Advancing Cross-Domain Fake News Detection: Enhanced Models to Improve Generalization and Tackle the Class Imbalance Problem

by

Mohammad Qasim Mohammad Alnabhan

Thesis submitted to the University of Ottawa
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in
Digital Transformation and Innovation

Examining Committee

The following served on the Examining Committee for this thesis.

External Member: Álvaro Figueira, Assistant Professor
Department of Computer Science
Faculty of Sciences
University of Porto

Telfer Member: Bijan Raahemi, Professor
Department of Management Information Systems and Technologies
Telfer School of Management
University of Ottawa

Internal Member(s): Emil Petriu, Professor
School of Electrical Engineering and Computer Science
Faculty of Engineering
University of Ottawa

Internal Member(s): Diana Inkpen, Professor
School of Electrical Engineering and Computer Science
Faculty of Engineering
University of Ottawa

Supervisor(s): Paula Branco, Assistant Professor
School of Electrical Engineering and Computer Science
Faculty of Engineering
University of Ottawa

Declaration of Authorship

I hereby certify that this thesis is entirely my own original work except where otherwise indicated. I am aware of the University of Ottawa regulations concerning plagiarism, including those regarding consequent disciplinary actions. Any use of the works of any other author, in any form, is properly acknowledged at their point of use.

Abstract

The rapid proliferation of fake news across domains, such as politics, health, and social media, poses a significant threat to the integrity of information dissemination, leading to misinformation that can affect public perception and decision-making. Detecting fake news is critical to preserving the credibility of media sources, protecting democratic processes, and ensuring public safety. However, as Artificial Intelligence (AI)-driven Fake News Detection (FND) systems become more prevalent, ensuring AI safety is equally important to prevent unintended biases, ethical concerns, and adversarial exploitation of these models. Several challenges hinder the effectiveness of current FND models. Among these, cross-domain generalization and class imbalance are two critical problems that considerably impact the performance of detection systems. Although there are multiple challenges in FND, this thesis focuses on these two problems due to their widespread influence across various datasets and domains.

Cross-domain generalization is a complex problem because the characteristics of fake news differ significantly between domains. For example, fake news in politics may be presented in a completely different manner when compared to health-related misinformation. This variation in linguistic and contextual features makes it difficult for models trained in one domain to generalize to another. Still, important patterns can be present across domains and might be useful to detect fake news on different domains, helping with model generalizability. This thesis adopts a domain classification approach to address this issue, allowing the model to classify the domain before applying a domain-specific FND strategy, thereby improving overall generalization.

The class imbalance, where fake news is underrepresented in datasets compared to real news, presents another challenge, as models often become biased toward the majority class. This thesis addresses this problem by incorporating advanced Deep Learning (DL) techniques, such as Transfer Learning (TL) and class imbalance mitigation strategies, to ensure that models learn the nuances of detecting fake news even when instances are scarce.

The contributions of this thesis are categorized into the following areas: the first contribution is a comprehensive Systematic Literature Review (SLR) of the current DL techniques used in FND. This review highlights the challenges existing methods face, particularly with respect to cross-domain generalization and class imbalance, laying the foundation for developing novel approaches. The second contribution is developing a two-tiered, multi-domain FND framework. This framework integrates domain classification with FND, enhancing the model's ability to generalize across diverse domains, in addition to the mitigation of the imbalance issue by navigating various techniques. The third contribution is

the introduction of a fine-tuned, prompt-tuned Llama 2 model. This model leverages pre-trained knowledge from Large Language Models (LLMs) and adapts it to the specific task of FND, improving performance by focusing on domain-specific characteristics. This thesis has an extensive evaluation of the proposed approaches using publicly available datasets from various domains, including politics, health, science, crime, and social media. The results demonstrate that our methods significantly improve the robustness and accuracy of the FND models, providing benchmarks for future research in this field. Additionally, the findings emphasize the importance of AI safety in FND, ensuring that models are not only effective but also fair, explainable, and resistant to adversarial manipulation. These contributions advance the state-of-the-art of FND and establish a more resilient framework for detecting fake news across multiple domains, addressing critical challenges that have previously limited the generalization and reliability of detection systems.

Acknowledgements

This thesis marks the culmination of a journey filled with challenges, triumphs, and personal growth. Along the way, I have experienced moments of happiness, sadness, loss, and resilience. I have been fortunate to receive unwavering support and encouragement from incredible individuals, some close, some far away, and others who have left this world, leaving behind words that continue to inspire me.

First and foremost, I thank Allah for His guidance and strength, which have led me through this journey.

Special thanks to the unknown message sender who inspired me to embark on this path and gave me the strength to persevere even in the most challenging times.

I am deeply grateful to my supervisor, Professor Paula Branco, for her guidance, patience, and insightful feedback. Her encouragement has been instrumental in shaping this work.

To the thesis committee members, thank you for your constructive criticism and valuable suggestions from the proposal to the final submission of the thesis, which have greatly enriched this research.

To my colleagues and friends from all over the world, I thank you for your camaraderie, stimulating discussions, and moral support throughout this journey.

To my big family, I owe my deepest gratitude. To my father, whose belief in me was unparalleled and who longed to see this day but left too soon. Your memory inspires me every day. To my mother, the strongest and most supportive woman I know, thank you for always believing in me and celebrating every small achievement. To my siblings, who have stood by me with unwavering care, and to my beautiful little family, who have been my pillars of strength, this milestone is as much yours as mine.

This thesis is dedicated to you all. Thank you for being part of my journey.

Table of Contents

List of Tables xii

List of Figures xiii

Abbreviations xvi

1 Introduction 1

1.1 Overview . . . 1
1.2 Problem Statement . . . 2
1.3 Research Questions and General Methodology . . . 3
1.4 Contributions . . . 4
1.5 Thesis Organization . . . 5
1.6 Publications List . . . 6
1.7 Summary . . . 6

2 Background 7

2.1 Notion of Fake News and its Implications . . . 7
2.2 Fake News Related Concepts . . . 10
2.3 The notion of Transfer Learning . . . 12
2.4 The notion of Class Imbalance . . . 13
2.5 Traditional Machine and Deep Learning methods . . . 14

