الجمهورية الجزائرية الديمقراطية الشعبية
People's Democratic Republic of Algeria
وزارة التعليم العالي و البحث العلمي
Ministry of Higher Education and Scientific Research
جامعة غرداية
University of Ghardaia
كلية العلوم والتكنولوجيا
Faculty of Science and Technology
قسم الرياضيات والاعلام الي
Department of Mathematics and Computer Science

Registration n°:
...../...../...../...../.....

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master

Domain: Mathematics and Computer
Field: Computer Science
Specialty: Intelligent Systems for Knowledge Extraction

Topic

AI Powered Fake News Classification through Advanced NLP Techniques

---

Presented by

Bouharkat Lina Djihane & *Mochten Ayat Elrrahmane*

Publicly defended on 06 21, 2026

Jury members:

Dr.Ahmed Saidi	MCA	Univ.Ghardaia	President
Dr.Mohamed El Amine Fekair	MCB	Univ.Ghardaia	Examiner
Dr.Messaoud Benguenane	MCB	Univ.Ghardaia	Supervisor

Academic Year: 2025/2026

Acknowledgment

First and foremost, we thank Almighty God for guiding our path and giving us strength for this work.

We extend our deepest appreciation to our supervisor, **Dr.Messaoud Benguenane**, for his unwavering support, patience, and invaluable guidance throughout the research and writing of this thesis.

We also want to extend our sincere appreciation & thanks to **Ms. Kaoutar Zita** for her help and guidance. May Allah bless you and make your wishes true.

We are also pleased to thank everyone who advised, guided, or contributed to us in preparing this research.

Dedication

To my beloved father, **Omar**, who exhausted himself and did the impossible for our happiness and success.
To my dear mother, **Naima**, who has always been my constant support and my greatest source of inspiration.
To my siblings, the stars of my sky:
To my beautiful sister **Hadil**, whom I am deeply proud of.
To my little princess before being my sister, my beloved **Wissal**.
To my dear brother, **Rayane**.
To all those who have been my companions on this journey, my support, my strength, my comfort, my joy, and my peace my dear friends **Meriam**, **Wissal**, **Doha**, and **Hayam**.
To my project partner and dear friend, **Ayat Elrrahmane**, whose companionship made this journey truly enjoyable. From the very beginning, our meeting was marked by remarkable harmony, mutual understanding, and shared determination. It was a pleasure and a privilege to share this experience with you, and I am sincerely grateful for your dedication, support, and friendship.
To **Kaoutar**, whose kindness and support carried me through the most difficult times. I will forever be grateful for your efforts,time and kindness.
To my dear students, who filled this year with joy and unforgettable moments. Your curiosity, energy, and smiles brought warmth and happiness to every day. Thank you for the wonderful memories and for the inspiration you have given me throughout this year.
I dedicate this humble work to all of you. By the grace and will of Allah, and thanks to your love, encouragement, and unwavering support, I am who I am today.

Lina Djihane

Dedication

To my beloved parents, **Slimane and Zohra**, whose love, sacrifices, and unwavering support made this journey possible. May Allah bless and reward you abundantly.

To my dear brothers, **Mouad and Anas**, for your constant support, kindness, and the joy you bring to my life. May Allah grant you success and happiness in all your endeavors.

To my beloved sister **Alaa**, the sunshine of our home and one of Allah's greatest blessings in my life. May your future be as beautiful as your heart.

To my beloved **grandmothers**, whose prayers have been a source of comfort, strength, and success. May Allah bless you with health, happiness, and long lives.

To my dear friends, **Ikram, Amina, and Maria**, thank you for your friendship, support, and the beautiful memories we shared. Distance may separate us, but you will always remain close to my heart.

To my partner and dear friend **Lina Djihane**, thank you for your support, dedication, and companionship throughout this journey. Your kindness and encouragement made this experience truly special.

To the **person** whose positive influence on this journey is greater than they may ever realize.

To **everyone** who offered support, encouragement, sincere prayers, or a helping hand along the way.

I dedicate this humble work to all of you. By the grace of Allah and through your love and support, I have reached this milestone. May Allah reward you all abundantly.

Ayat Elrrahmane

الملخص

أصبحت الأخبار الزائفة تمثل تحديًا كبيرًا في العصر الرقمي نتيجة الانتشار السريع للمعلومات المضللة عبر المنصات الإلكترونية ووسائل التواصل الاجتماعي، وما لذلك من تأثيرات مهمة على الرأي العام واتخاذ القرار وثقة المجتمع في مصادر المعلومات. تهدف هذه الدراسة إلى تحليل فعالية نماذج التعلم العميق المعتمدة على المحولات (Transformer) في الكشف عن الأخبار الزائفة، وذلك انطلاقًا من فرضية أن النماذج اللغوية المدربة مسبقًا قادرة على تعلم تمثيلات سياقية ودلالية غنية للنصوص، مما يسمح بتصنيف الأخبار بدقة إلى فئتين: حقيقية أو زائفة.

يتمثل الهدف الرئيسي في تصميم وتقييم إطار عمل قوي ومقارن يعتمد على أحدث معماريات المحولات، بما في ذلك BERT وRoBERTa وDeBERTa-v3 ونموذج تجميعي هجين، مع اختبار أدائها على مجموعتي بيانات معياريتين هما ISOT Fake News وWELFake. تعتمد المنهجية على ضبط (Fine-tuning) النماذج المدربة مسبقًا على بيانات معنونة، بعد تطبيق سلسلة موحدة من خطوات المعالجة المسبقة تشمل تنظيف البيانات، والترميز (Tokenization)، وتوحيد تسلسل النصوص، وترميز التصنيفات، ثم تقييم الأداء باستخدام مقاييس معيارية مثل الدقة (Accuracy)، ومقياس F1، ومساحة تحت منحنى ROC، ومصفوفة الالتباس.

تُظهر النتائج التجريبية أن النماذج المعتمدة على المحولات تحقق أداءً قويًا ومنافسًا في مهمة الكشف عن الأخبار الزائفة، حيث حقق نموذجRoBERTa أفضل النتائج على مجموعة بيانات ،WELFake بينما قدم نموذج BERT أكثر أداء مستقرًا على مجموعة بيانات ISOT محرزا أعلى نتيجة بـ F1-score 67%.95، في حين أظهر كل من DeBERTa-v3 والنموذج التجميعي الهجين أداءً أقل نسبيًا، مما يبرز تأثر فعالية النماذج بخصائص البيانات وإعدادات التدريب. بشكل عام، تؤكد النتائج فعالية الأساليب المعتمدة على المحولات في الكشف عن المعلومات المضللة، وتسلط الضوء على أهمية اختيار النموذج المناسب حسب طبيعة البيانات، مع إمكانية تطبيقها في أنظمة التحقق الآلي من الأخبار ومراقبة المحتوى على وسائل التواصل الاجتماعي.

الكلمات المفتاحية: كشف الأخبار الزائفة، معالجة اللغة الطبيعية، نماذج المحولات، BERT، RoBERTa، DeBERTa، المعلومات المضللة.

Abstract

Fake news has emerged as a major challenge in the digital era due to the rapid dissemination of misinformation across online platforms and social media, with significant implications for public opinion, decision-making, and societal trust in information sources. This study investigates the effectiveness of transformer-based deep learning models for fake news detection under the hypothesis that pre-trained language models can capture rich contextual and semantic representations of textual data, enabling accurate binary classification of news articles as real or fake. The main objective is to design and evaluate a robust and comparative framework based on state-of-the-art transformer architectures, including BERT, RoBERTa, DeBERTa-v3, and a hybrid ensemble model, and to assess their performance across two benchmark datasets, ISOT Fake News and WELFake. The adopted methodology consists of fine-tuning pre-trained models on labeled datasets following a unified preprocessing pipeline involving data cleaning, tokenization, sequence standardization, and label encoding, followed by evaluation using standard metrics such as accuracy, F1-score, ROC-AUC, and confusion matrix. The experimental results demonstrate that transformer-based models achieve strong and competitive performance in fake news detection, with RoBERTa achieving the best results on the WELFake dataset and BERT showing the most consistent performance on the ISOT dataset, achieving the best result with 95.67% F1-score, while DeBERTa-v3 and the hybrid ensemble model exhibit comparatively lower performance, highlighting the sensitivity of model effectiveness to dataset characteristics and training configurations. Overall, the findings confirm the effectiveness of transformer-based approaches for misinformation detection and underline the importance of model selection depending on dataset properties, with potential applications in automated fact-checking and social media monitoring systems.

Keywords: Fake News Detection, Natural Language Processing, Transformer Models, BERT, RoBERTa, DeBERTa, Misinformation.

Résumé

Les fausses nouvelles sont devenues un défi majeur à l'ère numérique en raison de la diffusion rapide de la désinformation à travers les plateformes en ligne et les réseaux sociaux, avec des conséquences importantes sur l'opinion publique, la prise de décision et la confiance sociétale envers les sources d'information. Cette étude examine l'efficacité des modèles d'apprentissage profond basés sur les transformeurs pour la détection des fausses nouvelles, en partant de l'hypothèse selon laquelle les modèles de langage pré-entraînés peuvent capturer des représentations contextuelles et sémantiques riches des données textuelles, permettant ainsi une classification binaire précise des articles de presse en vrais ou faux. L'objectif principal est de concevoir et d'évaluer un cadre robuste et comparatif basé sur des architectures de pointe de type transformeur, notamment BERT, RoBERTa, DeBERTa-v3 et un modèle d'ensemble hybride, et d'analyser leurs performances sur deux jeux de données de référence, ISOT Fake News et WELFake. La méthodologie adoptée consiste à affiner des modèles pré-entraînés sur des ensembles de données étiquetées en suivant un pipeline de prétraitement unifié comprenant le nettoyage des données, la tokenisation, la standardisation des séquences et l'encodage des étiquettes, suivi d'une évaluation à l'aide de métriques standard telles que l'exactitude , le score F1, l'AUC-ROC et la matrice de confusion. Les résultats expérimentaux montrent que les modèles basés sur les transformeurs obtiennent des performances élevées et compétitives dans la détection des fausses nouvelles, RoBERTa atteignant les meilleurs résultats sur le jeu de données WELFake et BERT présentant les performances les plus stables sur le jeu de données ISOT en obtenant le meilleur score F1 de 95,67%, tandis que DeBERTa-v3 et le modèle d'ensemble hybride affichent des performances relativement plus faibles, mettant en évidence la sensibilité de l'efficacité des modèles aux caractéristiques des jeux de données et aux configurations d'entraînement. Dans l'ensemble, les résultats confirment l'efficacité des approches basées sur les transformeurs pour la détection de la désinformation et soulignent l'importance du choix du modèle en fonction des propriétés des données,

avec des applications potentielles dans les systèmes automatisés de vérification des faits et de surveillance des réseaux sociaux.

Mots-clés : Détection de fausses nouvelles, traitement automatique du langage naturel (TALN), modèles transformeurs, BERT, RoBERTa, DeBERTa, désinformation.

Contents

List of Figures

List of Tables

List of Acronyms

FND	Fake News Detection
AI	Artificial Intelligence
ML	Machine Learning
DL	Deep Learning
RNN	Recurrent Neural Network
CNN	Convolutional Neural Network
LR	Logistic Regression
NB	Naive Bayes
ReLU	Rectified Linear Unit
tanh	Hyperbolic Tangent Function
LSTM	Long Short-Term Memory

Introduction

The rapid growth of the Internet, online news platforms, and social media has fundamentally transformed the way information is produced, distributed, and consumed. Digital technologies have enabled individuals to access vast amounts of information instantly and participate actively in content creation and dissemination. While these advancements have significantly improved global communication and information accessibility, they have also facilitated the widespread propagation of false, misleading, and manipulated information, commonly referred to as fake news [38].

Fake news has emerged as a major societal challenge with significant consequences for public opinion, political processes, public health, economic stability, and social cohesion. The ability of misleading information to spread rapidly through social networks often exceeds the capacity of traditional fact-checking mechanisms, making manual verification increasingly difficult and time-consuming. As a result, the development of automated and scalable fake news detection systems has become a critical research area within Artificial Intelligence (AI) and Natural Language Processing (NLP).

Early fake news detection approaches primarily relied on traditional machine learning techniques combined with handcrafted textual features. Although these methods achieved promising results, their effectiveness was often limited by their inability to capture complex semantic relationships and contextual information present in natural language. The emergence of deep learning techniques significantly improved detection capabilities by enabling automatic feature extraction from textual data. More recently, Transformer-based architectures such as BERT, RoBERTa, and DeBERTa have demonstrated remarkable success across a wide range of NLP tasks due to their ability to learn rich contextual representations and model long-range dependencies within text.

Motivated by these advances, this thesis investigates the effectiveness of Transformer-based models for automatic fake news detection. Several state-of-the-art architectures, including BERT, RoBERTa, DeBERTa-v3, and a hybrid BERT–RoBERTa model, are implemented and evaluated on publicly available benchmark datasets. The objective is

to analyze their performance and compare their strengths and limitations. To achieve this objective, the thesis is organized as follows:

- **Chapter 1** introduces the fundamental concepts related to fake news, fake news detection, Artificial Intelligence, Machine Learning, Deep Learning, Transformer architectures, and Natural Language Processing.
- **Chapter 2** presents a comprehensive review of existing literature on fake news detection, covering traditional machine learning methods, deep learning approaches, and recent Transformer-based models.
- **Chapter 3** describes the proposed methodology, including the system architecture, datasets, preprocessing pipeline, model implementation, experimental setup, and evaluation metrics.
- **Chapter 4** presents the experimental results and comparative analysis of the implemented models, followed by a discussion of their performance and effectiveness.
- Finally, the thesis concludes with a summary of the main findings, contributions, limitations, and potential directions for future research.

Through this work, we aim to contribute to the ongoing effort to combat misinformation by evaluating advanced Transformer-based approaches capable of accurately and efficiently detecting fake news in digital environments.

Chapter 1

Background

1.1 Introduction

The rapid expansion of digital communication technologies, online news sources, and social media has revolutionized how information is created, shared, and consumed. While these tools offer incredible access to information, they also enable widespread proliferation of fake news and misinformation. The faster, larger volume of false information spreading has raised serious concerns about its effects on individuals, organizations, political systems, and society as a whole.

Consequently, fake news detection has become a vital research area within Artificial Intelligence (AI) and Natural Language Processing (NLP). Developing automated detection systems is crucial for efficiently identifying misleading content and promoting the circulation of trustworthy information online.

This chapter introduces key concepts related to fake news and its detection. It starts with an overview of fake news, including its definition, main categories, impacts, significance, and detection challenges. The chapter then introduces Artificial Intelligence and its main paradigms, with particular emphasis on machine learning, deep learning, and transformer-based approaches used for FND. Finally, it discusses the role of Natural Language Processing, covering essential text preprocessing methods and text representation techniques that enable computational models to effectively analyze and classify textual information.

1.2 Overview of Fake News

Fake news refers to false or misleading information presented as legitimate news, either intentionally or unintentionally, with the purpose of deceiving audiences or influencing their perception of reality. It often imitates the style and structure of credible journalistic content and is widely disseminated through social media platforms and websites. The rapid diffusion of such content significantly affects public

opinion, emotional responses, and decision-making processes.

Fake news can be categorized into several forms based on its intent, structure, and mode of dissemination, as illustrated in Figure 1.1. These categories represent the most common manifestations of misinformation in digital and traditional media environments, and understanding them is essential for identifying and analyzing their impact on society [64, 65].

- **Fabricated News:** Completely false stories created without any factual basis. These are often intentionally produced to mislead audiences, generate political influence, or gain financial profit through online engagement.
- **Propaganda:** Information designed to promote a political or ideological agenda. It typically appeals to emotions and selectively presents facts to influence public perception.
- **Conspiracy Theories:** Interpretations of events as secret plots orchestrated by powerful groups without sufficient evidence. Such narratives often reject official explanations, leading to reduced trust in institutions.
- **Clickbait:** Sensationalized or misleading headlines designed to attract user attention and increase web traffic. In many cases, the headline exaggerates or distorts the actual content.

