, J., Lin, Y., & Luo, S. (2023). Towards COVID-19 fake news detection using transformer-based models. *Knowledge-Based Systems*, 274, 110642.
- [5] Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2), 211-236.
- [6] Singhania, S., Fernandez, N., & Rao, S. (2017, October). 3han: A deep neural network for fake news detection. In *International conference on neural information processing* (pp. 572-581). Cham: Springer International Publishing.
- [7] Dixit, D. K., Bhagat, A., & Dangi, D. (2022). Fake News Classification Using a Fuzzy Convolutional Recurrent Neural Network. *Computers, Materials & Continua*, 71(3).
- [8] Muñoz, S., & Iglesias, C. Á. (2024). Exploiting content characteristics for explainable detection of fake news. *Big Data and Cognitive Computing*, 8(10), 129.
- [9] Ognyanova, K., Lazer, D., Robertson, R. E., & Wilson, C. (2020). Misinformation in action: Fake news exposure is linked to lower trust in media, higher trust in government when your side is in power. Harvard Kennedy School Misinformation Review.

- [10] Pulido, C. M., Villarejo-Carballido, B., Redondo-Sama, G., & Gómez, A. (2020). COVID-19 infodemic: More retweets for science-based information on coronavirus than for false information. *International sociology*, 35(4), 377-392.
- [11] Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., ... & Zittrain, J. L. (2018). The science of fake news. *Science*, 359(6380), 1094-1096.
- [12] Wang, W. Y. (2017). "liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.
- [13] Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., & Stein, B. (2017). A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*.
- [14] Regacho, L. C. C. H., Matsushita, A., & Ceniza-Canillo, A. M. (2023). Automated Identification of Disaster News For Crisis Management Using Machine Learning. *arXiv preprint arXiv:2301.09896*.
- [15] Sneha, N., & Gangil, T. (2019). Analysis of diabetes mellitus for early prediction using optimal features selection. *Journal of Big data*, 6(1), 1-19.
- [16] Karim, A. A. J., Asad, K. H. M., & Azam, A. (2024). Strengthening Fake News Detection: Leveraging SVM and Sophisticated Text Vectorization Techniques. Defying BERT?. *arXiv preprint arXiv:2411.12703*.
- [17] Taffese, W. Z., & Espinosa-Leal, L. (2022). Prediction of chloride resistance level of concrete using machine learning for durability and service life assessment of building structures. *Journal of Building Engineering*, 60, 105146.
- [18] Almarashy, A. H. J., Feizi-Derakhshi, M. R., & Salehpour, P. (2023). Enhancing fake news detection by multi-feature classification. *IEEE Access*, 11, 139601-139613.
- [19] Roumeliotis, K. I., Tselikas, N. D., & Nasiopoulos, D. K. (2025). Fake News Detection and Classification: A Comparative Study of Convolutional Neural Networks, Large Language Models, and Natural Language Processing Models. *Future Internet*, 17(1).

- [20] Utku, A. (2024). Hybrid CNN-LSTM Model for Fake News Detection. *NATURENGS*, 5(2), 28-36.
- [21] Tian, Y., Xu, S., Cao, Y., Wang, Z., & Wei, Z. (2025). An Empirical Comparison of Machine Learning and Deep Learning Models for Automated Fake News Detection. *Mathematics*, 13(13), 2086.
- [22] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171-4186).
- [23] Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5), 1-40.
- [24] Owusu-Adjei, M., Ben Hayfron-Acquah, J., Frimpong, T., & Abdul-Salaam, G. (2023). Imbalanced class distribution and performance evaluation metrics: A systematic review of prediction accuracy for determining model performance in healthcare systems. *PLOS Digital Health*, 2(11), e0000290.
- [24] Hansrajh, A., Adeliyi, T. T., & Wing, J. (2021). Detection of online fake news using blending ensemble learning. *Scientific Programming*, 2021, 1–14.
- [26] Alghamdi, J., Lin, Y., & Luo, S. (2022). A Comparative Study of Machine Learning and Deep Learning Techniques for Fake News Detection. *Information*, 13(12), 576.
- [27] Kuntur, S., Krzywda, M., Wróblewska, A., Paprzycki, M., & Ganzha, M. (2024). Comparative Analysis of Graph Neural Networks and Transformers for Robust Fake News Detection: A Verification and Reimplementation Study. *Electronics*, 13(23), 4784.
- [28] Aslam, N., Khan, I. U., Alotaibi, F. S., Aldaej, L. A., & Aldubaikil, A. K. (2021). Fake Detect: A deep learning ensemble model for fake news detection. *Complexity*, 2021, Article 5557784.

- [29] Singhania, D., Fernandez, N., & Rao, S. (2023). 3HAN: A deep neural network for fake news detection. arXiv preprint arXiv:2306.12014.
- [30] Ajik, E. D., Obuandike, G. N., & Echobu, F. O. (2023). Fake news detection using optimized CNN and LSTM techniques. *Journal of Information Systems and Informatics*, 5(3), 1044–1056.
- [31] Shaker, K., & Alqudsi, A. (2024). Approach for detecting Arabic fake news using deep learning. *Iraqi Journal for Computer Science and Mathematics*, 5(3), 779–789.
- [32] Huang, T., Xu, Z., Yu, P., Yi, J., & Xu, X. (2025). A hybrid transformer model for fake news detection: Leveraging Bayesian optimization and bidirectional recurrent unit. arXiv preprint arXiv:2502.09097.
- [33] Karthik, G. M., Ahmad, K. S. F., Pamidimukkala, S. G., Sathe, A. P., Sirisha, G. N. V. G., Ram, S. M., & Koteswararao, C. (2025). Hybrid optimization driven fake news detection using reinforced transformer models. *Scientific Reports*, 15(1), 14782.
- [34] <https://www.kaggle.com/datasets/rahulogoel/isot-fake-news-dataset> Last accessed 14/12/2025