2.5.1 Traditional Machine Learning . . . 14
2.5.2 Deep Learning . . . 15
2.6 Summary . . . 20

3 Fake News Detection Using Deep Learning: A Systematic Literature Review 21
3.1 Introduction . . . 22
3.2 Research Methodology . . . 23
3.2.1 Search Strategy Overview . . . 23
3.2.2 Research Questions . . . 24
3.2.3 Source Databases and Search Query . . . 24
3.2.4 Inclusion Criteria . . . 26
3.2.5 Data Extraction . . . 26
3.2.6 Data Collection Summary . . . 27
3.3 Deep Learning algorithms used for fake news detection . . . 28
3.3.1 Architectures based on Convolutional Neural Networks . . . 30
3.3.2 Architectures based on Recurrent Neural Networks . . . 31
3.3.3 Architectures based on Graph Neural Networks . . . 37
3.3.4 Attention-based and BERT-based Architectures . . . 37
3.3.5 Architecture based on Feedforward Neural Networks . . . 39
3.3.6 Challenges related to deep learning methods for fake news detection 39
3.4 Datasets used for fake news detection . . . 41
3.4.1 Main datasets used for fake news detection . . . 41
3.4.2 Challenges related to the datasets used for fake news detection . . . 43
3.5 Strategies for transfer learning . . . 44
3.5.1 Transfer learning strategies applied to fake news detection . . . 44
3.5.2 Transfer learning challenges for fake news detection . . . 51
3.6 Strategies for dealing with imbalance . . . 51

3.6.1 The class imbalance problem in the context of fake news detection . 52
3.6.2 Challenges related to the class imbalance problem in fake new detection 54
3.7 Answers to Research Questions 56
3.8 Threats to Validity 59
3.9 Main gaps and open issues 59
3.10 Summary 61

4 Evaluating Deep Learning for Cross-Domains Fake News Detection 63
4.1 Introduction 64
4.2 Background and Literature Review 65
4.3 Research Methodology 66
4.3.1 Research Questions 66
4.3.2 Selected Datasets 67
4.3.3 Deep Learning Models 68
4.3.3.1 Bidirectional long short term memory (BiLSTM) 68
4.3.3.2 Hybrid Convolutional and Recurrent Neural Networks (CNN + RNN): 68
4.3.3.3 Convolutional Neural Networks (CNNs): 69
4.3.3.4 Hybrid CNN with LSTM (C-LSTM): 69
4.3.3.5 Bidirectional Encoder Representations from Transformers (BERT) 70
4.3.4 Experimental Settings 70
4.3.5 Computational Environment 71
4.3.6 Evaluation Metrics 71
4.4 Results and Discussions 72
4.4.1 Models' Performance Within the Same Dataset 72
4.4.2 Models' Performance Across Datasets 74
4.4.3 Research Questions' Answers 76
4.5 Summary 77

5 BERTGuard: Two-Tiered Multi-Domain Fake News Detection with Class Imbalance Mitigation **79**

5.1 Introduction . . . 80

5.2 Background and Related Work . . . 82

5.2.1 Fake News Datasets . . . 87

5.2.2 Dealing with Class Imbalance . . . 87

5.3 BERTGuard: Two-Tiered Multi-Domain Fake News Detection with Class Imbalance Mitigation . . . 88

5.4 Research Methodology . . . 92

5.4.1 Selected Datasets . . . 92

5.4.2 Experimental Settings . . . 95

5.4.3 Evaluation Metrics . . . 97

5.4.4 Computational Environment . . . 97

5.5 Results and Discussion . . . 98

5.5.1 Model Selection . . . 98

5.5.2 Single-Stage Fake News Detection Results . . . 100

5.5.3 Multi-Domain Fake News Approach . . . 101

5.5.4 Overall Impact of Class Imbalance Handling Strategies . . . 104

5.5.5 Impact of Balancing Datasets by Adjusting Weights . . . 105

5.5.6 Impact of Balancing Datasets by Upsampling . . . 108

5.5.7 Impact of Balancing Datasets by Downsampling . . . 109

5.6 Summary . . . 112

6 ProFineLlama: A Prompt and Fine-Tuned Transfer Learning Approach for Multi-Domain Fake News Detection **114**

6.1 Introduction . . . 114

6.2 Background and Literature Review . . . 118

6.2.1 Multi-Domain Fake News Detection . . . 118

6.2.2 Transfer Learning approaches for FND . . . 119

6.2.3 Prompt-Tuning Technique . . . 120
6.3 ProFineLlama: A Prompt and Fine-Tuned Transfer Learning Approach for Multi-Domain FND . . . 121
6.3.1 Computational Environment . . . 122
6.3.1.1 Hardware and Computing Resources . . . 123
6.4 Research Methodology . . . 123
6.4.1 Data Collection . . . 123
6.4.2 Experimental Settings . . . 124
6.4.3 Evaluation Metrics . . . 127
6.5 Results and Discussion . . . 127
6.5.1 Baselines Performance . . . 127
6.5.2 ProFineLlama Performance: Prompt-tuning the Fine-tuned Llama 2 7B . . . 129
6.5.3 Error Analysis . . . 131
6.5.4 Scalability and Deployment Considerations . . . 134
6.5.4.1 Computational Requirements . . . 134
6.5.4.2 Deployment Strategies . . . 135
6.5.5 Discussion of Different News Content . . . 135
6.6 Summary . . . 136

7 Conclusion and Future Work **137**
7.1 Conclusion . . . 137
7.2 Limitations . . . 138
7.3 Future Work . . . 139

References **142**

APPENDICES **180**

A Detailed Testing Results **181**
A.1 BERTGuard Detailed Results . . . 181

List of Tables

3.1 Main characteristics of the publicly available FND datasets. 42

4.1 Performance evaluation metrics. 72

5.1 The main transformer-based FND models. 84
5.2 Characteristics of the datasets used in our experiments. 93
5.3 Domain classification model results. 98
5.4 The average accuracies of various models on news domains. 99
5.5 Single-stage FND results—Baseline 1. 100
5.6 Overall results of Baseline 1, Baseline 2 (FakeBERT), and BERTGuard. . . 101
5.7 The proposed BERTGuard testing results. 102
5.8 The impact of class imbalance handling in BERTGuard. 105

6.1 Performance comparison between ProFineLlama and various baseline models. 128
6.2 Results per Dataset for Baseline 1, Baseline 2, and ProFineLlama 128
6.3 Examples of Misclassified Instances and Their Distribution 133

A.1 Testing Results on Snope 182
A.2 Testing Results on Covid-Claims 183
A.3 Testing Results on Pheme 184
A.4 Testing Results on Politifact 185
A.5 Testing Results on Climate 186
A.6 Testing Results on GossipCop 187

List of Figures

2.1 Transfer Learning Process 12
2.2 A Simple Neural Network 16

3.1 Search query used in our SLR. 25
3.2 PRISMA Chart of selecting and retrieving the articles 27
3.3 DL Models used for FND between the years of 2018 and 2023. 28
3.4 The General DL Framework that Used for FND 29
3.5 Taxonomy of the main Neural Networks (NN) categories used for FND. . . 30
3.6 Deep Learning Models for FND 30
3.7 An example of Convolutional Neural Network (CNN) architecture used in FND [287] 32
3.8 Recurrent Neural Network (RNN) utilization in FND. 33
3.9 An example of Gated Recurrent Unit (GRU) architecture used in FND [287]. 34
3.10 An example of CLSTM architecture used in FND [287] 35
3.11 An example of Bidirectional Encoder Representations from Transformers (BERT)-based architecture used in FND, FakeBERT 38
3.12 DL models in FND throughout the time. 40
3.13 Datasets used in the surveyed studies 42
3.14 Random undersampling 53
3.15 Random oversampling 53