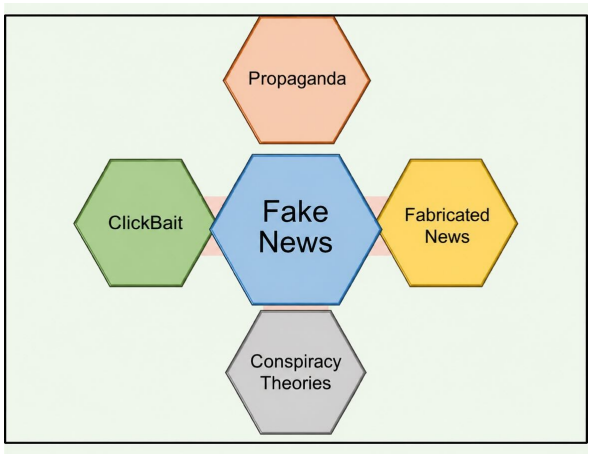


Figure 1.1: Types of Fake News [21].

1.2.1 Fake News Impact

The rapid expansion of digital media and social networking platforms has significantly amplified the dissemination and influence of fake news. Unlike traditional misinformation channels, online environments enable false information to reach large audiences within a short period of time, increasing its potential impact across multiple domains. The consequences of fake news extend beyond individual users and affect economic activities, political systems, and social cohesion [38].

- **Individual Impact:** Fake news can adversely affect individuals by spreading inaccurate information that damages personal and professional reputations, influences decision-making, and creates confusion regarding important issues. In sensitive domains such as healthcare, misinformation may lead individuals to adopt harmful behaviors or make unsafe decisions.
- **Economic Impact:** The dissemination of fake news can negatively influence economic activities by shaping consumer perceptions, damaging corporate reputations, and creating uncertainty in financial markets. Misleading information may also affect investment decisions and business performance.
- **Political and Democratic Impact:** Fake news can manipulate public opinion, influence electoral outcomes, and undermine trust in political institutions. By distorting public discourse and spreading misleading narratives, it poses significant challenges to democratic processes and informed civic participation.
- **Societal Impact:** The widespread circulation of misinformation contributes to social polarization and reinforces divisions among individuals and communities. It can erode public confidence in news media, governmental institutions, and other trusted sources of information, thereby weakening social cohesion.
- **Technological Amplification:** Digital platforms and social media networks facilitate the rapid dissemination of fake news through instant sharing mechanisms, recommendation algorithms, and personalized content delivery systems.

These technologies often create echo chambers that reinforce existing beliefs and increase users' exposure to misleading information.

1.2.2 Fake news detection

The rapid growth of online news platforms and social media has transformed the way information is produced, shared, and consumed. While these technologies facilitate access to information, they also enable the widespread dissemination of false or misleading content. Fake news often resembles legitimate news articles in terms of writing style, structure, and presentation, making it difficult for readers to distinguish between credible and deceptive information [34]. The problem of FND consists of automatically determining whether a news item is genuine or misleading based on the analysis of its content and related contextual information. This task is particularly challenging because misleading news creators continuously adapt their strategies to mimic trustworthy sources and exploit current events, public interests, and emotional narratives. Furthermore, the high volume, velocity, and diversity of online information make manual verification impractical. From a machine learning perspective, FND can be formulated as a binary classification problem in which a news article is assigned to one of two classes: *fake* or *real*. More advanced approaches may also consider multiclass settings that distinguish between different forms of misinformation, such as propaganda, satire, rumors, and fabricated content. Effective FND systems must therefore identify subtle linguistic patterns, semantic inconsistencies, stylistic characteristics, and contextual signals that differentiate misleading information from legitimate news. The development of accurate and robust detection models has become an essential research challenge for preserving information integrity and supporting informed decision-making in digital environments.

1.2.3 Importance of Fake News Detection

Fake News Detection (FND) has become a critical research area due to the rapid growth of online news consumption and social media platforms, which enable information to spread rapidly and reach a wide audience. The increasing circulation of false or misleading information across digital platforms can negatively affect individuals, organizations, political systems, and society as a whole, while also undermining public trust in the media. Therefore, effective fake news detection techniques are essential for identifying deceptive content, limiting its spread, preserving information credibility, and supporting informed decision-making. Based on the reviewed literature, the importance of Fake News Detection can be summarized as follows [4]:

- **Preventing the Spread of Misinformation:** Fake news can be rapidly disseminated through social media platforms, where users can easily share content without verification. Detecting fake news helps prevent the widespread circulation of false information before it reaches a large audience.
- **Protecting Individuals and Organizations:** Fake news is often used to damage the reputation of individuals, businesses, organizations, or political parties. Effective detection systems help identify and stop such harmful content, reducing reputational and social damage.
- **Reducing the Influence of False Information on Public Opinion:** Fake news can manipulate people's beliefs, opinions, and decisions. Detecting false information helps ensure that people make decisions based on accurate facts rather than misinformation.
- **Preventing Social and Political Harm:** The spread of fake news can negatively impact societies and political systems by creating confusion, division, and misinformation. Detecting fake news helps maintain social stability and protects democratic processes from manipulation.

- **Addressing the Limitations of Human Verification:** Manually identifying fake news is challenging because people often lack complete knowledge of a topic and may believe information without verification. Automated fake news detection systems can assist humans by efficiently analyzing large volumes of online content.

- **Controlling the Rapid Spread of Fake News:** Due to the sharing features of social media platforms, fake news can spread extremely quickly. Early detection is essential to stop misinformation before it reaches and influences a large number of users.

- **Enabling Automatic Detection through Machine Learning:** Machine learning algorithms can automatically analyze news content and classify it as real or fake. This automation makes fake news detection more scalable and effective than relying solely on human fact-checkers.

1.2.4 Challenges of Fake News Detection

FND is a complex and challenging task due to the dynamic and multifaceted nature of misinformation. False information can appear in various forms and is often designed to resemble legitimate news, making it difficult to distinguish. The widespread use of online news platforms and social media has further accelerated the creation and dissemination of misleading information, allowing it to reach large audiences within a short period of time. In addition, fake news continuously evolves in response to current events and public interests, increasing the difficulty of identifying and verifying its authenticity. As a result, several challenges must be addressed to effectively detect and combat fake news. Figure 1.2 shows the Most faced challenges.

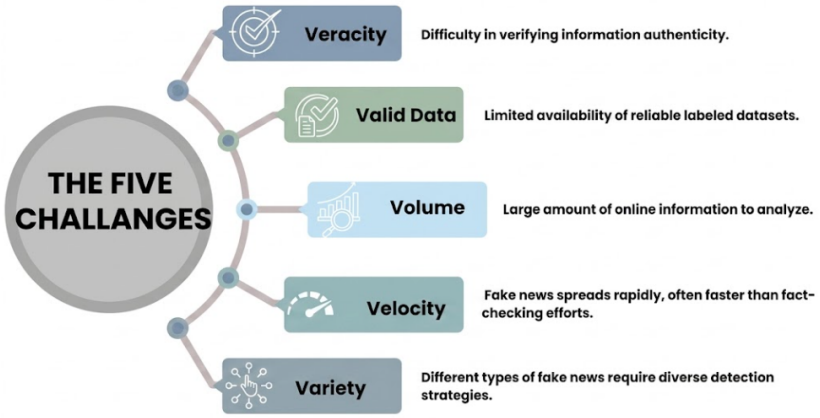


Figure 1.2: Fake News Detection Challenges.

- **Variety** Fake news exists in multiple forms, including propaganda, rumors, hoaxes, satire, and deepfakes. The diversity of these deceptive content types and their different objectives makes it difficult to identify and classify accurately [66].

- **Volume** The widespread use of social media and online platforms enables the rapid creation and dissemination of large amounts of fake news. The sheer volume of information shared daily makes monitoring and verifying content a significant challenge [66].

- **Velocity** Fake news spreads extremely quickly, especially during major events and breaking news situations. Its rapid propagation often outpaces verification efforts, making timely detection and mitigation difficult [66].

- **Veracity** Determining the truthfulness of news content is challenging because fake news is intentionally designed to appear credible. Assessing the authenticity of information often requires external verification beyond the content itself [59].

- **Valid Data** The development of reliable FND systems depends on the availability of high-quality datasets. However, obtaining comprehensive and accurately labeled data remains difficult due to data access restrictions and the evolving nature of misinformation [37].

Due to the several challenges of its detection faced over time, traditional fact-checking methods are no longer sufficient to manage this phenomenon effectively. Artificial Intelligence (AI) and Natural Language Processing (NLP) have become a powerful solution for automated detection of fake news in this context [53].

1.3 Artificial Intelligence (AI)

Artificial Intelligence (AI) refers to systems capable of performing tasks associated with human intelligence, such as learning, reasoning, problem-solving, and decision-making [49].

The primary objective of AI research is to design intelligent systems that can perceive their environment, learn from experience, reason effectively, and select optimal actions to achieve specific goals. Such systems are expected to adapt to new environments and perform cognitive tasks with increasing autonomy and efficiency. AI forms a hierarchical structure with Machine Learning (ML), and Deep Learning (DL). AI is the broadest domain, encompassing techniques that enable machines to perform tasks requiring human intelligence. Within AI, ML focuses on algorithms that learn patterns from data and improve their performance without explicit programming. DL is a specialized subset of ML that utilizes multilayer neural networks to learn complex representations from large-scale data. Figure 1.3 illustrates the hierarchical relationship between these three domains.

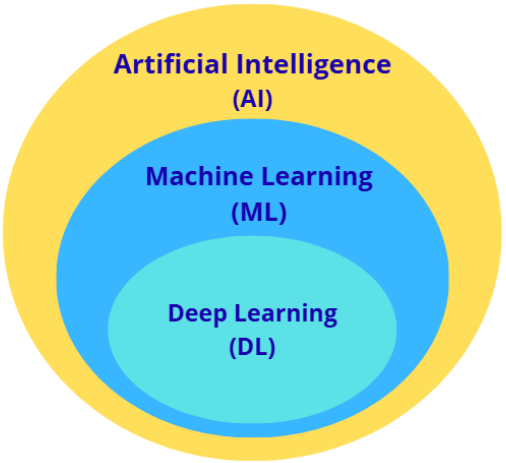


Figure 1.3: Relationship between Artificial Intelligence, Machine Learning, and Deep Learning

Over the years, various AI approaches have been proposed, ranging from traditional Machine Learning (ML) techniques to more advanced Deep Learning (DL) models and Transformer-based architectures. Each category offers unique capabilities and has contributed significantly to improving the accuracy and effectiveness of FND systems. This section provides an overview of these major AI-based approaches.

1.3.1 Machine Learning

Machine learning is a subset of artificial intelligence that enables systems to learn patterns from data and improve performance without being explicitly programmed. By detecting patterns, relationships, and trends in data, algorithms can forecast outcomes, categorize information, and assist in decision-making. These algorithms are trained on labeled or unlabeled datasets, enabling them to glean meaningful insights and generalize to new, unseen data [14].

ML is generally divided into three main paradigms: supervised, unsupervised, and semi-supervised learning. These differ based on the type and amount of labeled data involved during training. The suitability of each approach depends on the specific problem and the data available [20]. The main types of machine learning include:

Supervised learning : is an ML approach in which models are trained on labeled datasets of input-output pairs. This allows the models to learn the relationship between features and targets and make predictions on new, unseen data. It is generally categorized into two types: classification and regression. Classification involves predicting categorical outcomes by assigning data to predefined classes, such as detecting spam emails or diagnosing health conditions. Common algorithms include K-Nearest Neighbors (K-NN) [25], which classifies data based on the majority class among its nearest neighbors in the feature space, and Naïve Bayes (NB) [1], a probabilistic method based on Bayes' theorem that assumes feature independence and is effective for high-dimensional data. Regression, on the other hand, predicts continuous numerical values, such as house prices, sales, or stock trends, by modeling the relationship between inputs and a continuous target. Popular regression algorithms include Decision Trees, valued for their interpretability and for depicting decision processes through a hierarchical structure in which nodes represent features, branches contain decision rules, and leaves indicate predictions. Decision Trees also underpin ensemble methods like Random Forests, which enhance accuracy by aggregating multiple trees trained

on different subsets of the data [10].

Unsupervised learning: is an ML technique that detects hidden patterns, structures, and relationships in unlabeled data without relying on predefined target labels. It is commonly employed for data exploration, visualization, feature extraction, and dimensionality reduction. The primary categories of unsupervised learning are clustering and association. Clustering groups similar data points into meaningful clusters, helping to uncover natural structures in datasets. K-Means is a widely used clustering technique that partitions data into a fixed number of clusters by assigning points to their nearest centroid and updating the centroids to minimize within-cluster variance. Although it is computationally efficient, K-Means requires specifying the number of clusters beforehand and can perform poorly with irregularly shaped clusters or varying data densities. Another popular method is Mean Shift, a non-parametric, density-based algorithm that finds clusters by moving data points toward regions of higher density until reaching local maxima. Unlike K-Means, Mean Shift does not require specifying the number of clusters in advance, it automatically determines the number from the data, making it suitable for datasets with complex structures and unclear boundaries [42].

Semi-Supervised Learning: is an ML approach that combines both labeled and unlabeled data during training, making it particularly useful when labeled data is scarce, costly, or time-consuming to obtain [67]. By leveraging the large amount of available unlabeled data alongside a small labeled dataset, semi-supervised learning can improve model performance and generalization. It is commonly applied in tasks such as sentiment analysis and text classification. Popular semi-supervised learning algorithms include **self-training**, where a model iteratively labels unlabeled data using its own predictions **co-training**, where multiple models collaboratively exchange predictions to enhance learning, and **multi-perspective learning**, which integrates different feature representations or viewpoints to better capture data characteristics [40] [15].

The diagram in Figure 1.4 illustrates how different learning approaches use data during model training: supervised learning uses fully labeled data, semi-supervised learning combines a small amount of labeled data with a large amount of unlabeled data, and unsupervised learning relies entirely on unlabeled data to train the model.

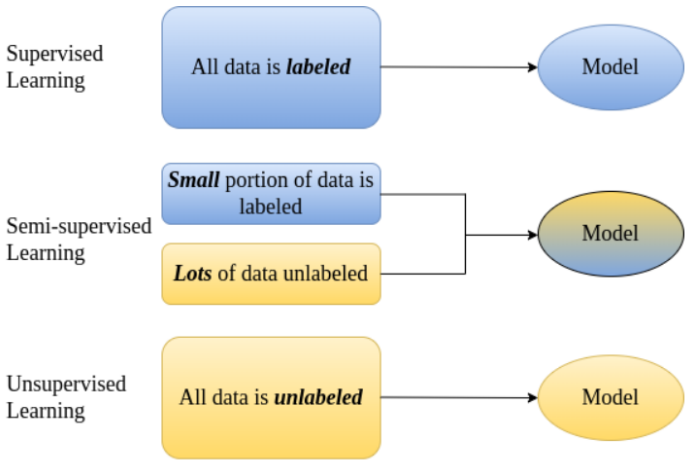


Figure 1.4: Machine Learning Paradigms.

1.3.2 Deep Learning

Deep learning is a branch of ML that allows models to learn multiple levels of representation and abstraction directly from raw data. This enables them to automatically identify meaningful features for tasks like detection and classification. Usually based on artificial neural networks with several hidden layers, it can model highly complex and nonlinear relationships between inputs and outputs. Deep learning techniques have achieved impressive results across numerous real-world applications, often outperforming traditional machine learning methods in prediction accuracy. It is commonly used in areas such as pattern recognition, natural language processing, image analysis, and time-series forecasting, where understanding complex structures from large datasets is crucial [9]. RNN, LSTM, and GRU are some of the most used architectures in NLP and text classification domain.

Recurrent Neural Network (RNN) RNNs are a class of neural networks specifically designed to process sequential data by maintaining information from previous inputs through recurrent connections. Unlike traditional feedforward neural networks, RNNs possess an internal memory that enables them to capture temporal dependencies and contextual information within a sequence. This characteristic makes them particularly suitable for tasks involving sequential data, such as language modeling, machine translation, speech recognition, and time-series forecasting [44]. The core component of an RNN is the hidden state, which is updated at each time step based on the current input and the previous hidden state. Formally, the hidden state h_t at time step t is computed as a nonlinear transformation of the input vector x_t and the preceding hidden state h_{t-1}:

$$h_t = f(W_{ih}x_t + W_{hh}h_{t-1} + b_h)[44] \tag{1.1}$$

where
W_{ih}, W_{hh} = trainable weight matrices,
b_h = bias vector,
f = nonlinear activation function, such as tanh or ReLU.

By continuously updating its hidden state throughout the sequence, the RNN is able to capture temporal dependencies and contextual information, making it effective for modeling sequential patterns in data.