4.1 A CNN-based architecture for FND. 69

4.2 A C-LSTM architecture for FND. 70
4.3 BiLSTM Performance 73
4.4 Hybrid CNN+RNN Performance 73
4.5 CLSTM Performance 73
4.6 CNN Performance 73
4.7 BERT Performance 74
4.8 Models Trained on Fake News 75
4.9 Models Trained on PHEME 75
4.10 Models Trained on Fake News Detection 75
4.11 Models Trained on ISOT 75
4.12 Models Trained on LIAR 76
4.13 Models Trained on FoR 76
4.14 Models Trained on FA-KES 76

5.1 BERTGuard FND architecture. 90
5.2 F1-scores for both baselines against BERTGuard. 104
5.3 The effect of handling the imbalance on the main baseline (Baseline 1) F1-scores by adjusting the class weights. 106
5.4 The effect of handling the imbalance on the BERTGuard F1-scores by adjusting the class weights. 106
5.5 F1-scores of both baselines and BERTGuard after treating the imbalance by adjusting class weights. 107
5.6 The effect of random oversampling on BERTGuard F1-scores. 108
5.7 F1-score comparison between upsampling and class weight adjustments on BERTGuard. 109
5.8 The effect of random undersampling on BERTGuard F1-scores. 110
5.9 The effect of random undersampling on Baseline 1 F1-scores. 111
5.10 The effect of using upsampling, downsampling, and class weight adjustment vs. no handling technique on BERTGuard. 111

6.1 ProFineLlama FND Architecture 121
6.2 F1-scores of the alternative baselines vs. the proposed ProFineLlama . . . 129
6.3 F1-scores of the Baselines vs. the Proposed ProFineLlama 131

Abbreviations

ADASYN Adaptive Synthetic Sampling 53

AI Artificial Intelligence iv, v, 3

ALBERT A lite version of BERT 19, 20, 83, 84, 91

ANN Artificial Neural Networks 14, 17, 39

BERT Bidirectional Encoder Representations from Transformers xiii, 2, 13, 15, 19, 20, 29, 36, 38, 39, 43, 49, 50, 54, 56, 57, 61, 66, 70, 71, 73–77, 79, 81–86, 88–92, 95, 97, 100, 112, 116, 118, 119, 121, 122, 124, 130

BiGRU Bidirectional Gated Recurrent Unit 33, 34, 43, 84

BiLSTM Bidirectional Long Short-Term Memory 4, 15, 29, 31, 32, 34, 36, 38, 39, 42, 43, 56, 57, 68, 71, 73, 74, 77, 84, 88, 92, 138

CNN Convolutional Neural Network xiii, 14–19, 28–32, 34, 36, 38, 39, 43, 48, 50, 54, 56, 57, 61, 66, 68, 69, 71, 73, 75, 77, 84, 85, 88, 92, 118, 124

DistilBERT Distilled version of BERT 83, 84, 90, 91, 96, 99, 116, 121, 122, 125

DL Deep Learning iv, xiii, 1–7, 11–15, 17–25, 28, 29, 31, 37, 39, 40, 43–46, 51, 55–57, 59–66, 68, 76–79, 82, 84, 85, 115, 118, 119, 122, 137, 138, 140

DNNs Deep Neural Networks 14, 16, 36, 39, 54

FND Fake News Detection iv, v, xii–xv, 1–6, 11, 13, 14, 20–26, 28–44, 49–52, 54–66, 69, 70, 75–92, 96, 97, 100, 101, 103–105, 109, 112–122, 124, 127, 129–133, 135–140

FNN Feedforward Neural Network 18, 28, 29, 39, 57

G-Mean Geometric-mean 71, 72, 100–103, 127, 129, 130

GAT Graph Attention Networks 37, 43, 84

GCN Graph Convolutional Networks 37

GloVe Global Vectors for Word Representation 18

GNN Graph Neural Network 29, 37, 43, 57, 66, 86

GPT Generative Pre-trained Transformer 49, 50

GRU Gated Recurrent Unit xiii, 15, 17, 18, 29, 31, 33, 34, 36, 37, 56, 57

KNN K-Nearest Neighbours 15, 18

LLMs Large Language Models v, 2, 3, 6, 115, 116, 120, 122, 123, 127, 137

LSTM Long Short-Term Memory 15, 17–19, 29, 31–34, 36–39, 54, 56, 57, 61, 68, 69, 71, 84, 92, 118

ML Machine Learning 5, 7, 12, 14, 15, 18–20, 23, 25, 44–46, 49, 51, 61, 66, 80, 85, 115

MLM named Masked Language Modeling 19

NLP Natural Language Processing 14, 15, 19, 21, 49, 51, 61, 64, 82, 83, 115, 134

NN Neural Networks xiii, 15, 16, 18, 28–30, 39, 45, 47, 66, 87

NSP Next Sentence Prediction 19, 20

RL Reinforcement Learning 84, 87

RNN Recurrent Neural Network xiii, 14, 15, 17, 18, 29, 31–34, 37, 43, 54, 56, 57, 61, 66, 68, 71, 73, 75, 77, 84, 92, 118

RoBERTa Robustly Optimized BERT Pretraining Approach 19, 50, 82, 84, 91, 119

SFND Single-Domain Fake News Detection 80

SGT Sequence Graph Transform 37

SLR Systematic Literature Review iv, xiii, 5, 20, 21, 23–26, 59, 61, 63, 137

SMOTE Synthetic Minority Oversampling Technique 53, 54, 58, 88

SVM Support Vector Machine 14, 15, 53

TL Transfer Learning iv, 2–7, 12, 13, 20–24, 27, 45–51, 57, 58, 60, 61, 77, 114, 117–119, 124, 128, 136–138, 140

USA United States of America 8, 22, 80, 85, 115

XGBoost eXtreme Gradient Boosting 15

Chapter 1

Introduction

In this chapter, we introduce the thesis topic and objectives, outline the problem statement addressed in this thesis, outline the research questions and goals, highlight the contributions and publications that fill gaps in the current literature and resolve existing issues, and conclude with the thesis outline.

1.1 Overview

The rapid rise of social media and online platforms has revolutionized how information is disseminated, but it has also enabled the rapid spread of fake news. Fake news can distort public perception and influence critical areas such as political outcomes, health decisions, and social stability. The challenge of detecting fake news lies in its diverse nature across different domains (politics, health, science) and the imbalance between fake and true news articles. This makes it difficult for machine learning models to generalize effectively, especially when faced with content from new or unseen domains.