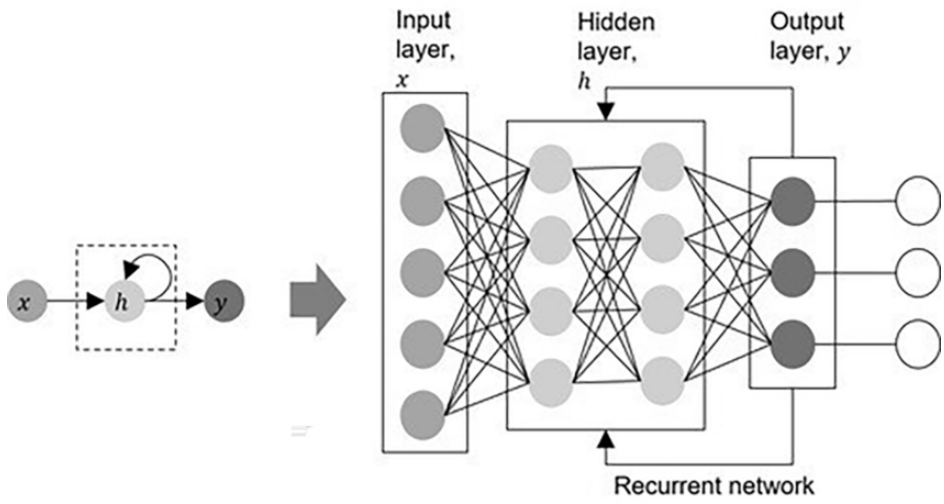


Figure 1.5: RNN Architecture [46].

Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) designed to capture long-term dependencies in sequential data through the use of a memory cell that stores information over extended periods [11]. The memory cell is updated at each time step based on the current input and the previous hidden state, with the update process regulated by three gating mechanisms: **the input gate, forget gate, and output gate**. The input gate controls how much new information is stored in the memory cell, the forget gate determines which information should be retained or discarded, and the output gate regulates how much information from the memory cell is exposed to the hidden state and network output. These mechanisms enable LSTM networks to selectively preserve relevant information while eliminating unnecessary data, making them effective for sequence modeling tasks [26].

The architecture of an LSTM unit, including its memory cell and gating mechanisms, is illustrated in Figure 1.6.

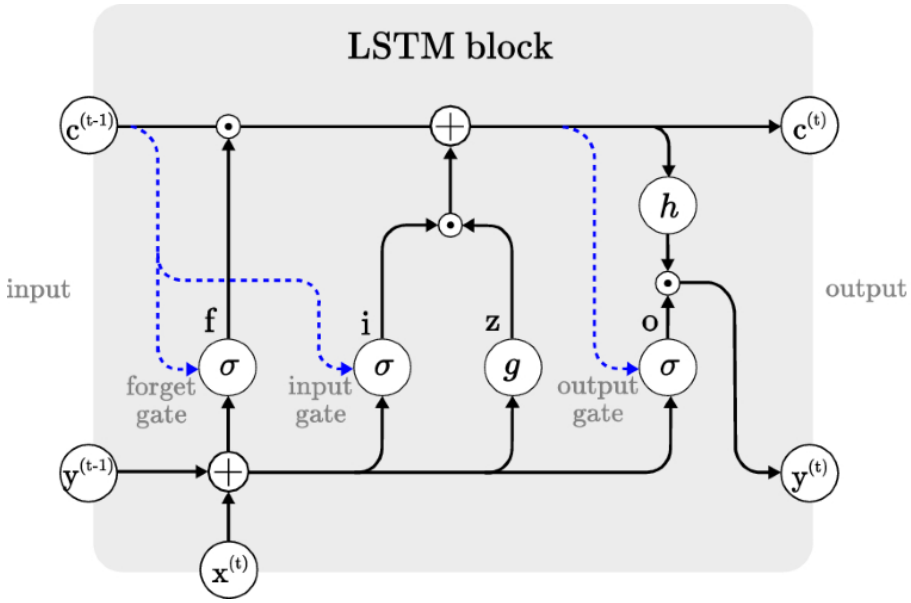


Figure 1.6: LSTM unit with input and output connections [61].

Gated Recurrent Unit (GRU) GRU is a recurrent neural network designed to model sequential data and capture long-term dependencies. Unlike LSTM, GRU does not use a separate memory cell; instead, it employs an update gate and a reset gate to control information flow through the hidden state. This simpler architecture reduces the number of parameters while maintaining effective sequence modeling performance [16]. The GRU computation proceeds through the following steps:

Update Gate regulates the balance between retaining past hidden information and incorporating new input, allowing the model to capture long-term dependencies through selective updates of the hidden state (Eq. (1.2)).

$$z_t = \sigma(W_z x_t + U_z h_{t-1}) \text{[16]} \tag{1.2}$$

where

W_z, U_z = trainable weight matrices,

x_t = input at time t,

h_{t-1} = previous hidden state,

σ = sigmoid activation function.

Reset Gate controls how much of the previous hidden state is ignored when computing the candidate hidden state, enabling the model to focus on the most relevant past information (Eq. (1.3)).

$$r_t = \sigma(W_r x_t + U_r h_{t-1})[16] \tag{1.3}$$

where

W_r, U_r = trainable weight matrices,

x_t = input at time t,

h_{t-1} = previous hidden state,

σ = sigmoid activation function.

Candidate Hidden State and Hidden State Update is computed by using the reset gate to filter past information (Eq. (1.4)), while the final hidden state is obtained by interpolating between the previous hidden state and the candidate state via the update gate (Eq. (1.5)).

$$\tilde{h}_t = \tanh\left(W_h x_t + U_h(r_t \odot h_{t-1})\right)[16] \tag{1.4}$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t[16] \tag{1.5}$$

where

W_h, U_h = trainable weight matrices,

r_t = reset gate,

z_t = update gate,

$\odot$ = element-wise multiplication,

$\tanh$ = hyperbolic tangent activation function.

The architecture of a GRU unit, including its update and reset gates, is illustrated in Figure 1.7.

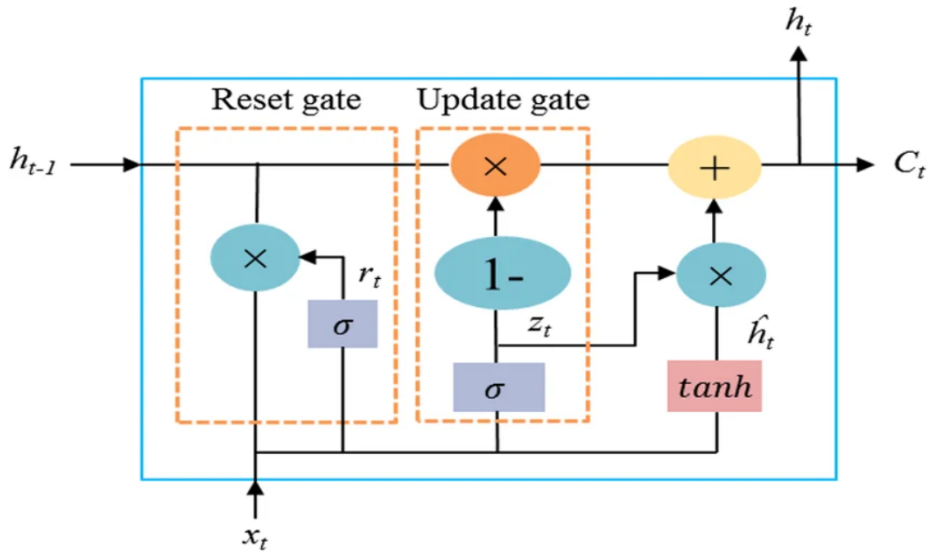


Figure 1.7: The Architecture of basic Gated Recurrent Unit (GRU) [45]

While RNNs, and their variants (LSTM, GRU...) have achieved impressive results in natural language processing, they remain constrained by their sequential processing and focus on local features. This makes it challenging for them to capture long-range dependencies efficiently. Recurrent models, in particular, process data step by step, which complicates modeling global context and increases computational costs for long sequences.

As a result, more advanced architectures have been developed, with the Transformer being a notable example that replaces recurrence with self-attention mechanisms to model relationships between all tokens in a sequence directly.

1.3.3 Transformers

The Transformer is a neural network architecture based on an encoder–decoder structure, as shown in Figure 1.8, that is composed of stacked layers of self-attention mechanisms and feed-forward neural networks [62].

The encoder consists of six identical layers, each containing a Multi-Head Self-Attention mechanism followed by a Position-wise Feed-Forward Network. Residual connections and layer normalization are applied around each sublayer to facilitate sta-

ble training. Similarly, the decoder is composed of six layers that include masked multi-head self-attention, encoder–decoder attention, and feed-forward networks. The decoder has a similar structure but introduces an additional encoder–decoder attention layer that enables the model to attend to the encoder outputs. Furthermore, masked self-attention is employed in the decoder to prevent access to future tokens during sequence generation. Through this architecture, the Transformer effectively captures both local and global dependencies while enabling highly parallelized computation, making it more efficient than traditional recurrent architectures. The effectiveness of the Transformer architecture has inspired the development of several widely used pre-trained language models, including BERT, RoBERTa, and DeBERTa. These models extend the original Transformer framework through various architectural modifications and training strategies.

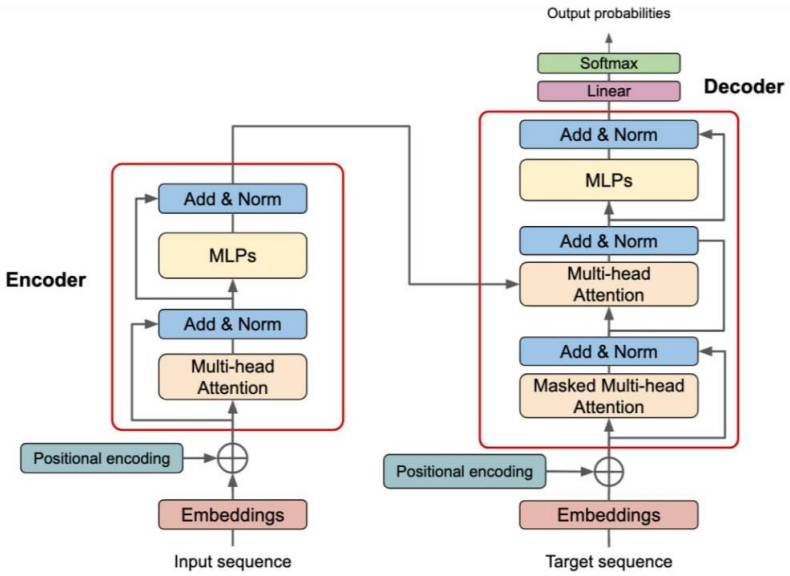


Figure 1.8: Transformer Architecture[62].

Self-Attention Mechanism

The fundamental operation of the Transformer architecture is the Scaled Dot-Product Attention mechanism, which enables each token in a sequence to attend to all other tokens and capture contextual dependencies regardless of their distance. Given the query matrix $\mathbf{Q}$, key matrix $\mathbf{K}$, and value matrix $\mathbf{V}$, the attention function is

defined as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V} \tag{1.6}$$

where d_k denotes the dimensionality of the key vectors. The scaling factor $\sqrt{d_k}$ is introduced to prevent excessively large dot-product values, which could lead to extremely small gradients after the softmax operation.

To enhance the model's ability to capture different types of relationships, the Transformer employs Multi-Head Attention. Instead of computing a single attention operation, multiple attention heads are executed in parallel, allowing the model to learn complementary representations from different subspaces of the input sequence.

Positional Encoding

Unlike recurrent and convolutional neural networks, the Transformer architecture does not inherently encode information about token order. To preserve sequential information, positional encodings are added to the input embeddings before they are processed by the model. By incorporating positional information into token embeddings, the Transformer can effectively model word order and sequential dependencies while retaining the benefits of parallel processing.

BERT Bidirectional Encoder Representations from Transformers) is a Transformer-based language model introduced by Devlin et al. It is designed to learn deep bidirectional contextual representations by jointly considering both left and right contexts during training.[18]. BERT is pre-trained using two self-supervised learning objectives. The first is Masked Language Modeling (MLM), in which 15% of the input tokens are randomly masked and the model learns to predict the missing words based on their surrounding context. The second objective is Next Sentence Prediction (NSP), where the model determines whether two sentences appear consecutively in the original text.

These pre-training tasks enable BERT to acquire rich linguistic knowledge from large unlabeled corpora.

After pre-training, BERT can be adapted to downstream tasks through fine-tuning. In this process, a task-specific classification layer is attached to the pre-trained encoder and the entire model is optimized on labeled data. For text classification tasks such as fake news detection, the contextual representation associated with the special `[CLS]` token is used as a sequence-level representation and fed to a classification layer to generate the final prediction. The architecture of BERT is illustrated in Figure 1.9.

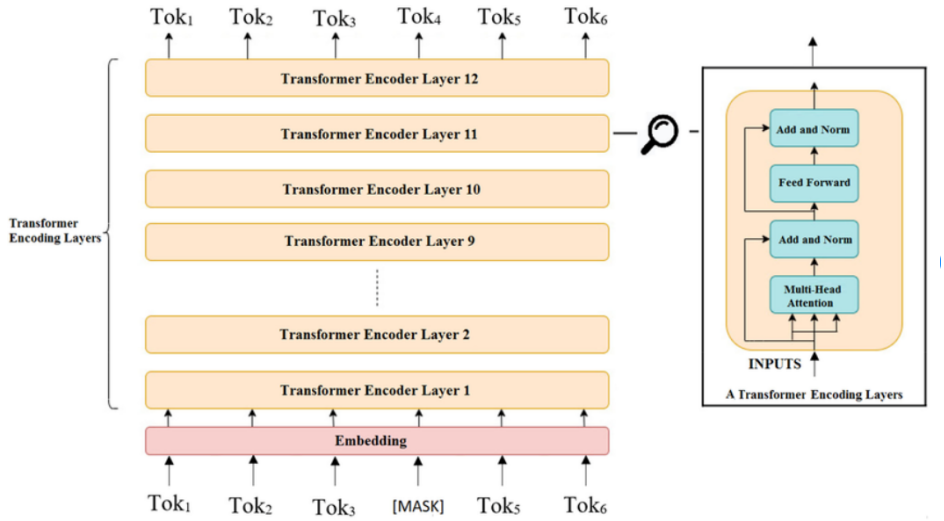


Figure 1.9: BERT Architecture[48].

RoBERTa (Robustly Optimized BERT Pretraining Approach) was introduced by Liu et al. [41] as an enhanced version of BERT. Rather than proposing a new architecture, RoBERTa improves the training strategy used during pre-training, resulting in stronger language representations and improved performance across a wide range of Natural Language Processing tasks.

RoBERTa introduces several important modifications to the original BERT framework. First, it employs *dynamic masking*, where masked tokens are regenerated at each training epoch instead of remaining fixed throughout the training process. This exposes the model to a larger variety of prediction tasks and improves its ability to learn

contextual representations.

Second, RoBERTa removes the Next Sentence Prediction (NSP) objective used in BERT. Experimental studies showed that NSP contributes little to downstream performance and may even hinder the learning process. Consequently, RoBERTa relies solely on the Masked Language Modeling (MLM) objective during pre-training.

Third, the model is trained on substantially larger corpora, totaling approximately 160 GB of text compared to the 16 GB used for BERT. In addition, RoBERTa is trained for more iterations using larger mini-batches, allowing it to better exploit the available training data. Figure 1.10 illustrates the architecture of the RoBERTa model. 1.10.

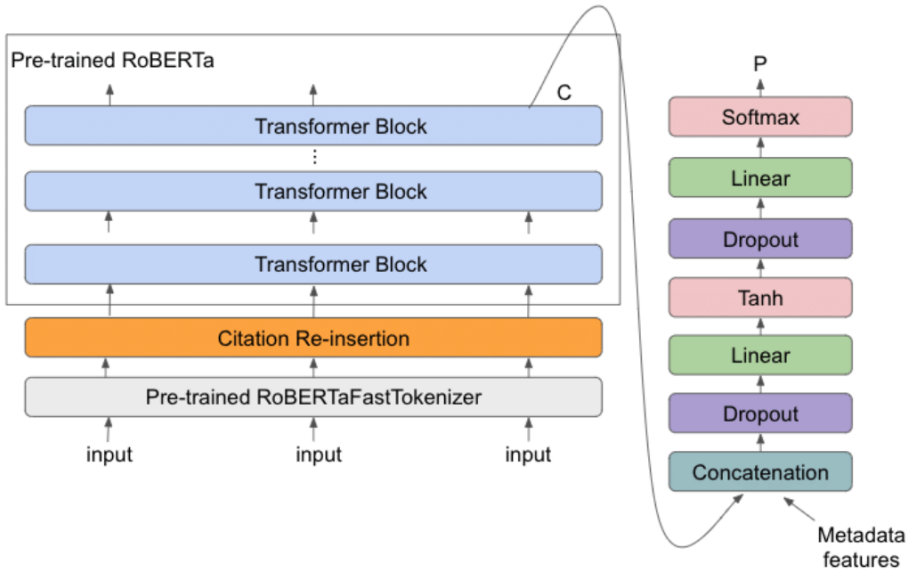


Figure 1.10: RoBERTa Architecture[28].

DeBERTa (Decoding-enhanced BERT with Disentangled Attention) is an enhanced Transformer-based language model that extends BERT and RoBERTa by introducing a disentangled attention mechanism [23], which combines content and positional information into a single representation. DeBERTa separates these two sources of information and processes them independently through a disentangled attention mechanism. In DeBERTa, each token is represented using both a content embedding and a position embedding. The attention score between two tokens is computed by considering interactions between content and relative positional information rather than relying solely

on content representations. The architecture of DeBERTa is illustrated in Figure 1.11.

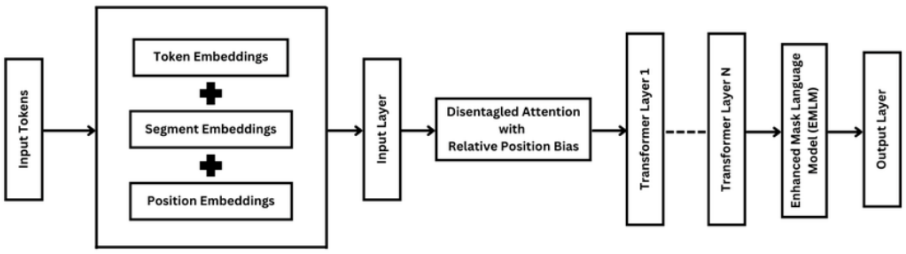


Figure 1.11: DeBERTa Architecture [2].

1.4 Natural Language Processing (NLP)

Natural Language Processing (NLP) is a field at the intersection of AI and computational linguistics that enables computers to analyze, understand, and generate human language. By transforming unstructured textual data into structured and meaningful representations, NLP supports a wide range of language-related tasks, including text classification, sentiment analysis, machine translation, and information extraction. In the context of FND, the rapid spread of misleading information across social media platforms has exposed the limitations of traditional fact-checking approaches, which rely on human experts such as journalists to verify claims against credible evidence. While effective, these manual methods are often time-consuming, costly, and difficult to scale. Automated fake news classification systems based on NLP techniques address these challenges by efficiently processing large volumes of textual content, reducing both time and resource requirements while improving the accuracy and effectiveness of identifying deceptive or fabricated information [24][54]. Although NLP can be applied to various forms of language data, including text and speech, this work focuses on textual data, which constitutes the primary source of information in FND. To effectively analyze and classify such data, several processing stages are required before it can be used by machine learning or deep learning models. These

stages mainly include text preprocessing, which cleans and normalizes raw text, and text representation, which converts textual information into numerical representations that can be efficiently processed by computational models. .

1.4.1 Text-preprocessing

Raw text is often noisy and unstructured (punctuation, capitalization, URLs, etc.). However, text pre-processing helps to make text easier to analyze and represent numerically. Text-preprocessing includes common steps such as Tokenization, Stop-word removal, Stemming, and Lemmatization [50].

- **Tokenization:** The initial step involves dividing raw text into basic units (tokens) such as words, sub-words, or characters.
- **Stop-word removal:** Removes high-frequency function words like "the" and "and" that have minimal semantic significance, thereby reducing noise for many models.
- **Stemming:** Reduces words to their root forms using heuristic rules (e.g., "running" → "run"), prioritizing speed over linguistic precision.
- **Lemmatization:** Maps words to their lemma based on morphological analysis (e.g., "better" → "good"), preserving part-of-speech information.

1.4.2 Text-representation techniques

Text representation is a fundamental component of NLP that aims to convert textual data into numerical representations that can be processed by machine learning and deep learning models. Since computers cannot directly understand human language. Over the years, various text representation techniques have been developed, as shown in the figure 1.12, ranging from traditional methods such as Bag-of-Words (BoW) and TF-IDF to distributed word embeddings like Word2Vec and GloVe, and

more recently, contextualized representations generated by transformer-based models such as BERT and RoBERTa. The effectiveness of an NLP system largely depends on the quality of its text representation, as it directly influences the model's ability to understand context, semantics, and relationships between words.

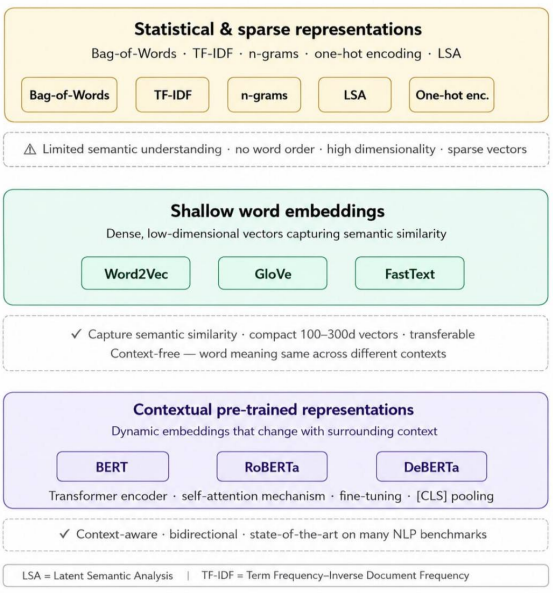


Figure 1.12: The historical development of text representation in NLP

1.4.2.1 Statistical & Spare (Classic) Representations

- **N-grams:** Represent text as contiguous sequences of n words or characters. N-grams capture limited local word order information and are widely used in traditional NLP and text classification tasks.
- **One-Hot Encoding:** Represents each word as a high-dimensional binary vector in which only one position is activated. While simple to implement, it produces sparse representations and cannot capture semantic relationships between words.
- **Bag-of-Words (BoW):** Represents a document as a sparse vector of word counts, ignoring word order[32][54].
- **Term Frequency-Inverse Document Frequency (TF-IDF):** TF–IDF is a widely used text representation technique that measures the importance of a term within a document relative to a collection of documents (corpus). It

assigns higher weights to terms that occur frequently in a specific document while reducing the influence of terms that appear in many documents[51][54].

The *Term Frequency (TF)* component quantifies how often a term appears in a document. It is calculated as shown in equation (1.7).

$$TF(t,d) = \frac{f_{t,d}}{\sum_k f_{k,d}} \tag{1.7}$$

where $f_{t,d}$ denotes the frequency of term t in document d.

The *Inverse Document Frequency (IDF)* component measures the rarity of a term across the corpus. It is calculated as shown in equation (1.8).

$$IDF(t) = \log\left(\frac{N}{df_t}\right) \tag{1.8}$$

where N is the total number of documents in the corpus and df_t is the number of documents containing the term t.

The TF–IDF weight of a term is obtained by multiplying its TF and IDF values, as it is calculated as shown in equation (1.9).

$$TF\text{-}IDF(t,d) = TF(t,d) \times IDF(t) \tag{1.9}$$

This representation highlights discriminative words and is commonly used in text classification and information retrieval tasks.

1.4.2.2 Word Embeddings

Dense, low-dimensional vectors that encode semantic and syntactic relationships between words from large corpora like Word2Vec and GloVe [32].

- **Word2Vec**

Word2Vec is a neural-network-based word embedding technique that learns dense vector representations of words from large text corpora. It captures semantic and syntactic relationships by placing words with similar contexts close to each other in the vector space. Word2Vec is commonly implemented using two architectures: Continuous Bag of Words (CBOW), which predicts a target word from its surrounding context, and Skip-Gram, which predicts surrounding words from a target word [51].

- **GloVe**

  GloVe (Global Vectors for Word Representation) is a word embedding method that combines the advantages of matrix factorization and context-based learning. Unlike Word2Vec, which relies on local context windows, GloVe leverages global word co-occurrence statistics from the entire corpus to learn vector representations. This approach enables the model to capture meaningful semantic relationships between words while efficiently handling large-scale datasets [51].

- **FastText**

  FastText is an extension of Word2Vec that represents words as collections of character n-grams rather than treating each word as a single unit. By incorporating subword information, FastText can generate meaningful embeddings for rare and unseen words and better capture morphological patterns. This characteristic makes it particularly effective for languages with rich morphology and for tasks involving limited vocabulary coverage [51].

1.4.2.3 Advanced Contextualized Representations

Unlike static embeddings, contextualized representations generate dynamic word vectors whose meanings depend on the surrounding context. Recent transformer-based models discussed in the previous section, such as **BERT**, **RoBERTa**, and **DeBERTa** have significantly improved NLP performance by capturing long-range dependencies and deep semantic information.

1.5 Conclusion

This chapter outlined key concepts related to fake news and how it can be detected. It defined fake news and its main types, discussed its effects on individuals and society, and emphasized the importance of identifying misinformation. The chapter also examined the main challenges involved in this task. Additionally, it provided an overview of Artificial Intelligence and its paradigm, and the concepts & techniques of NLP in text preprocessing to text representation.

The concepts and technologies covered here form the basis for understanding current FND systems. The next chapter offers a thorough review of existing research and state-of-the-art methods in the field, with particular emphasis on Machine Learning, Deep Learning, and Transformer-based techniques.

Chapter 2

State of the Art

2.1 Introduction

Over the years, various computational approaches have been proposed to automatically detect and classify fake news. The objective of this chapter is to review and analyze some relevant studies in this field, categorized into three main sections: section for machine learning-based approaches, section for deep learning-based approaches, and section for advanced DL approaches that are Transformer-based approaches. Table 2.1 presents an overview of the reviewed works and facilitates their analysis.

2.2 Machine Learning Approaches

Traditional machine learning approaches pioneered automated FND by relying on handcrafted statistical features like TF-IDF and n-grams, which were then fed into classifiers such as SVM, Naïve Bayes, and Random Forest. While computationally efficient and foundational, their heavy dependence on manual feature engineering limits their ability to capture deep contextual semantics.
The following studies present some of the relevant research works that employed traditional machine learning techniques for fake news detection.

Puranik et al [55] proposed a supervised learning framework that compares a Passive Aggressive (PA) online classifier with a Naïve Bayes (NB) model. In this study, the objective is to identify the most effective algorithm for classifying fake versus real articles while maintaining high accuracy on a balanced dataset from Kaggle named Fake News dataset that contains several domains of news(politics, technology, sports, and entertainment news). Linguistic features were extracted using Bag of Words (BOW) and TF IDF representations, and both classifiers were trained with scikit-learn. The PA model achieved 93% accuracy, surpassing NB model.

Jehad et al. [31] compared Decision Tree (J48) and Random Forest classifiers for fake news detection, using a Kaggle dataset of 20,761 articles. The methodology included data preprocessing and TF-IDF feature extraction. J48 and Random Forest were trained. Results show J48 achieves 89.11% accuracy, outperforming Random Forest at 84.97% accuracy.

Sajid Khan et al. [33] presented a domain-agnostic system that combines NLP preprocessing with a Count Vectorizer (CV) and TF-IDF features, followed by a linear SVM classifier. Evaluated on Kaggle and ISOT datasets, the CV + SVM model achieved 89.66% and 99.57% accuracy, outperforming TF-IDF + SVM (87.22% / 92.48%) and baselines like Logistic Regression and Random Forest.

Ahmad et al. [3] proposed an FND framework based on traditional machine learning and ensemble learning techniques (XGBoost, AdaBoost, Bagging, Voting, RF, LR, LSVM). The study evaluated several classifiers on four fake news datasets collected from web sources. Experimental results showed that ensemble methods outperformed individual classifiers, with XGBoost achieving the best performance, reaching up to 99% across all metrics. However, the approach relies on handcrafted textual features and lacks contextual semantic understanding.

2.3 Deep Learning-Based Approaches

To overcome the semantic limitations of traditional methods, deep learning architectures like CNN, RNN, and LSTM were introduced. These models automatically learn hierarchical features directly from raw text, significantly improving the ability to capture sequential and contextual patterns without the need for manual feature extraction.

The following research works illustrate how deep learning architectures have been ap-

plied to fake news detection.

Muhammad Umer et al. [60] proposed a hybrid deep learning architecture that combines Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM), The model was evaluated on the Fake News Challenge (FNC-1) dataset. where Principal Component Analysis (PCA) and Chi-Square feature selection were employed to enhance feature representation before generating Word2Vec embeddings. Experimental results demonstrated the effectiveness of the proposed approach, achieving an F1-score of 97.8%.

Jaybhaye et al. [30] introduced a deep learning–based system for FND, using a Kaggle dataset. Traditional models like Logistic Regression and Random Forests are evaluated first, followed by a Bidirectional LSTM with embedding layers, ReLU activations, and the Adam optimizer. Results show deep learning significantly outperforms traditional approaches, with accuracy reaching 99% with LSTM.

Emmy Ajik et al. [6] aim to develop highly optimized deep learning models for automatic FND. The evaluation uses two datasets, US election news articles and political news. It combines CNN and LSTM models with advanced preprocessing steps, including tokenization, lemmatization, stop-word removal, and word embeddings. Additionally, HyperOpt's Bayesian optimization fine-tunes critical hyperparameters, such as dropout rate, batch size, and number of epochs. The optimized models show substantial improvements over non-optimized versions, with 95.16% accuracy & F1-score.

Abdullah Ali et al. [8] proposed an approach that uses a three-phase, two-stage architecture: NLP preprocessing with TF-IDF bi-grams, followed by an ensemble of deep learning binary classifiers, then an MLP for multi-class classification. It employs tokenization, lemmatization, stemming, ReLU, dropout, and SoftMax. On LIAR and ISOT datasets, the model improves results, achieving an average F1-score of 61% on

LIAR and 100% on ISOT.

Nasir et al.[47] highlight the limitations of traditional machine learning in FND, especially their difficulty in capturing complex textual relationships. To enhance classification accuracy, a hybrid CNN–LSTM model is introduced. The proposed approach is tested on two datasets: FA-KES and ISOT. It integrates a CNN with 100-dimensional GloVe embeddings for local feature extraction and an LSTM to understand long-term dependencies. Results show the model achieves 59% F1-score on FA-KES. On the larger ISOT dataset, the model performed well, with 99% F1-score.

Hager Saleh et al. [58] presented an Optimized Convolutional Neural Network for FND (OPCNN-FAKE), tested on four datasets(FA-KES, ISOT, dataset1 & Fake News Net). Using 200-dimensional GloVe embeddings and Hyperopt for parameter tuning, the architecture includes embedding, dropout, convolutional layers with ReLU, max pooling, flatten, and sigmoid output layers with the ADAM optimizer. Results show OPCNN-FAKE outperforms traditional models and two deep learning models (RNN, LSTM), achieving 99.99% F1 score on ISOT.

2.4 Transformer-Based Approaches

Transformers are a deep learning architecture designed to process sequential data (especially text) using a mechanism called self-attention. They have become the dominant approach in NLP. By generating dynamic, context-aware embeddings, these architectures can capture subtle linguistic dependencies and complex semantic relationships far more effectively than both traditional ML and standard deep learning models. The following studies review some recent transformer-based approaches proposed for the fake news detection task.

Nabeel Raza et al. [56] addressed the shortcomings of traditional FND methods that rely on superficial linguistic features or external metadata, which are often missing or unreliable. This study proposed a BERT-based transformer framework focused solely on content for fake news classification. The model is tested on the large-scale WELFake dataset. It employs the transformer architecture with multi-head self-attention to grasp deep contextual meaning. A notable enhancement is a progressive, episode-based training strategy, combined with the AdamW optimizer and learning rate scheduling, to improve stability and generalization compared to standard fine-tuning. It achieved 95.3% F1-score, outperforming typical BERT fine-tuning.

Kumar et al. [36] propose the Graph-Augmented Transformer Ensemble (GETE), a new hybrid model designed for efficient and intelligent FND. This study utilizes the FakeNewsNet and LIAR benchmarking datasets to evaluate a framework that integrates linguistic and relational features. Specifically, the method combines the semantic capabilities of transformer models, such as BERT and RoBERTa, with the structural insights provided by Graph Neural Networks (GNNs), which are constructed from user-news interactions and source credibility graphs. Results show that GETE achieved a 96.5% F1-score.