While many DL models have been applied in FND, they typically perform well within specific datasets but struggle to generalize when tested across domains. Moreover, due to the imbalance in datasets where fake news is often outnumbered by true news, models tend to skew their predictions toward the majority class, reducing their effectiveness in identifying false information.

This thesis aims to address these core challenges by proposing innovative detection solutions that improve cross-domain generalization and mitigate class imbalance, ultimately leading to more reliable FND systems in diverse contexts.

1.2 Problem Statement

In this thesis, we tackle two key challenges of FND:

1. Cross-Domain Generalization: Existing FND models struggle to generalize effectively across domains [25]. For example, models trained on political news may fail when tested on health-related misinformation. This lack of robustness is a significant hurdle in real-world applications, where fake news spans multiple domains and topics [25]. The importance of cross-domain generalization comes from the fact that fake news does not follow a fixed structure or linguistic pattern, and fake news varies significantly depending on the domain. A model trained solely on one domain may not recognize deception in another, limiting its real-world applicability. Since new topics and narratives constantly emerge, FND models must be adaptable to unseen domains to remain effective. Ensuring cross-domain generalization is crucial for building scalable and reliable detection systems that can handle fake news across diverse sources and contexts.

2. Class Imbalance: Fake news datasets often suffer from a class imbalance, where fake news articles are vastly outnumbered by true news articles [6, 25]. This imbalance causes models to favor the majority class (true news), leading to poor detection of the minority class (fake news), which is the primary focus of detection systems.

While traditional DL models have made progress in addressing these issues, they often fall short in multi-domain environments. To overcome this, transformer-based architectures such as BERT contribute to this process by providing a robust mechanism for contextual understanding in text analysis tasks. Furthermore, LLMs, such as Llama 2, take this further by leveraging their extensive pre-training on vast and diverse datasets. Through TL, these LLMs can be fine-tuned for domain-specific FND tasks, enabling them to better capture linguistic nuances and contextual variations across multiple domains. However, the challenge remains to fully leverage these models to enhance cross-domain detection and address class imbalance effectively.

This thesis tackles these issues by using LLMs for TL, combining fine-tuning and prompt-tuning techniques to enhance model adaptability across different domains. Additionally, we introduce methods to mitigate class imbalance and improve the model's ability to detect underrepresented fake news instances.

This research is inherently interdisciplinary, aligning with the Digital Transformation and Innovation (DTI) program by addressing both technological and real-world challenges

associated with fake news. Although this thesis focuses mainly on DL models to enhance FND, this constitutes an applied research thesis, as the problem being tackled extends beyond computer science into domains such as digital media, policy making, and public trust in information systems.

Fake news has significant social, psychological, and economic consequences, influencing public perception, elections, healthcare decisions, and financial markets. The rapid spread of fake news can contribute to public panic, polarization, and economic instability, making its detection not just a technical challenge but a societal necessity.

The increasing reliance on AI-driven FND highlights the need for cross-domain solutions that go beyond purely computational methods. Digital transformation involves not only advancing technology, but also ensuring its adaptability, ethical application, and real-world deployment. By tackling key challenges such as cross-domain generalization and class imbalance, this research contributes to the broader innovation landscape in AI-driven FND, making it highly relevant to the DTI context.

1.3 Research Questions and General Methodology

This research is driven by the following questions:

- RQ1: How can we change DL models to improve their generalizability across multiple fake news domains?
- RQ2: What techniques can effectively mitigate the class imbalance problem in FND?
- RQ3: How can TL and LLMs be leveraged to enhance FND across multiple domains?
- RQ4: What are the trade-offs between different approaches (e.g., fine-tuning, prompt-tuning) to improve FND performance?

To address these questions, this thesis aims to achieve the following research goals:

1. Conduct a comprehensive literature review to identify gaps in current FND models.
2. Compare different strategies for FND, analyzing their effectiveness, scalability, and applicability.
3. Develop and evaluate a two-tiered multi-domain FND framework that integrates domain classification and class imbalance mitigation strategies.

4. Introduce a more generalized and novel TL approach that utilizes fine-tuned and prompt-tuned Llama 2 models to enhance FND performance.

This research follows the Design Science Research (DSR) methodology [341], which emphasizes designing, refining, and evaluating artifacts to solve FND. The thesis presents two primary artifacts, BERTGuard and ProFineLlama, developed iteratively to enhanceFND. The research follows these key DSR stages:

- Problem Identification and Motivation: Defining the challenges in FND, such as domain shifts and class imbalance.
- Objectives of the Solution: Establish performance benchmarks for improved FND models.
- Artifact Development: Designing BERTGuard and ProFineLlama to enhance generalization and mitigation of class imbalances.
- Demonstration and Evaluation: Testing the models across multiple datasets and comparing them with baselines.
- Communication: Presenting findings to contribute to both academic research and real-world applications.

1.4 Contributions

This thesis makes the following four key contributions:

1. **Systematic Literature Review**: We conduct a comprehensive review of the existing literature on FND, highlighting the strengths and limitations of various DL models, TL techniques, and strategies to address class imbalance. The review sets the stage for the novel contributions presented in later chapters.
2. **Evaluation of Cross-Domain Models**: Various DL architectures, such as BERT and Bidirectional Long Short-Term Memory (BiLSTM), are evaluated for their performance in cross-domain FND. We assess how well these models perform across multiple datasets and highlight the limitations of current approaches, particularly in terms of their inability to generalize to new domains.

3. **Two-Tiered Multi-Domain Framework**: We introduce, BERTGuard, a novel two-tiered FND system designed to improve cross-domain generalization and mitigate class imbalance. This approach leverages feature extraction combined with imbalance handling techniques to deliver superior results across multiple domains.

4. **Transfer Learning with Prompt-Tuning and Fine-Tuning**: We present, ProFineLlama, a refined TL strategy that incorporates prompt-tuning and fine-tuning of Llama 2 models. This approach demonstrates significant improvements in FND, particularly in tackling the class imbalance problem.

1.5 Thesis Organization

The thesis is structured as follows:

- Chapter 2: Background on FND, Machine Learning (ML), DL, TL and class imbalance techniques.
- Chapter 3: An SLR of FND methods and key challenges in the field.
- Chapter 4: Evaluation of various DL models for cross-domain FND.
- Chapter 5: Proposal of a two-tiered multi-domain detection system addressing class imbalance and domain generalization, BERTGuard.
- Chapter 6: Development of a framework based on TL using prompt tuning and fine tuning techniques, ProFineLlama.
- Chapter 7: Conclusion, limitations, and future directions for improving FND models.