Vladislav Kolev et al. [35] proposed a FOREAL (FND based on REAL emotional profiles) framework that addresses this by exploring fine-grained emotional classification of text as a primary detection vector. The methodology involves a two-stage hybrid approach: first, a pre-trained RoBERTa model (REAL) is fine-tuned on an amalgamated corpus based on a taxonomy of six emotions plus a neutral category to generate emotional probability vectors for article titles. These vectors then serve as the feature set for a Random Forest binary classifier trained to predict news veracity. Experimental results demonstrate that the emotional classifier achieves approximately 90% accuracy, while the final integrated model reaches 88% accuracy in fake news detection.

Alghamdi et al. [7] presented a comprehensive transformer-based framework for

COVID-19 FND, focusing on the impact of domain-specific pre-training and hybrid neural architectures. This study systematically compared several transformer models, including BERT, RoBERTa, and COVID-Twitter-BERT (CT-BERT), combined with downstream CNN, LSTM, and GRU networks. In particular, the CT-BERT+BiGRU architecture achieved an F1-score of 98.54% on the COVID-19 fake news dataset. These results highlight the importance of domain-adapted pre-training and fine-tuning strategies in fake news detection, establishing a strong benchmark for transformer-based misinformation classification systems.

Ishraquzzaman et al. [29] proposed a comprehensive fake and AI-generated news detection framework based on transformer architectures and model optimization techniques. The study introduces a large-scale dataset containing real, fake, and AI-generated news titles(Custom D6 dataset) collected from multiple sources. The proposed work is to develop an ensemble transformer model combining RoBERTa and DeBERTa using a confidence-based weighting strategy. Experimental results showed that the proposed ensemble achieved an F1-score of 96.66%, outperforming all individual models.

Li et al. [39] proposed a heterogeneous graph-based Transformer framework for FND (HetTransformer) that represents one of the most advanced state-of-the-art approaches in the field. Unlike traditional content-only models, HetTransformer jointly models multi-modal content information (text, images, and user features) and the heterogeneous propagation structure of news in social networks, where entities such as news articles, posts, and users are interconnected through complex relational edges. The model introduces an encoder-decoder Transformer architecture in which the encoder aggregates information from heterogeneous neighborhoods using sampled graph structures, while the decoder refines representations by focusing on target news nodes and their direct news-type neighbors. Experimental results on benchmark datasets such as PolitiFact, GossipCop, and PHEME demonstrate that HetTransformer significantly outperforms existing baselines, achieving up to 97.5% accuracy on PolitiFact.

Table 2.1: Summary of Previous Studies on Fake News Detection

Category	Ref	Model / Method	Dataset	Best Results	Main Limitation
ML	[55]	Passive Aggressive, Naïve Bayes	Kaggle Fake News dataset	Accuracy = 93% F1 score = 93%	Text-only detection; unable to process multimedia content.
	[31]	Decision Tree, Random Forest, TF-IDF	Kaggle Fake News	Accuracy = 89.11% F1 score= 88.74%	English-only dataset; limited features.
	[33]	SVM + TF-IDF + Count Vectorizer	Kaggle fake News dataset + ISOT	Accuracy = 99.57%	Ignores semantic and contextual information.
	[3]	Ensemble Learning (XGBoost, AdaBoost, Bagging, Voting, RF, LR, LSVM)	Four fake news datasets (including Kaggle datasets)	F1 score= 99% (XGBoost)	Relies on handcrafted textual features
DL	[30]	LSTM	Kaggle News dataset	Accuracy = 99%	Small dataset and inconsistent evaluation reporting.
	[6]	Optimized CNN and LSTM	Election News + Political News	Accuracy = 95.16%	Uses only textual information.
	[60]	Hybrid CNN-LSTM	FNC-1	Accuracy = 97.8% F1 score= 97.8%	dependent on the quality of feature selection.
	[8]	Ensemble Deep Classifiers + MLP	LIAR + ISOT	Accuracy= 100% F1 score=100%(ISOT)	Dataset dependency and classifier ambiguity.
	[47]	Hybrid CNN-RNN	FA-KES + ISOT	Accuracy = 99%	Requires large datasets and lacks interpretability.
	[58]	OPCNN-Fake	FA-KES + ISOT +dataset1 + Fake News Net	Accuracy = 99.99% F1 score= 99.98% (ISOT)	limited set of datasets and languages.
Transformer-Based	[56]	BERT + Progressive Training	WELFake	Accuracy = 95.2% F1 score= = 95.3%	Potential dataset imbalance issues.
	[36]	GETE (BERT/RoBERTa + GNN)	LIAR + FakeNewsNet	Accuracy = 96.5% F1 score= = 96.5%	Additional computational overhead.
	[35]	RoBERTa + Random Forest	Fake and Real News	Accuracy = 88% F1 score= 88%	emotion combinations dependency & computational and time constraints.
	[7]	CT-BERT + BiGRU	COVID-19 Fake News Dataset	Accuracy: 98.46% F1-Score: 98.54%	Domain limited (COVID-19)
	[29]	Ensemble (RoBERTa + DeBERTa)	Custom D6 dataset	Accuracy: 96.65% F1-score: 96.66%	High computational cost & dataset dependency
	[39]	HetTransformer	FakeNewsNet + PHEME	Accuracy= 99.0% F1 score= 97.9%	Requires heterogeneous social network graph data

Table 2.1 illustrates the progression of fake news detection methods from traditional machine learning techniques to deep learning and, more recently, Transformer-based models. Traditional machine learning approaches achieved competitive performance but relied heavily on handcrafted features and extensive feature engineering, limiting their ability to capture complex semantic relationships. Deep learning models improved detection accuracy through automatic feature extraction and representation learning; however, they often require large labeled datasets, substantial computational resources, and may offer limited interpretability.

Transformer-based models, particularly BERT and its variants, represent a significant advancement by leveraging bidirectional contextual representations to better understand linguistic and semantic dependencies within text. Although some deep learning models report the highest performance on specific benchmark datasets, these results are often dataset-dependent and may not generalize well to unseen or real-world data. Furthermore, exceptionally high reported accuracies in certain studies

may reflect dataset bias or overfitting rather than superior generalization.

Based on this analysis, the present study focuses on BERT not solely because of its reported performance, but because of its strong contextual language understanding, transfer learning capabilities, and proven effectiveness across a wide range of natural language processing tasks. Fine-tuning BERT for binary fake news classification provides an opportunity to investigate whether these advantages can achieve a balance between classification accuracy, generalization, and computational efficiency, making it a suitable and promising approach for this research [56].

2.5 Conclusion

This chapter presented an overview of existing fake news detection approaches, including machine learning, deep learning, and Transformer-based approaches. A comprehensive analysis of previous studies revealed the strengths and limitations of each category of methods. Motivated by these findings, the next chapter presents our proposed methodology, including the dataset, the preprocessing pipeline, system architecture adopted in this work.

Chapter 3

Methodology

3.1 Introduction

This chapter presents the methodological framework of the proposed transformer-based fake news detection system. It outlines the overall system architecture, the datasets used for training and evaluation, and the preprocessing pipeline applied to prepare textual data for model ingestion. The objective is to provide a clear and structured overview of the experimental design that supports the implementation and evaluation stages presented in the following chapter.

3.2 System Architecture

The proposed FND system, as illustrated in Figure 3.1, is based on a transformer architecture designed for binary classification of news articles into real and fake categories. The processing pipeline begins with raw news input, which undergoes preprocessing to normalize and structure the text. The cleaned text is then tokenized using model-specific tokenizers. The resulting token sequences are fed into the transformer encoder, which generates contextualized embeddings capturing semantic and syntactic relationships within the input text. These representations are then passed to a classification head. Finally, the classification layer outputs a probability score used to assign each news article to one of two classes: fake or real.

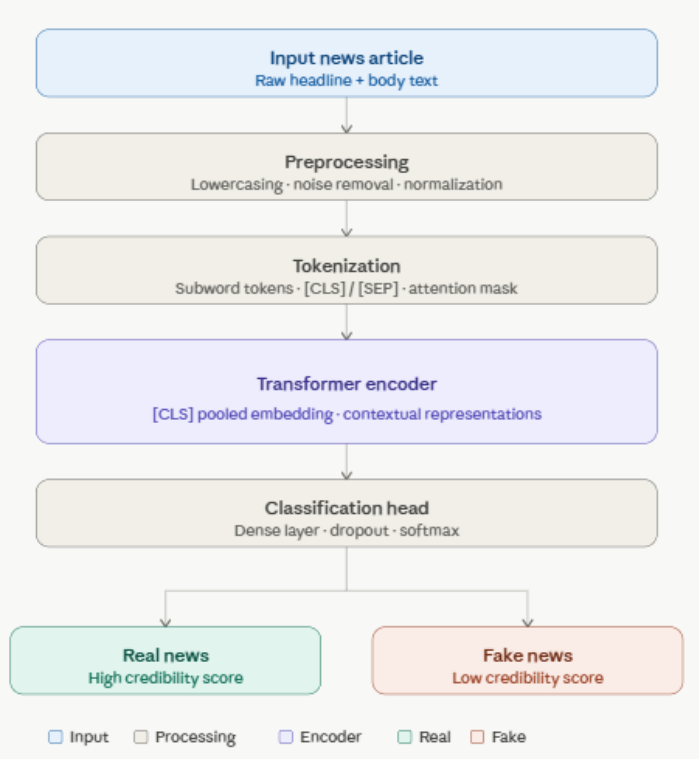


Figure 3.1: Proposed System Architecture

Following this architecture, the implemented framework leverages several transformer-based models [57], including BERT [56], RoBERTa [35], DeBERTa-v3 [29], as well as a hybrid BERT–RoBERTa model. Each model independently processes the input text through its respective encoder to generate contextual embeddings, which are then used for downstream classification in the FND task.

3.3 Datasets

To evaluate the performance and generalization capability of the proposed FND models, this study leverages two widely used publicly available benchmark datasets: ISOT Fake News and WELFake, which are illustrated in Table 3.1. These datasets are selected due to their diversity in size, domain coverage, and annotation strategies, which enable a comprehensive assessment of misinformation detection models across different real-world scenarios.

Table 3.1: Overview of benchmark datasets used in Our study.

Dataset	Lang.	Total Samples	Real	Fake	Features
ISOT Fake News [1]	EN	44,919	21,417	23,502	Title, Text, Subject, Date
WELFake [2]	EN	72,134	37,106	35,028	Title, Text, Label

ISOT Fake News Dataset[3] The ISOT Fake News Dataset comprises 44,919 news articles, including 23,502 fake and 21,417 real samples. The real news articles were primarily collected from Reuters, while the fake news articles were gathered from a variety of unreliable online sources. Each entry contains the article title, full text, subject category, and publication date [5].

Primary Core Features

- **Article Title:** The headline of the news piece. It is frequently utilized independently in natural language processing (NLP) to detect misinformation based on clickbait or sensationalism.
- **Text Body:** The full content of the article.
- **Subject:** Broad category tags (e.g., politics, world news, general).
- **Date: The dates in the ISOT Fake News Dataset typically range from 2016 to 2017.**

WELFake Dataset[4] [5] The WELFake dataset contains 72,134 news articles, consisting of 37,106 real and 35,028 fake samples. It was constructed by merging multiple datasets, including Kaggle, McIntire, Reuters, and BuzzFeed Political datasets, to enhance diversity and reduce dataset bias. Each instance includes a title, full article text, and a binary label indicating its veracity [63].

Primary Core Features

[3] https://www.kaggle.com/datasets/csmalarkodi/isot-fake-news-dataset

[4] https://zenodo.org/record/4561253

[5] https://www.kaggle.com/datasets/studymart/welfake-dataset-for-fake-news

- **title:** The headline or title of the news article. This feature captures immediate attention-grabbing elements and is heavily utilized for title-only model testing.
- **text:** The main body and content of the news article. This is the primary independent variable for most natural language processing models, containing the semantic narrative.
- **label:** The target or output variable for classification, represented by binary integers(0 = fake and 1 = real).

3.4 Data Preprocessing

Data preprocessing plays a crucial role in FND by transforming raw textual data into a structured format suitable for transformer-based models. In this study, a unified preprocessing pipeline is designed for all models, with minor adaptations depending on architectural requirements.

Feature Construction and Input Representation

For transformer-based models trained on full news articles, such as BERT [56], RoBERTa[35], DeBERTa [29], and the proposed hybrid approach, multiple textual fields were systematically merged to preserve rich contextual information. In particular, the title and the article body were concatenated into a single input sequence. This design allows the model to jointly encode both the concise summary provided by the headline and the detailed narrative of the article, thereby improving contextual understanding and semantic coherence.

In the hybrid architecture, additional metadata specifically the subject field was also incorporated into the input representation. This extended concatenation further enhances the model's ability to capture topical context and inter field relationships, en-

abling more robust feature learning. Overall, this strategy improves the transformer's capacity to model dependencies between headlines, content, and auxiliary metadata, leading to stronger contextual representations.

Data Cleaning and Missing Value Handling

The datasets were first examined for missing, incomplete, or inconsistent entries. Samples lacking essential fields such as *text*, *title*, or *label* were removed to ensure data integrity and avoid errors during tokenization and training. For optional attributes (e.g., *subject* in the hybrid model), missing values were replaced with a default placeholder token, `"Unknown"`, rather than discarding the samples, in order to preserve dataset size. Additionally, extremely short or empty texts were filtered out, as they do not provide sufficient linguistic information for reliable classification.

Label Preparation

A binary classification scheme was adopted across all experiments, where each sample is assigned one of two possible labels: fake news is mapped to 0, while real news is mapped to 1.

For the ISOT dataset, labels were assigned during the data loading phase based on the originating source of each file. This approach ensures consistent and reliable supervision across heterogeneous datasets. Since transformer-based classification heads directly accept integer-encoded labels, no additional encoding schemes, such as one-hot encoding, were required.

Dataset Splitting Strategy

The dataset is split into training, validation, and test sets using stratified sampling to preserve class balance. Two splitting configurations are used depending on the

model, as summarized in Table 3.2.

Table 3.2: Dataset splitting strategy across different models

Model	**Training**	**Validation**	**Testing**
RoBERTa	70%	15%	15%
BERT / DeBERTa-v3 / Hybrid Model	80%	10%	10%

Tokenization

Each transformer model employs its own tokenizer and preprocessing pipeline. Tokenization parameters are selected according to the architectural design of each model.

As summarized in Table 3.3, BERT uses WordPiece tokenization, RoBERTa uses Byte-Pair Encoding (BPE), and DeBERTa-v3 employs a subword-based tokenizer. The hybrid model combines WordPiece and BPE strategies to support both encoders.

Table 3.3: Tokenizer configurations for transformer-based models

Setting	**BERT**	**DeBERTa-v3**	**RoBERTa**	**Hybrid**
Tokenizer Type	bert-base-uncased	DeBERTa tokenizer	roberta-base	Combined setup
Tokenization Method	WordPiece	Subword	BPE	WordPiece + BPE
Special Tokens	[CLS], [SEP]	[CLS], [SEP]	<s>, </s>	Model-dependent
Max Sequence Length	128	128	128	256
Padding/Truncation	Enabled	Enabled	Enabled	Enabled
Output Format	TensorFlow tensors	Framework tensors	PyTorch tensors	Multi-stream tensors

Sequence Padding and Truncation

All tokenized sequences are padded or truncated to a fixed maximum length to ensure uniform input dimensions. Specifically, BERT, RoBERTa, and DeBERTa-v3 use a maximum length of 128 tokens, and the hybrid model uses 256 tokens.

Attention masks are applied to distinguish real tokens from padding tokens, ensuring that padding does not influence the self-attention mechanism.

Tensor Preparation and Data Loading

After tokenization, the data is converted into tensor-based formats compatible with TensorFlow or PyTorch frameworks. Each sample consists of *input_ids*, *attention_mask*, and *labels*.

The resulting datasets are organized using DataLoaders to enable mini-batch training. During training, random shuffling is applied, while sequential loading is used for validation and testing to ensure consistent evaluation. The batch size is adjusted according to model complexity and available computational resources.

3.5 Conclusion

This chapter presented the overall methodology of the proposed fake news detection system, including the system architecture, benchmark datasets, and data preprocessing pipeline. It established a consistent and structured framework for transforming raw text into model-ready inputs and set the foundation for the experimental evaluation conducted in the following chapter.