Each chapter in this thesis corresponds to a standalone research paper published in peer-reviewed journals and conferences. As a result, some repetition is inevitable, particularly in literature reviews and methodology descriptions. However, the structuring of this thesis as a collection of individual studies enhances its contribution by ensuring that each section is self-contained and directly accessible to researchers focused on specific aspects of FND.

1.6 Publications List

The research conducted during this thesis has contributed significantly to the field of FND, particularly in addressing the challenges of cross-domain generalization and class imbalance using DL techniques. The following peer-reviewed publications, chronologically ordered, present the results and innovations developed as part of this thesis:

1. M. Q. Alnabhan and P. Branco, "Evaluating Deep Learning for Cross-Domains Fake News Detection," in Foundations and Practice of Security. FPS 2023, Lecture Notes in Computer Science, vol. 14552, Springer, Cham, 2024. doi: 10.1007/978-3-031-57540-2_4

2. M. Q. Alnabhan and P. Branco, "Fake News Detection Using Deep Learning: A Systematic Literature Review," IEEE Access, vol. 12, pp. 114435-114459, 2024. doi: 10.1109/ACCESS.2024.3435497

3. M. Q. Alnabhan and P. Branco, "BERTGuard: Two-Tiered Multi-Domain Fake News Detection with Class Imbalance Mitigation," Big Data and Cognitive Computing, vol. 8, no. 8, p. 93, 2024. doi: 10.3390/bdcc8080093

4. M. Q. Alnabhan and P. Branco, "ProFineLlama: A Prompt and Fine-Tuned Transfer Learning Approach for Multi-Domain Fake News Detection," in Foundations and Practice of Security. FPS 2024, Lecture Notes in Computer Science, Springer, Cham, 2024.

1.7 Summary

In summary, this thesis addresses the core challenges of cross-domain FND and class imbalance through the use of advanced DL methods and LLMs. It proposes innovative solutions that lead to significant improvements in detection accuracy across a variety of domains. Throughout this thesis, several key contributions were made to advance the field of FND using DL techniques. First, we present a comprehensive review of the state-of-the-art in FND. Second, in addressing cross-domain challenges, this thesis offers insights into handling domain shifts in FND systems. Third, the development of BERTGuard, a two-tiered multi-domain FND system, introduces a robust model designed for multi-domain detection tasks. Lastly, we present ProFineLlama, TL, and Prompt-Tuning for FND which explores innovative TL methods and prompt-tuning to improve detection accuracy across multiple domains.

Chapter 2

Background

This chapter provides background information for this thesis. We explain in Section 2.1 the notion of fake news and its impacts in different domains. Other terminology used interchangeably with fake news is discussed in Section 2.2. The background notions on TL and the class imbalance are provided in Section 2.3 and Section 2.4, respectively. In Section 2.5, a description of different ML and DL architectures is provided. Finally, Section2.6 summarizes the chapter.

2.1 Notion of Fake News and its Implications

Fake news can be defined differently in different contexts. For example, "fake news is fabricated information that mimics news media content" [18]. This definition emphasizes the appearance and format of fake news, highlighting its intention to imitate legitimate news content to gain credibility and deceive audiences. It is often applied in contexts where the focus is on the structural characteristics of fake news, such as the case in journalism or media analysis.

Another definition of fake news is "news articles that are intentionally and verifiably false and could mislead readers" [218]. This definition emphasizes the intentionality of the information. It is more aligned with psychological and behavioral research, where the focus is on how fake news impacts readers.

The ability of fake news to alter people's beliefs is becoming a major concern. It can have significant effects on many aspects, such as consumer decision-making [355], elections [24], education [266], and health [46]. The reasons behind the spread of fake

news can vary significantly depending on the context. While some individuals may share fake news out of a desire for community acceptance or to align with societal views, there are often more malicious intentions at play. For instance, fake news can be deliberately crafted to bias public opinion before elections, manipulate perceptions, or achieve political or economic goals.

Moreover, during the COVID-19 pandemic, fake news was frequently used to target specific groups, spreading misinformation with malicious intent, such as undermining trust in health authorities or promoting harmful agendas. These examples highlight that the spread of fake news is not limited to social conformity but can also stem from deliberate attempts to mislead, harm, or exploit individuals and communities.

The popularity of using social media made it possible to disseminate fake news at low to no cost. Moreover, it also enabled the rapid dissemination of messages with the assistance of various dissemination tools. These tools include social robots, trolls, and cyborgs [218]. Such social robots are created as computer algorithms to share content and interact with people in the online environment. These bots manipulate news and conversations to be shared with readers. This was clear during the 2016 US presidential election when statistics showed that more than 19% of online conversations were found to have manipulated the information of the presidential election results to spread false news and disrupt online communities [49].

Fake news is considered one of the greatest threats in the world due to its ability to be found in many domains. It was recorded as a multi-disciplinary threat that can exist in various domains such as social, economic, commerce, health, journalism, and democracy all over the world, with huge direct damages [165]. Financially, United States of America (USA) governmental financial reports showed that about $130 billion loss in the stock market was the direct result of fake news that reported that USA president Barak Obama got injured in an explosion [259]. On the commerce and health levels, fake news was the reason behind the sudden shortage of salt in Chinese supermarkets after claiming that iodized salt would help counteract the effects of radiation after the Fukushima nuclear leak in Japan [344]. In addition, an escalation of tensions between India and Pakistan began with the fake reporting of the Balakot strike and resulted in the deaths of military personnel and the loss of expensive military equipment [230]. In another instance, during the 2023 wildfire crisis in Canada, fake news circulated widely on social media, spreading false claims about evacuation orders and fire containment measures [308]. This confusion led to delays in evacuation efforts, putting lives at risk and impeding emergency responses. Such incidents highlight how fake news during critical events exacerbates public safety risks and undermines trust in official communications.

By obtaining this fake information, we cannot make good decisions regarding worldwide matters or even personal decisions. Researchers categorized the primary impacts of fake news into four groups as follows:

- *Impact on People Psychology and Social Life*: A specific person or group of people may be harassed, insulted, and threatened by fake news in social media, which may lead to serious life consequences such as committing suicide and leading to a high level of depression [130]. Social media readers must not rely on this invalid news when judging others or building relationships through social media.

- *Impact on Health*: The internet has become the main source of health-related information. Fake health news has a serious impact on people's lives and existence [353]. The impact of fake news on health has had an incredible impact over the last year [39]. The pressure from doctors, health agencies, health advocates, and lawmakers has forced social media platforms to change their policies to ban fake news publishers and limit the spread of health fake news [46]. During the recent COVID-19 pandemic, social media platforms have experienced a remarkable increase in usage as primary health resources. Recent statistics indicate that the number of active users on social media has surpassed 3.8 billion [123]. People are now getting more involved in these platforms, especially on Facebook, Twitter, Instagram, etc., and expressing their thoughts, news, opinions, and information related to COVID-19. The sources of this COVID-19 gathered information are other news media or social media platforms. This information is being shared without verifying and fact-checking it. As a result, this fake news has scared people and affected their mental health in addition to the panic it shared. This fake news has also affected people's behaviours toward the validity of the virus and the vaccination side effects [130].

- *Financial Impact*: Fake news is currently a crucial problem in industries and the business world. Fake news is being disseminated by dishonest businessmen in the form of reviews that will raise their profits. This fake news is the reason for the drop in stock prices. This fake information can also destroy the reputation of a business and ruin its fame. The fake reviews will also create an unethical business mentality and impact the overall customer experience [130].

- *Democratic and political impact*: Research and scientists have investigated the significant effect of fake news on political rights after the vital role that fake news played in the American presidential election [104]. This was a major drop in political and democratic life. It has been argued that without fake news targeting the opponents of former President Donald Trump, he would have never won the election [104].

2.2 Fake News Related Concepts

The term "fake news" refers to a broad range of concepts and is frequently used in conjunction with the widespread practice of disseminating misleading information. In the related literature, misleading information is frequently referred to in various forms, including misinformation, disinformation, rumors, satire, and clickbait [230]. These terms can be used interchangeably; however, there are some differences in terms of the purpose of disseminating this news and the context of using it as follows:

- *Misinformation* is false information that is inaccurate or misleading information [350] [17]. There is no intention behind sharing this kind of false news, and it usually results from an honest mistake [116] or knowledge update without the purpose of misleading.

- *Disinformation* is false information that is controlled by a defined reason and intention (e.g., to deceive people [181] and/or to promote a biased agenda [340]) of sharing this misleading information. Disinformation is close to a narrow definition of fake news. In contrast to misinformation, it spreads because of a deliberate attempt to deceive or mislead others [116]. The survey conducted by Guo et al. [105] differentiates between misinformation and disinformation by using the term “deception” to highlight disinformation and fake news but not misinformation [373].

- *Rumor* is an early stage of the information disseminated from one person to another and is unverified and relevant information that is circulated. This kind of information can later be confirmed as true, fake news, or left unconfirmed [243] [55]. The concept of “rumor detection” was used to classify false news before the term “fake news” became popular [376].

- *Hoaxes* are information intentionally created and presented as the truth [181]. Because it involves rather complicated and substantial fabrications, it frequently results in serious material damage to the target [267].

- *Satire*, is meant to amuse or critique the readers and frequently uses irony and humor [53]. Some websites, such as SatireWire.com [207] and The Onion [73], routinely publish these news pieces. If parody news is widely disseminated regardless of the context, it might be damaging.

- *Hyperpartisan news* is news that, in a political context, is incredibly skewed or one-sided [246]. Biased in and of itself does not imply being fake, but according to certain publications [117] [368] it has a high likelihood of being false in hyperpartisan news

networks and is extensively disseminated in the alt-right community on sites like 4chan [74] and Gab [75].

- *Propaganda*: Through the controlled dissemination of one-sided messages, propaganda is a type of convincing that seeks to affect the feelings, attitudes, opinions, and behaviours of certain target audiences for political, ideological, and religious reasons [193]. In the recent past, it has frequently been employed to sway political opinion and election outcomes.

- *Spam* is made up of fake information, ranging from self-promotion to fake product announcements. Spam on review sites offers unreasonably good ratings to unfairly promote products or unjustified bad reviews to harm the reputation of rival products [140]. Spam in the social ecosystem targets users to spread malware and promote affiliate websites [185].

- *Clickbait* is a tale with a catchy headline written to draw readers and earn from advertising [340] [69]. Clickbait is one of the less severe forms of false information because of the difference between the headline and the substance, which is the major element of clickbait.

From the existing literature, we observe that many researchers aimed to define fake news using some or all of the aforementioned definitions interchangeably (e.g., in [373], [351], [105]). In addition, there are no formal definitions for these terms, similar to the term "fake news", which is why they are being used interchangeably. For the context of this thesis, we will consider the definition of fake news as false information shared with a specific purpose, such as deceiving people or promoting a biased agenda. We may sometimes use other fake news synonyms, such as misinformation, interchangeably. It is worth mentioning that this thesis does not aim to detect or analyze disinformation, as that would require methods beyond text classification, such as source credibility analysis, intent detection, and network-based propagation tracking. Instead, this research focuses on automated FND using DL models, emphasizing cross-domain generalization, class imbalance handling, and model performance optimization. The proposed approaches assess the factual integrity of news content based on textual characteristics without evaluating the intent behind misinformation spread.

2.3 The notion of Transfer Learning

ML and DL techniques have been applied to numerous real-world applications. These methodologies typically assume that the input feature space and data distribution are consistent, as both the training and the testing data are drawn from the same domain [354]. However, this assumption may not hold in many real-world scenarios. In some cases, collecting training data can be expensive or difficult. Therefore, it becomes necessary to develop high-performance models trained on data that is more readily available from different domains. This approach is known as TL, as depicted in Figure 2.1.

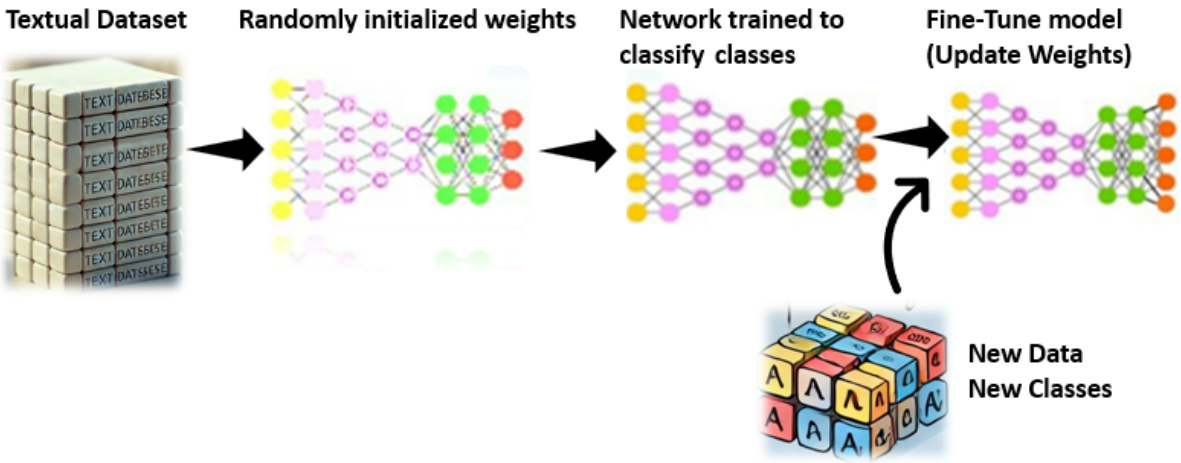

Figure 2.1: Transfer Learning Process

Many applications in which it is possible to extract patterns from historical information (training data) to forecast future outcomes have successfully and widely used data mining and ML [356]. Training and testing data for traditional ML typically share the same input feature space and data distribution. The performance of a predictive learner may suffer when the distribution of data between training and test sets differs [296]. In some circumstances, it can be challenging and expensive to find training data that match the test data's feature space and predicted data distribution properties. Consequently, a high-performance learner for a target domain that has been taught from a similar source domain is required. This is the motivation for using TL.