Chapter 4

Implementation and Experiments

4.1 Introduction

This chapter outlines the implementation process and experimental setup for this work. It begins with an overview of the development environment, programming language, and libraries utilized to create the models. Next, it details the implementation of the BERT, RoBERTa, DeBERTa-v3, and Hybrid models, including their training configurations and hyperparameters. Additionally, the chapter covers the evaluation metrics used to assess model performance. Finally, it includes a results and discussion section where we present & discuss our proposed work results & performance.

4.2 Experimental Setup

The experiments conducted in this work were carried out using a set of well-established development tools and computational resources, selected for their reliability, accessibility, and suitability for deep learning and natural language processing tasks.

- **Google Colaboratory (Google Colab)**[1]: A cloud-based platform built on Jupyter Notebook that enables Python code execution through a web browser. It is widely used in machine learning research due to its free access to GPU resources, collaborative features, and ease of use.
- **Python**[2]: A high-level programming language used for implementing all components of the proposed system. Python is extensively adopted in data science, machine learning, and deep learning because of its simplicity, readability, and rich ecosystem of scientific computing libraries.

The following Python libraries and frameworks were employed throughout the implementation process:

[1] https://colab.research.google.com/
[2] https://www.python.org/

- **Hugging Face Transformers**[3]: Provides pre-trained transformer models and tools for fine-tuning natural language processing tasks.
- **PyTorch**[4]: An open-source deep learning framework used for implementing and training transformer-based neural networks.
- **TensorFlow**[5]: A comprehensive deep learning framework that supports large-scale model training and deployment [19].
- **Keras**[6]: A high-level neural network API that facilitates rapid model development and experimentation.
- **NumPy**[7]: A fundamental library for numerical computation and efficient manipulation of multidimensional arrays [22].
- **Pandas**[8]: A data analysis and manipulation library used for data preprocessing, cleaning, and organization [43].
- **Scikit-learn**[9]: A machine learning library used for data preprocessing, performance evaluation, and implementation of traditional machine learning utilities [52].

4.3 Implementation

This section describes the implementation of the transformer-based models used in this study. The objective is to evaluate and compare the effectiveness of different state-of-the-art transformer architectures for fake news detection. Four models were implemented, namely BERT, RoBERTa, DeBERTa-v3, and a hybrid BERT+RoBERTa

[3] https://huggingface.co/docs/transformers/index
[4] https://pytorch.org/
[5] https://www.tensorflow.org/
[6] https://keras.io/
[7] https://numpy.org/
[8] https://pandas.pydata.org/
[9] https://scikit-learn.org/stable/

ensemble model. The implementation process included dataset preparation, text preprocessing, tokenization, model fine-tuning, prediction generation, and performance evaluation.

4.3.1 BERT Implementation

The proposed FND model is built upon the pre-trained BERT-base-uncased architecture and adopts a transfer learning framework for binary text classification. In this setup, raw news articles are first processed using the BERT tokenizer, which transforms the text into subword tokens and generates the corresponding input IDs and attention masks. All input sequences are standardized through padding or truncation to a fixed maximum length, ensuring uniform tensor dimensions across the dataset. The tokenized representations are then fed into the BERT encoder, which produces deep contextual embeddings by modeling bidirectional dependencies within the input text. Among the outputs, the hidden state corresponding to the special `[CLS]` token is extracted and used as a fixed-length representation of the entire sequence. This sentence-level embedding is subsequently passed to a task-specific classification head composed of regularization and dense layers, followed by a sigmoid-activated output layer that predicts whether the input article is real or fake. Figure 4.1 illustrates the overall architecture of the proposed model, highlighting the flow of information from tokenized news articles through the BERT encoder and classification layers to the final prediction output.Table 4.1 summarizes the complete set of hyperparameter configurations used during training. To ensure a fair and consistent evaluation across datasets, the same hyperparameter settings were applied throughout all BERT-based experiments.

Table 4.1: Hyperparameter Configuration of the BERT-Based Model

Hyperparameter	**Value**
Pretrained Model	BERT Base Uncased
Maximum Sequence Length	128
BERT Hidden Size	768
Dense Layer Units	64
Dense Activation	Tanh
Dropout Rate 1	0.5
Dropout Rate 2	0.5
Output Units	1
Output Activation	Sigmoid
Optimizer	Adam
Learning Rate	$1 * 10^{-5}$
Loss Function	Binary Cross-Entropy
Batch Size	30
Maximum Epochs	20
Early Stopping Patience	3
Reduce LR Patience	2
LR Reduction Factor	0.5
Minimum Learning Rate	$1 * 10^{-7}$

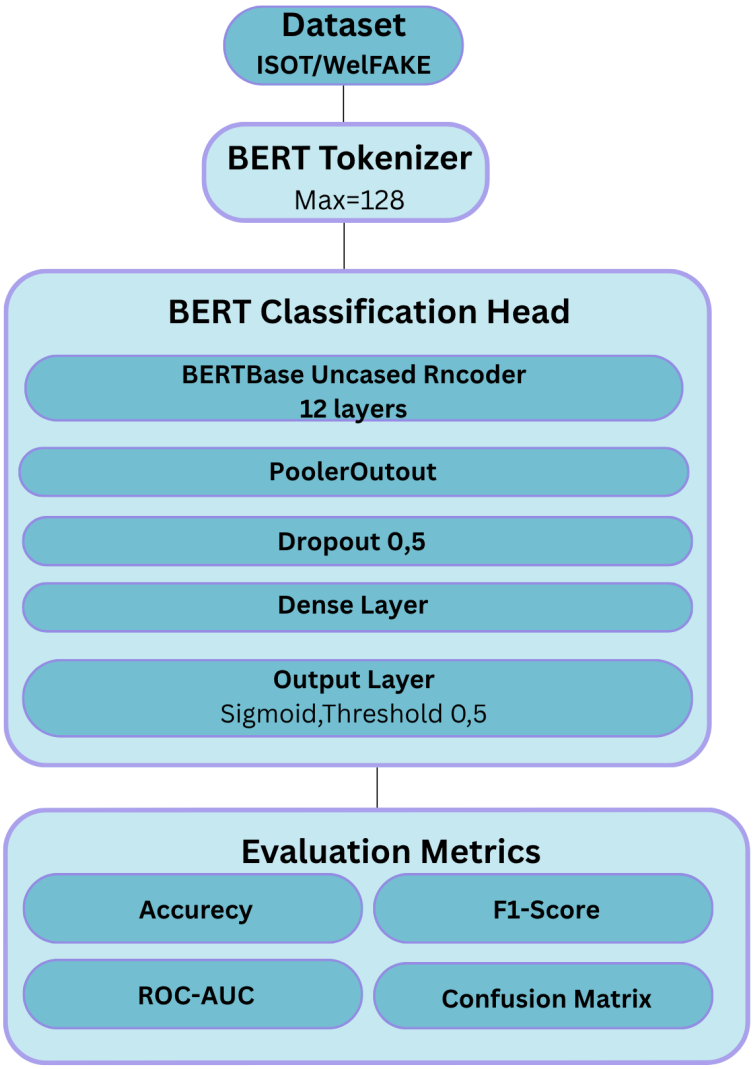


Figure 4.1: Overall architecture of the proposed BERT-base model.

4.3.2 RoBERTa Implementation

The proposed FND model is based on the pre-trained RoBERTa-base architecture and employs a transfer learning approach for binary classification. News titles and texts are concatenated, then tokenized and converted into contextual representations through the RoBERTa encoder. To enrich the semantic information learned by the transformer, a set of handcrafted linguistic and stylistic features associated with fake news characteristics is extracted and fused with the contextual embedding of the input sequence.

The resulting representation is fed into a feedforward classification network to distinguish between real and fake news articles. The model is trained using a progressive fine-tuning strategy, where the classification layers are optimized before jointly fine-tuning the entire transformer encoder. Figure 4.2 illustrates the overall architecture of the model, while the complete hyperparameter configuration is summarized in Table 4.2.

Table 4.2: Hyperparameter Summary for RoBERTa-based Model

Hyperparameter	**Value**
Transformer Backbone	RoBERTa-base
Hidden Size	768
Maximum Sequence Length	128
Handcrafted Features	6
Fusion Dimension	774
Dense Layer 1	256
Dense Layer 2	64
Output Classes	2
Activation Function	GELU
Dropout Rate	0.30
Batch Size	16
Phase 1 Learning Rate	$5 * 10^{-4}$
Phase 2 Base Learning Rate	$2 * 10^{-5}$
Weight Decay	0.01
Label Smoothing	0.05
Gradient Clipping	1.0
Early Stopping Patience	3
Phase 2 Maximum Epochs	8

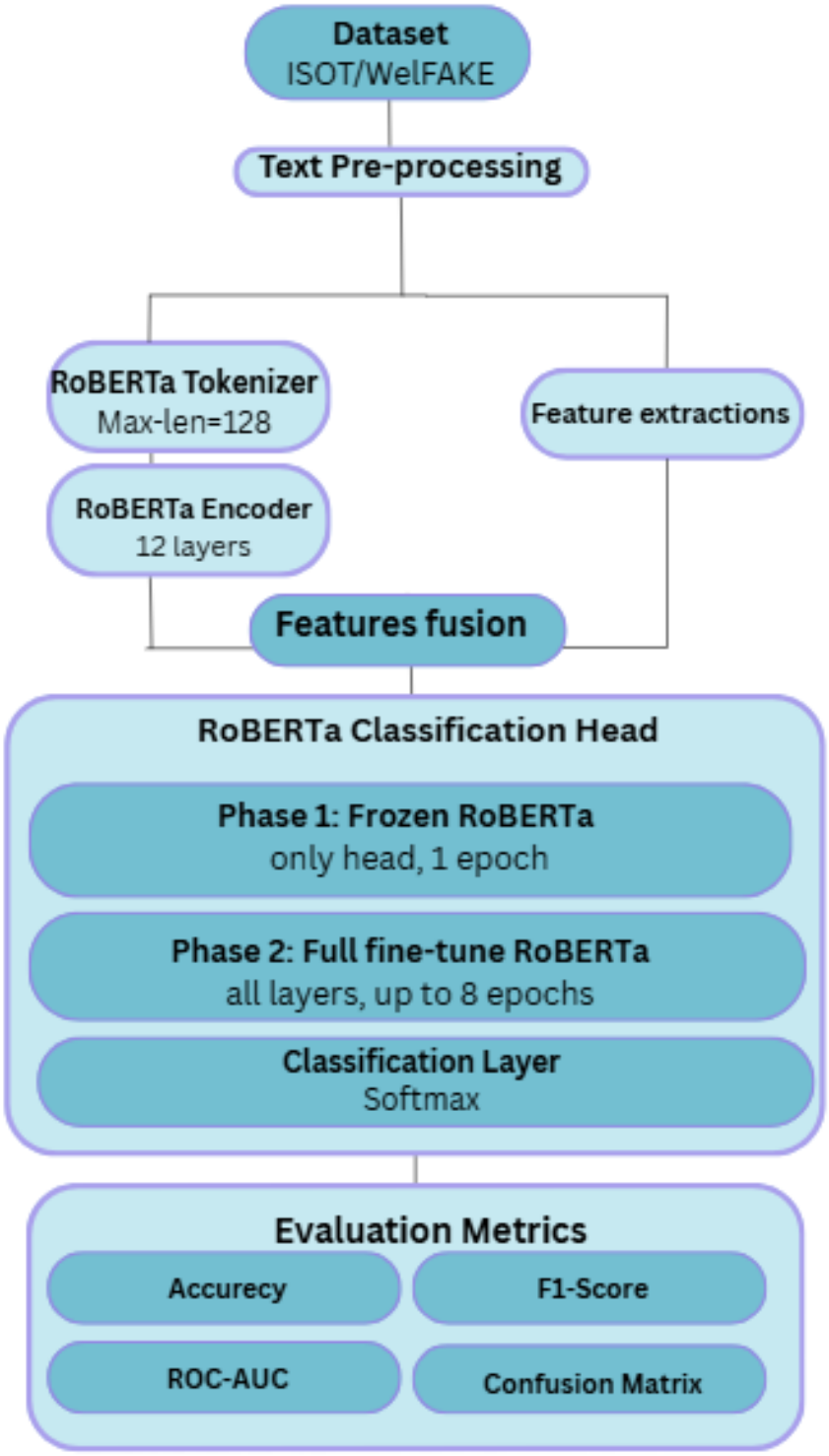


Figure 4.2: Overall architecture of the proposed RoBERTa-based model.

4.3.3 DeBERTa-v3 Implementation

The proposed model is based on the pre-trained DeBERTa-base architecture and adopts a transfer learning approach for binary text classification. News articles are first tokenized using the DeBERTa tokenizer and transformed into input representations suitable for the transformer encoder. The encoded inputs are then processed through DeBERTa's disentangled self-attention mechanism, which separately models content and positional information, enabling the capture of complex semantic relationships within the text.

The contextual representation corresponding to the classification token is extracted and passed to a sequence classification layer that predicts whether a news article is real or fake. The model is fine-tuned end-to-end on the target datasets.

The overall architecture of the proposed DeBERTa-based model is illustrated in Figure 4.3, while the complete hyperparameter configuration is presented in Table 4.3. To ensure a fair and consistent comparison across datasets, the same experimental settings were maintained throughout all DeBERTa-based experiments.

Table 4.3: Hyperparameter configuration of the DeBERTa-based model.

Hyperparameter	**Value**
Pre-trained Model	`deberta-base`
Number of Classes	2
Maximum Sequence Length	128
Learning Rate	$2 * 10^{-5}$
Training Epochs	3
Train Batch Size	2
Evaluation Batch Size	2
Gradient Accumulation Steps	4
Effective Batch Size	8
Weight Decay	0.01
Optimizer	AdamW
Gradient Checkpointing	Enabled
Maximum Gradient Norm	1.0
Evaluation Frequency	Every 500 steps
Model Saving Frequency	Every 500 steps
Best Model Metric	F1-score
Mixed Precision (FP16)	Disabled
Numerical Precision	FP32
Dynamic Padding	Enabled
Padding Multiple	8

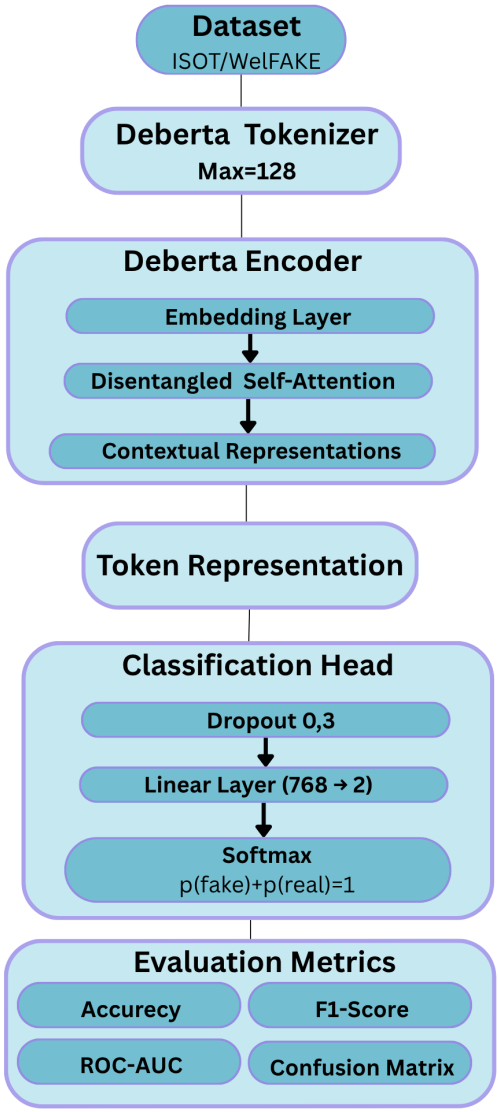


Figure 4.3: Overall architecture of the proposed DeBERTa-base model.

4.3.4 Hybrid Implementation

The proposed hybrid model combines two pre-trained transformer architectures, BERT-base and RoBERTa-large, within an ensemble learning framework for binary fake news classification. After tokenization and input standardization, the same news text is processed independently by the BERT and RoBERTa branches. Each transformer generates contextualized representations and produces a probability distribution over the target classes. The outputs of the two models are subsequently combined using a weighted soft-voting ensemble strategy. Let P_{BERT} and P_{RoBERTa} denote the predicted probability distributions from BERT and RoBERTa, respectively. The final prediction is computed as:

$$P_{\text{ensemble}} = w_{\text{BERT}} P_{\text{BERT}} + w_{\text{RoBERTa}} P_{\text{RoBERTa}}$$

subject to:

$$w_{\text{BERT}} + w_{\text{RoBERTa}} = 1$$

Where the final prediction is obtained by aggregating the probabilities generated by both classifiers. Figure 4.4 illustrates the overall architecture of the proposed hybrid model, while the complete hyperparameter configuration is presented in Table 4.4.

Table 4.4: Hyperparameter Configuration of the Proposed Hybrid BERT+RoBERTa Model

Hyperparameter	**Value**
BERT Model	BERT-Base-Uncased
RoBERTa Model	RoBERTa-Large
Number of Classes	2
Maximum Sequence Length	256
Batch Size	16
Epochs	4
Learning Rate	$2 * 10^{-5}$
Optimizer	AdamW
Epsilon	$1 * 10^{-8}$
Weight Decay	0.01
Warmup Ratio	10%
Gradient Clipping	1.0
Label Smoothing	0.1
Random Seed	42
Ensemble Method	Weighted Soft Voting
Weight Search Step	0.05
Output Labels	Fake / Real

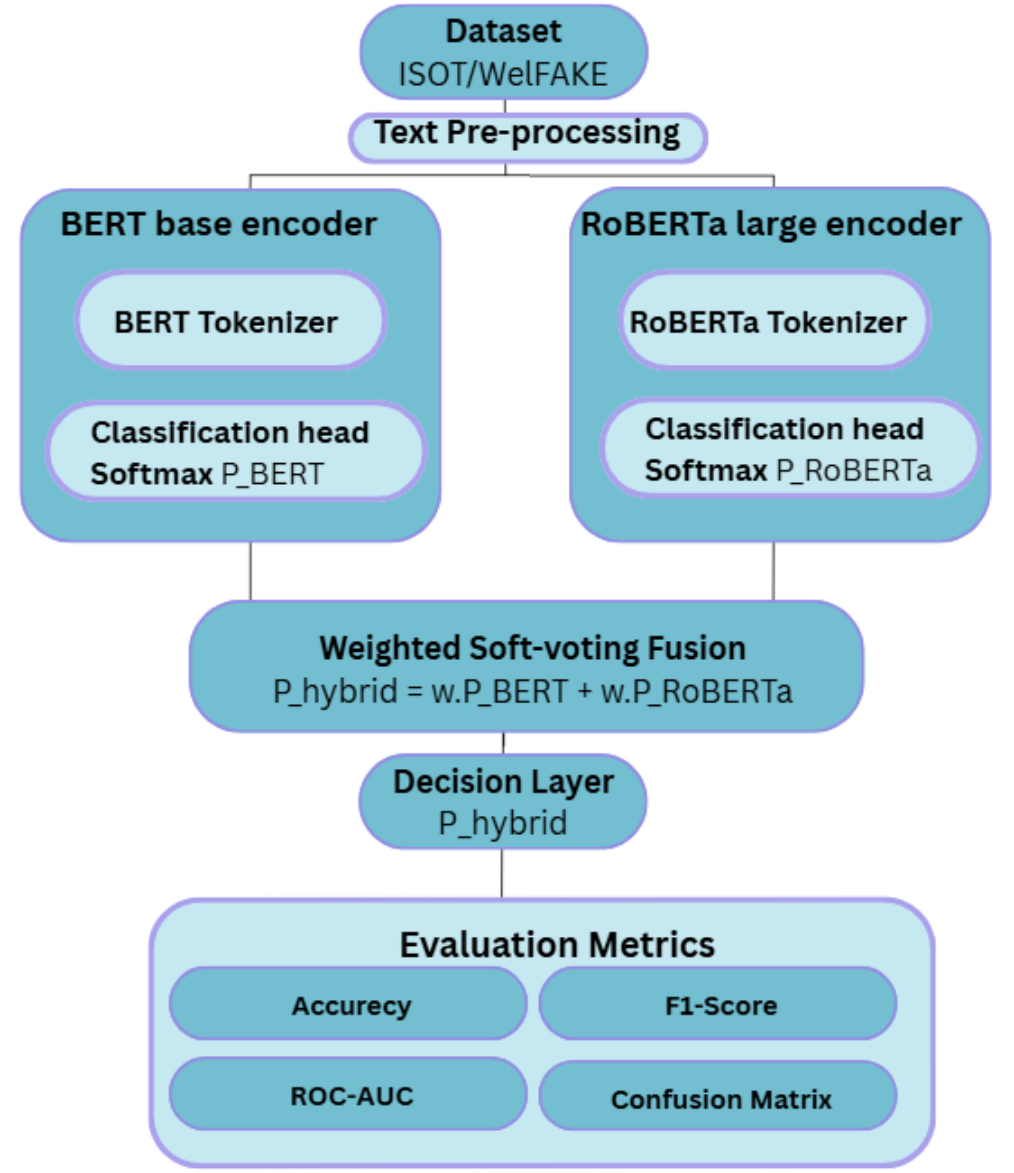


Figure 4.4: Overall architecture of the proposed Hybrid model.

4.4 Evaluation Metrics

To assess the effectiveness of the proposed fake news detection system, widely used evaluation metrics were employed on the test set. Table 4.5 summarizes the evaluation metrics used in this study along with their definitions and interpretations.

Table 4.5: Evaluation Metrics

Metric	**Formula / Description**	**Interpretation**
Accuracy [17]	$\frac{TP+TN}{TP+TN+FP+FN}$	Measures the overall proportion of correctly classified instances among all observations.
Precision [17]	$\frac{TP}{TP+FP}$	Measures the proportion of correctly predicted positive instances among all positive predictions.
Recall [17]	$\frac{TP}{TP+FN}$	Measures the ability of the model to correctly identify actual positive instances.
F1-Score [17]	$\frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$	Provides a balanced measure by combining Precision and Recall through their harmonic mean.
ROC–AUC [12, 27]	Area under the Receiver Operating Characteristic (ROC) curve	Evaluates the model's ability to distinguish between classes across different decision thresholds.
Log Loss [13]	$-[y\log(p)+(1-y)\log(1-p)]$	Measures the quality of predicted probabilities by penalizing confident– incorrect predictions.

4.5 Results and Discussion

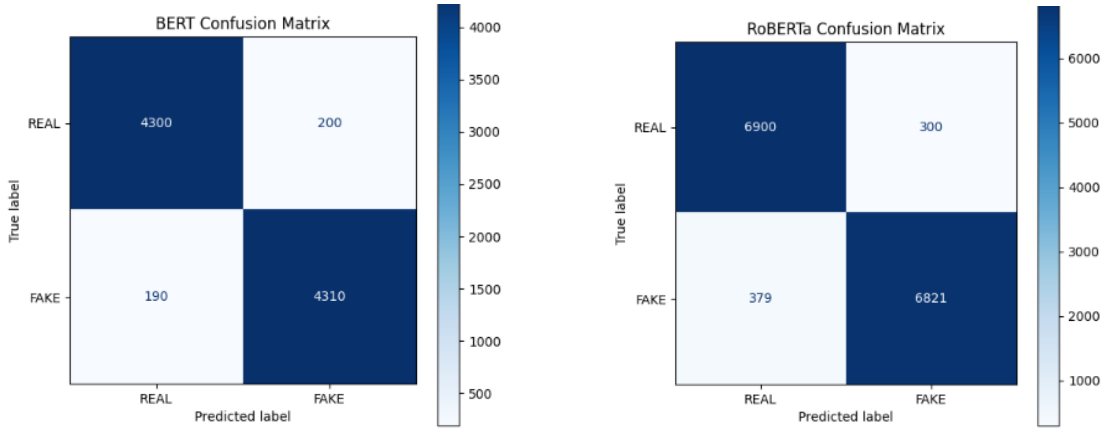
To evaluate the effectiveness of the selected transformer-based models for fake news detection, a series of experiments was conducted on two datasets: WELFake and ISOT. The performance of each model was assessed using two widely adopted evaluation metrics, Accuracy and F1-score. Table 4.6 summarizes the obtained results and provides a comparative analysis of the models across the different datasets.

Table 4.6: Performance comparison (%) of transformer-based models across datasets.

Model	WELFake		ISOT	
	Acc.	**F1**	**Acc.**	**F1**
BERT	91.23	91.21	**95.66**	**95.67**
RoBERTa	**95.29**	**95.29**	86.41	86.40
DeBERTa-v3	80.50	80.48	86.83	86.83
BERT+RoBERTa	65.52	65.50	66.40	66.39

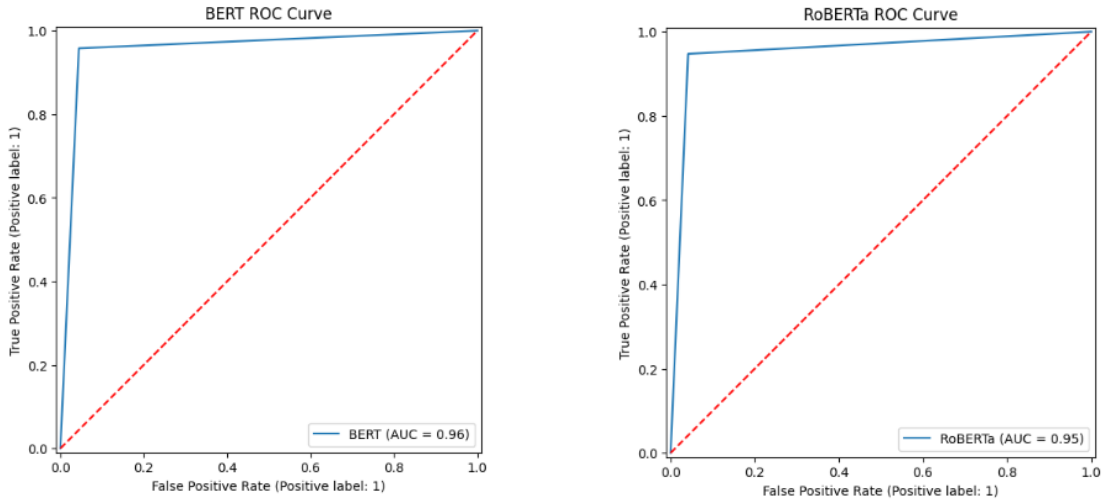
Figures 4.5 present the confusion matrices obtained by the BERT and RoBERTa models, as they achieved the best performance across datasets. Both models demonstrate strong classification performance, with most instances concentrated along the main diagonal, indicating correctly classified samples.

Figures 4.6 show the ROC curves for both models, indicating remarkable classification performance, with AUC values approaching 1 across both the WELFake and ISOT datasets. However, the BERT model exhibits fewer misclassified instances on the ISOT dataset, which explains its higher accuracy and F1-score compared to RoBERTa.



(a) ISOT dataset (b) WELFake dataset

Figure 4.5: Confusion matrices of the best-performing models across datasets



(a) ISOT dataset (b) WELFake dataset

Figure 4.6: ROC curves of the best-performing models across datasets

4.5.1 Discussion

The experimental results as shown in Table 4.6, reveal several remarkable points that reflect both the strengths and limitations of the evaluated models and experimental configurations. RoBERTa achieved the highest performance on the WELFake dataset, attaining an accuracy and F1-score of 95.29%. This outcome can be attributed to its progressive fine-tuning strategy and the integration of handcrafted linguistic features, which effectively capture stylistic markers prevalent in the multi-source WELFake corpus. However, RoBERTa's performance dropped considerably on the ISOT dataset with an F1-score of 86.40%, suggesting that the handcrafted features, while beneficial for stylistically diverse data, introduced noise when applied to ISOT's more homogeneous journalistic register.

In contrast, BERT demonstrated the most consistent performance across both datasets, achieving an F1-score of 91.23% on WELFake and 95.67% on ISOT. Despite being architecturally less advanced than RoBERTa and DeBERTa-v3, its simpler fine-tuning setup (a single learning rate, standard cross-entropy loss, and no auxiliary features) resulted in more robust generalization. This finding supports the hypothesis that model complexity does not always correlate with cross-domain performance, and

that architectural Simplicity can serve as an implicit regularizer.

The underperformance of DeBERTa-v3, which achieved only 80.50% on WELFake, despite being the most architecturally advanced model evaluated, is primarily attributable to constraints in its experimental configuration rather than model limitations. Specifically, a training batch size of 2, combined with only three training epochs and disabled mixed precision, likely prevented the model from converging to an optimal solution. These constraints, presumably motivated by hardware limitations, disproportionately affected DeBERTa-v3 due to its larger parameter space and greater sensitivity to batch-level gradient noise. Future work should re-evaluate DeBERTa-v3 under more favorable training conditions to obtain a fairer comparison.

The hybrid BERT–RoBERTa ensemble achieved the lowest performance among the evaluated models on both datasets, obtaining F1-scores close to 65%. Although ensemble approaches are generally expected to benefit from the strengths of multiple models, this was not observed in the present study. One possible explanation is that the two models were trained using different output configurations, which likely produced probability estimates that were not directly comparable. As a result, combining their predictions through weighted averaging may have introduced inconsistencies rather than improving classification performance. In addition, the limited range of weight combinations explored during the ensemble design may have prevented the identification of a more effective configuration. These results indicate that simply combining model outputs is not always sufficient to enhance performance. More advanced ensemble strategies, such as probability calibration techniques or meta-learning approaches, may be required to fully exploit the complementary capabilities of different transformer models.

Overall, the results demonstrate that dataset characteristics, fine-tuning strategy, and training configuration are at least as influential as model architecture in determining FND performance. No single model consistently outperformed the others across both datasets, which highlights the importance of dataset-aware model selection and

configuration.

4.6 Conclusion

This chapter detailed the implementation process and experimental setup for the proposed fake news detection models. It described the development environment, software tools, model architectures, and training configurations. Additionally, the evaluation metrics used to measure model performance were introduced. It also covered the experimental results and discussed them.

Conclusion and Perspectives

This thesis addressed the growing problem of fake news detection in the context of rapidly expanding digital information platforms. The main research question guiding this work was how to effectively leverage transformer-based deep learning models to automatically distinguish between real and fake news articles with high accuracy and robustness. This problem is particularly relevant due to the increasing spread of misinformation, which can significantly influence public opinion, political processes, and societal trust in media.

The interest of this research lies in the critical need for reliable automated systems capable of detecting misinformation at scale. Traditional methods based on manual verification or classical machine learning approaches are insufficient when dealing with large volumes of textual data. Therefore, transformer-based models provide a promising solution due to their ability to capture deep contextual and semantic relationships in text.

The contribution of this work lies in the design and evaluation of a framework based on multiple state-of-the-art transformer architectures, including BERT, RoBERTa, DeBERTa-v3, and a hybrid ensemble model. The research methodology consisted of several key stages: data collection from publicly available datasets (ISOT and WELFake), preprocessing of textual data through cleaning, tokenization, and normalization, fine-tuning of pre-trained transformer models, and evaluation using standard metrics such as accuracy, F1-score, ROC-AUC, and confusion matrices.

Despite the fact that the experiment design was based on advanced models, this study has certain limitations. First, the work is restricted to English-language datasets, which limits its applicability to multilingual environments. Second, the models were trained and evaluated on relatively static benchmark datasets, which may not fully reflect the dynamic nature of real-world misinformation. Third, computational constraints limited the exploration of larger transformer architectures and more complex ensemble strategies.

The results obtained demonstrate that transformer-based models are effective for

fake news detection tasks. RoBERTa achieved the best performance on the WELFake dataset, while BERT showed the most stable and consistent results on the ISOT dataset. DeBERTa-v3 and the hybrid ensemble model performed less effectively, highlighting the sensitivity of model performance to dataset characteristics and training configurations.

From a scientific perspective, this study contributes to the growing body of research on natural language processing for misinformation detection by providing a systematic comparative analysis of multiple transformer architectures under a unified experimental setting. From a practical perspective, the findings demonstrate the potential of these models to be integrated into real-world systems for automated news verification, social media monitoring, and decision-support tools.

Finally, several perspectives can be considered for future work. These include extending the study to multilingual and cross-domain datasets, exploring more advanced transformer architectures such as large-scale language models, and integrating multimodal information such as images and metadata for improved detection performance. Additionally, future research could focus on developing more robust ensemble strategies and improving model interpretability to enhance transparency and trust in automated fake news detection systems.