TL is a technique used to advance a learner from one domain by transferring knowledge from a related domain. Real-world, non-technical experiences can help us comprehend why TL is feasible. Take the case of two individuals who wish to learn how to play the piano.

One person has no prior musical training, whereas the other plays the guitar and has a wealth of musical expertise. By applying previously acquired musical information to the goal of learning to play the piano, a person with a strong musical background will be able to learn the piano more quickly and effectively [239]. One can employ knowledge from a task they have already mastered to help them learn a new related task.

Consider the challenge of predicting text sentiment of product reviews when there is an abundance of labeled data from digital camera reviews as a specific example from the domain of DL. DL approaches are employed to generate strong prediction results when the training and target data are derived from digital camera reviews. The prediction results, however, are likely to deteriorate due to the different domain data when the target data comes from food reviews and the training data comes from digital camera reviews. Digital camera reviews and food reviews still have several characteristics in common, if not exactly the same. They both convey opinions regarding purchased goods and are written in text form using almost the same vocabulary. Since the digital camera and food reviews are related domains, TL can be utilized to potentially improve the results of a target learner [239]. An alternative way to view the data domains in a TL environment is that the training data and the target data exist in different sub-domains linked by a high-level common domain. A musician domain, for instance, has subdomains for piano players and guitarists. In addition, reviews of food and digital cameras are subdomains of the review domain. How the subdomains are related is determined by the high-level common domain.

In FND, TL typically involves (1) pretraining on large datasets: models like BERT and Llama are pre-trained on vast corpora of general text data, allowing them to learn useful language patterns and representations; and (2) fine-tuning: after pretraining, the model is fine-tuned on a smaller dataset specific to the task at hand, such as FND, adapting the pre-trained knowledge to the new domain [84].

2.4 The notion of Class Imbalance

Another challenge in FND is class imbalance. In many fake news datasets, the number of true news articles vastly outnumbers the number of fake news articles. This imbalance can skew the model predictions toward the majority class (true news), making it difficult to accurately identify fake news.

To mitigate this issue, various strategies have been proposed, including: Oversampling: Increasing the number of fake news instances in the dataset by duplicating or generating new examples, Undersampling: Reducing the number of true news instances to balance

the dataset, and Cost-sensitive learning: Adjusting the model's loss function to penalize misclassifications of the minority class (fake news) more heavily, encouraging the model to focus on detecting fake news [27, 51].

In this thesis, various techniques are integrated into DL models to address the class imbalance problem and improve the performance of FND.

2.5 Traditional Machine and Deep Learning methods

DL models have seen an extraordinary evolution in recent years due to their promising success in many fields, including communication [236] and networking [5], computer vision [371], intelligent transportation [336], speech recognition [153], as well as Natural Language Processing (NLP) [169].

DL systems have advantages over traditional ML methods. DL is a subfield of ML strategies, which shows high precision and exactness in FND. ML methods are based on hand-crafted features, which may introduce biased features since the feature extraction task is considered a challenging and slow task [250]. They also work best with small datasets, which made them fail in the context of FND, which needs large datasets to train the DL models. In addition, ML approaches produce high-dimensional representations of linguistic information, resulting in the curse of dimensionality. The existing neural network-based models have outperformed the traditional models in terms of their performance due to their exceptional feature extraction capability [250]. In contrast, DL systems can acquire hidden representations from less complex inputs. The hidden features can be extracted from both the news content and context varieties. A study by Hiramath and Deshpande [118] showed that Deep Neural Networks (DNNs) need less computational time than other ML-based classification algorithms, such as Logistic Regression, Random Forest, Support Vector Machine (SVM), etc. However, DNNs use more memory. CNN and RNN are two highly utilized ideal architectures for DL that pioneered Artificial Neural Networks (ANN). We provide a brief description of ML and DL-based detection strategies in the following subsections.

2.5.1 Traditional Machine Learning

Traditional ML refers to the methods that have been used consistently for many years. These traditional ML methods are often used to solve classification [152], regression [96], clustering [88], etc. These traditional ML methods contain Logistic Regression, SVM,

Naïve Bayes, K-Nearest Neighbours (KNN), Decision Tree, Random Forest, and eXtreme Gradient Boosting (XGBoost).

The SVM is one of the most used types of algorithms in terms of classification problems. Creates a hyperplane or a set of hyperplanes in a high-dimensional space for the classification of the data point based on the set of features [76]. The number of features determines the dimensions of the hyperplane. For instance, the hyperplane will be exclusively a single line if there are two input features. On the other hand, the hyperplane will be a two-dimensional plane if there are three input features. Hyperplanes help in the classification of the data points. The main objective is to identify the optimal hyperplane that differentiates the data points with maximum margin (the maximum distance between the data points of both classes), as there could be many possibilities for a hyperplane to exist in an N-dimensional space.

The Naïve Bayes model is considered a simple probabilistic model [141]. NB is assumed from the Bayes theorem. It involves a conditional independence assumption. The KNN is a popular lazy learner [8]. In KNN, a data point is classified according to the class of the majority of the k closest data points. It is a non-parametric method used for classification and regression [29].

The Decision Tree is an algorithm that is supervised and can be used when dealing with both classification and regression problems [253]. A Decision Tree model can be broken down into several parts (branches) since a tree is composed of nodes. The outcome, which can be categorical or continuous, is represented through the leaf of the tree. The tree is built by an attribute selection method and split criteria. The attribute selection can be done in several ways, but most allocate a quality measure directly to the attribute and split. After the tree is grown, a pruning method is applied to simplify the tree structure and deal with the overfitting problem.

The Random Forest creates output predictions by merging the outcome sequence of multiple decision trees [52]. All trees in the forest have the same distribution and are built separately using a random vector selected from the input data. As such, a tree is made up of nodes that consist of a combination of features or randomly selected features.

2.5.2 Deep Learning

DL is the most recent type of ML being used and explored. When dealing with NLP tasks, DL is becoming the most preferred method to be used [169]. The frequently used DL methods include NN, CNN, RNN with both Long Short-Term Memory (LSTM) and BiLSTM, GRU, and BERT. We will now briefly describe these methods.