Bibliography

[1] Dildar Masood Abdulqader, A Mohsin Abdulazeez, and Diyar Qader Zeebaree. Machine learning supervised algorithms of gene selection: A review. *Machine Learning*, 62(03):233–244, 2020.

[2] Temesgen Abraha and Amril Nazir. Evaluation of transformer-based neural language models for writing feedback and automated essay scoring, 02 2024.

[3] Iftikhar Ahmad, Muhammad Yousaf, Suhail Yousaf, and Muhammad Ovais Ahmad. Fake news detection using machine learning ensemble methods. *Complexity*, 2020(1):8885861, 2020. doi: https://doi.org/10.1155/2020/8885861. URL https://onlinelibrary.wiley.com/doi/abs/10.1155/2020/8885861.

[4] Alim Al Ayub Ahmed, Ayman Aljabouh, and Myung Choi. Detecting fake news using machine learning : A systematic literature review, 02 2021.

[5] Hadeer Ahmed, Issa Traore, and Sherif Saad. Detecting opinion spams and fake news using text classification. *Security and Privacy*, 1:e9, 12 2017. doi: 10.1002/spy2.9.

[6] Emmy Ajik, Georgina Obunadike, and Faith Echobu. Fake news detection using optimized cnn and lstm techniques. *Journal of Information Systems and Informatics*, 5:1044–1057, 08 2023. doi: 10.51519/journalisi.v5i3.548.

[7] Jawaher Alghamdi, Yuqing Lin, and Suhuai Luo. Towards covid-19 fake news detection using transformer-based models. *Knowledge-Based Systems*,

274:110642, 2023. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2023.110642. URL https://www.sciencedirect.com/science/article/pii/S0950705123003921.

[8] Abdullah Ali, Fuad Ghaleb, Bander Al-rimy, Fawaz Alsolami, and Asif Khan. Deep ensemble fake news detection model using sequential deep learning technique. *Sensors*, 22:6970, 09 2022. doi: 10.3390/s22186970.

[9] Ioannis Antonopoulos, Valentin Robu, Benoit Couraud, Desen Kirli, Sonam Norbu, Aristides Kiprakis, David Flynn, Sergio Elizondo-Gonzalez, and Steve Wattam. Artificial intelligence and machine learning approaches to energy demand-side response: A systematic review. *Renewable and Sustainable Energy Reviews*, 130:109899, 2020. ISSN 1364-0321. doi: https://doi.org/10.1016/j.rser.2020.109899. URL https://www.sciencedirect.com/science/article/pii/S136403212030191X.

[10] Nihel Fatima Baarir and Abdelhamid Djeffal. Fake news detection using machine learning. In *2020 2nd International workshop on human-centric smart environments for health and well-being (IHSH)*, pages 125–130. IEEE, 2021.

[11] Pritika Bahad, Preeti Saxena, and Raj Kamal. Fake news detection using bi-directional lstm-recurrent neural network. *Procedia Computer Science*, 165:74–82, 2019.

[12] Andrew P Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

[13] François Castagnos, Martin Mihelich, and Charles Dognin. A simple log-based loss function for ordinal text classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4604–4609, 2022.

[14] Nitin Kumar Chauhan and Krishna Singh. A review on conventional machine learning vs deep learning. In *2018 International Conference on Computing, Power*

and Communication Technologies (GUCON), pages 347–352, 2018. doi: 10.1109/GUCON.2018.8675097.

[15] Mingcai Chen, Yuntao Du, Yi Zhang, Shuwei Qian, and Chongjun Wang. Semi-supervised learning with multi-head co-training. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36:6278–6286, 06 2022. doi: 10.1609/aaai.v36i6.20577.

[16] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[17] Hercules Dalianis. Evaluation metrics and evaluation. In *Clinical Text Mining: secondary use of electronic patient records*, pages 45–53. Springer, 2018.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL https://arxiv.org/abs/1810.04805.

[19] Joshua V Dillon, Ian Langmore, Dustin Tran, Eugene Brevdo, Srinivas Vasudevan, Dave Moore, Brian Patton, Alex Alemi, Matt Hoffman, and Rif A Saurous. Tensorflow distributions. *arXiv preprint arXiv:1711.10604*, 2017.

[20] Reinaldo Padilha França, Ana Carolina Borges Monteiro, Rangel Arthur, and Yuzo Iano. Chapter 3 - an overview of deep learning in big data, image, and signal processing in the modern digital age. In Vincenzo Piuri, Sandeep Raj, Angelo Genovese, and Rajshree Srivastava, editors, *Trends in Deep Learning Methodologies*, Hybrid Computational Intelligence for Pattern Analysis, pages 63–87. Academic Press, 2021. ISBN 978-0-12-822226-3. doi: https://doi.org/10.1016/B978-0-12-822226-3.00003-9. URL https://www.sciencedirect.com/science/article/pii/B9780128222263000039.

[21] Sakshini Hangloo and Bhavna Arora. Combating multimodal fake news on social media: methods, datasets, and future perspective. *Multimedia Systems*, 28:2391–2422, 07 2022. doi: 10.1007/s00530-022-00966-y.

[22] Charles R Harris, K Jarrod Millman, Stéfan J Van Der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al. Array programming with numpy. *nature*, 585(7825): 357–362, 2020.

[23] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention, 2021. URL https://arxiv.org/abs/2006.03654.

[24] Julia Hirschberg and Christopher Manning. Advances in natural language processing. *Science (New York, N.Y.)*, 349:261–266, 07 2015. doi: 10.1126/science.aaa8685.

[25] Ismail Hmeidi. Stock price prediction using k-nearest neighbor (knn) algorithm. 2013.

[26] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[27] Jin Huang and Charles X Ling. Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering*, 17(3):299–310, 2005.

[28] Zihan Huang, Charles Low, Mengqiu Teng, Hongyi Zhang, Daniel Ho, Mark Krass, and Matthias Grabmair. Context-aware legal citation recommendation using deep learning, 06 2021.

[29] Md. Ishraquzzaman, Mohammed Ashraful Islam Chowdhury, Shahreen Rahman, and Riasat Khan. Ensemble transformer–based detection of fake and ai–generated news. *Applied Computational Intelligence and Soft Computing*, 2025

(1):3268456, 2025. doi: https://doi.org/10.1155/acis/3268456. URL https://onlinelibrary.wiley.com/doi/abs/10.1155/acis/3268456.

[30] Sangita M Jaybhaye, Vivek Badade, Aryan Dodke, Apoorva Holkar, and Priyanka Lokhande. Fake news detection using lstm based deep learning approach. In *ITM web of conferences*, volume 56, page 03005. EDP Sciences, 2023.

[31] Reham Jehad and Suhad A. Yousif. Fake news classification using random forest and decision tree (j48). *Al-Nahrain Journal of Science*, 23(4):49–55, Nov. 2020. URL https://www.anjs.edu.iq/index.php/anjs/article/view/2306.

[32] Dinara Kaibassova, Bigul Mukhametzhanova, Dinara Tokseit, and Murat Kozhanov. Document analysis via combined vectorization and machine learning approaches. *International Journal of Innovative Research and Scientific Studies*, 8:2195–2204, 07 2025. doi: 10.53894/ijirss.v8i4.8356.

[33] Sajid Khan, Mehmoon Anwar, Huma Qayyum, Farooq Ali, and Marriam Nawaz. Fake news classification using machine learning: Count vectorizer and support vector machine. *Journal of Computing amp; Biomedical Informatics*, 4(01):54–63, Dec. 2022. doi: 10.56979/401/2022/85. URL https://jcbi.org/index.php/Main/article/view/85.

[34] Pravin P Kharat, Sanyam Sharma, Sayali Tambe, and Deepali Vora. Fake news detector: Fnd. *International Journal of Computer Applications*, 975:8887.

[35] Vladislav Kolev, Gerhard Weiss, and Gerasimos Spanakis. Foreal: Roberta model for fake news detection based on emotions. pages 429–440, 01 2022. doi: 10.5220/0010873900003116.

[36] Chanchal Kumar, Mani Bansal, Mohd Anas Khan, Vinay Kaushik, Md Arquam, and Abdulatif Alabdultif. Graph-augmented transformer ensemble framework for robust and scalable fake news detection in social media ecosystems. *Scientific Reports*, 2025.

[37] Manish Kumar Singh, Jawed Ahmed, Afshar Alam, Kamlesh Kamlesh, and Sachin Kumar. A comparative study of computational fake news detection on social media, 08 2022.

[38] Prosper Kuorsoh and Ebenezer Nkrumah. Analyzing the impact of false information on social media: Implications for society. *Journal of Trends in Arts and Humanities*, 2:3007–7370, 12 2025. doi: 10.61784/jtah3037.

[39] Tianle Li, Yushi Sun, Shang-Ling Hsu, Yanjia Li, and Raymond Wong. Fake news detection with heterogeneous transformer, 05 2022.

[40] Ximing Li, Yuanzhi Jiang, Changchun Li, Yiyuan Wang, and Jihong Ouyang. Learning with partial labels from semi-supervised perspective, 11 2022.

[41] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.

[42] Bradley C. Love. Comparing supervised and unsupervised category learning. *Psychonomic Bulletin amp; Review*, 9(4):829–835, December 2002. ISSN 1531-5320. doi: 10.3758/bf03196342. URL http://dx.doi.org/10.3758/BF03196342.

[43] Wes McKinney et al. pandas: a foundational python library for data analysis and statistics. *Python for high performance and scientific computing*, 14(9):1–9, 2011.

[44] Ibomoiye Mienye, Theo Swart, and George Obaido. Recurrent neural networks: A comprehensive review of architectures, variants, and applications. *Information*, 15:517, 08 2024. doi: 10.3390/info15090517.

[45] Saeed Mohsen. Recognition of human activity using gru deep learning algorithm. *Multimedia Tools and Applications*, 82, 05 2023. doi: 10.1007/s11042-023-15571-y.

[46] Nurul Nasarudin, Fatma Al Jasmi, Richard Sinnott, Nazar Zaki, Hany Alashwal, Elfadil Mohamed, and Mohd Mohamad. A review of deep learning mod-

els and online healthcare databases for electronic health records and their use for health prediction. *Artificial Intelligence Review*, 57, 08 2024. doi: 10.1007/s10462-024-10876-2.

[47] Jamal Nasir, Osama Khan, and Iraklis Varlamis. Fake news detection: A hybrid cnn-rnn based deep learning approach. *International Journal of Information Management Data Insights*, 1:100007, 04 2021. doi: 10.1016/j.jjimei.2020.100007.

[48] Ali Noorian, Ali Harounabadi, and M. Hazratifard. A sequential neural recommendation system exploiting bert and lstm on social media posts. *Complex Intelligent Systems*, 10, 08 2023. doi: 10.1007/s40747-023-01191-4.

[49] P Russel Norvig and S Artificial Intelligence. A modern approach. *Prentice Hall Upper Saddle River, NJ, USA: Rani, M., Nayak, R., & Vyas, OP (2015). An ontology-based adaptive personalized e-learning system, assisted by software agents on cloud storage. Knowledge-Based Systems*, 90:33–48, 2002.

[50] BM Patel and M Sule. Tokenization techniques in nlp: A comprehensive review. *International Journal of Advance Research and Innovative Ideas in Education*, 9 (1):1873–1892, 2023.

[51] Rajvardhan Patil, Sorio Boit, Venkat Gudivada, and Jagadeesh Nandigam. A survey of text representation and embedding techniques in nlp. *IEEE Access*, 11: 36120–36146, 2023. doi: 10.1109/ACCESS.2023.3266377.

[52] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[53] Omkar Reddy Polu. Ai-based fake news detection using nlp. *INTERNATIONAL JOURNAL OF ARTIFICIAL INTELLIGENCE MACHINE LEARNING*, 3:231–239, 10 2024. doi: 10.34218/IJAIML_03_02_019.

[54] Omkar Reddy Polu. Ai-based fake news detection using nlp. *Journal ID*, 9339: 1263, 2024.

[55] Sakshi Puranik, Suzan Khan Pathan, Ronak Patel, and Yogita Shelar. Fake news detection using passive aggressive and naïve bayes. 2022.

[56] Nabeel Raza, Said Jadid Abdulkadir, Yawar Abbas Abid, Sami S. Albouq, Ayed Alwadain, Asad Ur Rehman, Ebrahim Hamid Sumiea, and Muhammad Farhan. Enhancing fake news detection with transformer-based deep learning: A multi-disciplinary approach. *PLOS ONE*, 20(9):1–32, 09 2025. doi: 10.1371/journal.pone.0330954. URL https://doi.org/10.1371/journal.pone.0330954.

[57] Alexander Rush. The annotated transformer. In Eunjeong L. Park, Masato Hagiwara, Dmitrijs Milajevs, and Liling Tan, editors, *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 52–60, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-2509. URL https://aclanthology.org/W18-2509/.

[58] Hager Saleh, Abdullah Alharbi, and Saeed Hamood Alsamhi. Opcnn-fake: Optimized convolutional neural network for fake news detection. *IEEe Access*, 9: 129471–129489, 2021.

[59] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data*, 8:171–188, 06 2020. doi: 10.1089/big.2020.0062.

[60] Muhammad Umer, Zainab Imtiaz, Saleem Ullah, Arif Mehmood, Gyu Sang Choi, and Byung-Won On. Fake news stance detection using deep learning architecture (cnn-lstm). *IEEE Access*, 8:156695–156706, 2020. doi: 10.1109/ACCESS.2020.3019735.

[61] Greg Van Houdt, Carlos Mosquera, and Gonzalo Nápoles. A review on the long short-term memory model. *Artificial Intelligence Review*, 53(8):5929–5955, De-

cember 2020. ISSN 0269-2821, 1573-7462. doi: 10.1007/s10462-020-09838-1. URL https://link.springer.com/10.1007/s10462-020-09838-1.

[62] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

[63] Pawan Verma, Prateek Agrawal, Ivone Amorim, and Radu Prodan. Welfake: Word embedding over linguistic features for fake news detection. *IEEE Transactions on Computational Social Systems*, PP:1–13, 04 2021. doi: 10.1109/TCSS.2021.3068519.

[64] Liang Wu, Fred Morstatter, Kathleen M Carley, and Huan Liu. Misinformation in social media: definition, manipulation, and detection. *ACM SIGKDD explorations newsletter*, 21(2):80–90, 2019.

[65] Savvas Zannettou, Michael Sirivianos, Jeremy Blackburn, and Nicolas Kourtellis. The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans. *Journal of Data and Information Quality*, 11, 04 2018. doi: 10.1145/3309699.

[66] Xichen Zhang and Ali Ghorbani. An overview of online fake news: Characterization, detection, and discussion. *Information Processing Management*, 57, 03 2019. doi: 10.1016/j.ipm.2019.03.004.

[67] Xiaojin Jerry Zhu. Semi-supervised learning literature survey. 2005.

Appendix

الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire
وزارة التعليـــم العالـــي والبحـــث العلمـــــي
Ministère de l'Enseignement Supérieur et de La Recherche Scientifique

كلية العلوم والتكنولوجيا
قسم الرياضيات والإعلام الآلي

Faculté des Sciences et de la Technologie

Département Des Mathématiques & de l'Informatique

جامعـــة غردايـــة

الرقم .....2026/ ق.ر.ا /ك.ع.ت/ج.غ

شهادة الترخيص بالإيداع

أنا الأستاذ: ***Saidi Ahmed***

بصفتي رئيس [1]- والمسؤول عن تصحيح مذكرة تخرج ماستر الموسومة بـــــ

AI Powered Fake News Classification through Advanced NLP Techniques

من انجاز الطالبة: ***Bouharkat Lina Djihane***

والطالبة: ***Mochten Ayat Elrrahmane***

الكلية: العلوم والتكنولوجيا.
القسم: الرياضيات والاعلام الالي.
الشعبة: اعلام الي.
التخصص: الأنظمة الذكية لاستخراج المعارف.
تاريخ التقييم/المناقشة: **21/06/2026**
أشهد ان الطالب (الطلبة) قد قام (قاموا) بالتعديلات والتصحيحات المطلوبة من طرف لجنة المناقشة وان المطابقة بين النسخة الورقية والالكترونية استوفت جميع شروطها.

امضاء المسؤول عن التصحيح

مصادقة رئيس القسم

رئيس قسم الرياضيات والإعلام الآلي
الحاج موسى بلعيد

جامعة غرداية
قسم الرياضيات والاعلام الالي
كلية العلوم والتكنولوجيا

[1] شطب الصفة غير ملزمة