A NN is a highly complex system composed of multiple layers of interconnected nodes, known as neurons, which process and transform data through mathematical operations. Inspired by the structure of the human brain, NNs are designed to recognize patterns and learn hierarchical representations of data [152]. At a fundamental level, a NN consists of an input layer, one or more hidden layers, and an output layer. Each neuron in a layer receives inputs, applies a weighted transformation, passes the result through a non-linear activation function, and forwards the output to the next layer. Through backpropagation and gradient descent, the network adjusts its weights iteratively to minimize the error in predictions, making it capable of learning complex relationships within data [294]. These neurons are linked with the next layer of neurons. Although some linkages have a higher weight, all inputs are multiplied with the weight in which the total value is processed in the activation function that is in the hidden layer neurons. At this stage, the neuron will either be activated or shut down. The output layer is then in charge of producing the outcome. Figure 2.2 represents the function of the NN method.

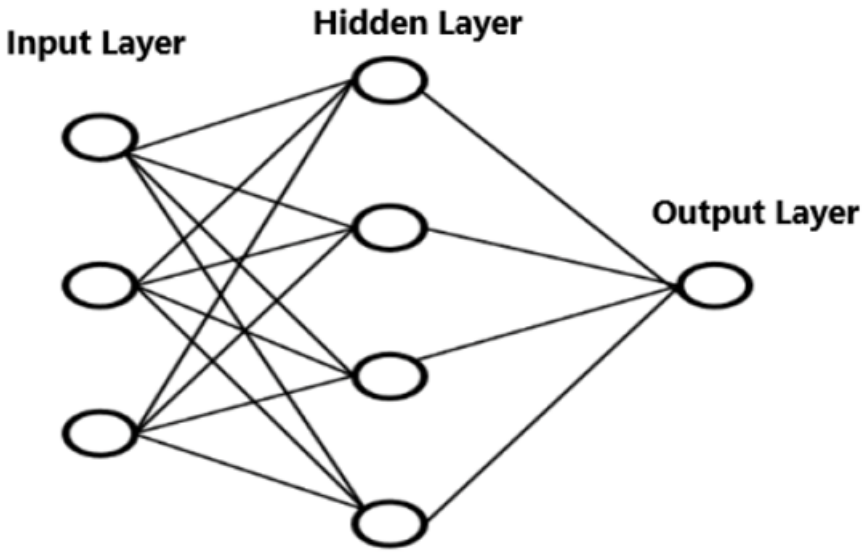

Figure 2.2: A Simple Neural Network

The CNN is a class of DNNs that is designed to process data that has a grid-like structure, such as an image [71]. The CNN contains a series of layers: an input layer, a convolutional layer consisting of various filters, pooling layers, and fully connected layers.

The input layer represents the pixel matrix of the image. It will then extract and convert the text input into a matrix form, also known as a word embedding vector, ensuring that each vector is at a fixed size.

As the name suggests, the CNN contains filters that execute the convolution task on the input. This layer is a complex process in which a mathematical operation is conducted between two inputs, for instance, between the matrix layer and the convolution filter. The convolution filter glides over the input matrix layer to extract a dot product, which is located between the filter and the input matrix. Once the matrix of the input layer passes through this layer, a feature map with a single column is obtained as the output [71]. Upon the conclusion of this process, a bias value is incorporated into the acquired output matrix. The pooling layer performs dimensionality reduction of its input feature vectors [71].

This is achieved by sub-sampling onto the convolutional layer's output vectors in collaboration with near elements. Although many types of pooling can operate in the pooling layer, max pooling appears to be the most frequently used. Max pooling functions by extracting the greatest value from the feature map within the convolutional layer. Average pooling, amongst others, is one of several types of pooling that can be used in this layer. The final layer found within the CNN structure is the Fully Connected layer. This layer consists of a minimum of one layer. The output information from the pooling layer must be flattened before being processed in the Fully Connected layer. The softmax or sigmoid activation functions determine the final output in the Fully Connected layer. Each activation function serves a different purpose. The softmax activation function is often used for a multi-class classification problem and the sigmoid activation function is used in the binary classification task [250].

Another DL method is the RNN, which is a class of ANN that employs sequential information in the network, which is important in the application where the implanted structure in the data sequence carries beneficial knowledge [23]. When looking at the traditionally used neural networks, one would notice that all inputs are treated separately from one another. This will be deemed impractical in instances where anticipating the following word in each sentence is the target, as the preceding words must be considered. As such, the RNN network became a preferred choice as it resolves the issue found with the traditional RNN processes, each word in a sentence sequence the same by considering its preceding words. Once each word is separately processed until the entire sequence is complete, the output will be given. The RNN can then use the information collected to predict other word sequences. When considering the basic RNN model, one must be aware of the difficulties the model has with long-term dependencies, which in turn causes a vanishing gradient problem. This can be solved by using two variants of RNN, such as the LSTM [120] and GRU [70]. LSTM features feedback connections as opposed to typical

Feedforward Neural Network (FNN). Such a RNN may analyze whole data sequences (such as speech or video) in addition to single data points (such as photos). A cell, an input gate, an output gate, and a forget gate compose a typical LSTM unit. The three gates control the flow of information into and out of the cell, while the cell holds values over arbitrary time intervals. Given that there may be lags of uncertain length between significant events in a time series, LSTM networks are well-suited to categorizing, analyzing, and making predictions based on time series data [120]. It is used to solve the issue of vanishing gradients that can arise when training conventional RNN. The weights of a NN are updated via gradients. The gradient reduces as it travels through time, which is known as the "vanishing gradient problem." A gradient value doesn't add much learning if it gets very small. The vanishing problem can also be solved by using GRU. The GRU is a more recent version of RNN and resembles an LSTM in many ways. GRU got rid of the cell state and transferred information using the hidden state. A reset gate and an update gate are the only gates it possesses. GRU train slightly more quickly than LSTM because they have fewer tensor operations.

Many researchers investigated various DL models for detecting fake news in social media. Nasir et al. [230] developed a hybrid model that integrates a CNN and LSTM. The CNN extracts the features, while LSTM captures the long-term dependencies. The approach uses Global Vectors for Word Representation (GloVe), which are pre-trained word embeddings that represent words as vectors. This methodology surpassed seven established ML techniques in terms of performance: Logistic Regression, Random Forest, Multinomial Naïve Bayes, KNN, Decision Tree, CNN, and RNN.

Kaliyar et al. [148] developed a DL network based on pre-trained GloVe word embeddings called FNDNet. Three convolutional pooling layers gathered characteristics from word embedding vectors, which were then classified using additional convolutional pooling and dense layers.

Saleh et al. [274] proposed an optimized DL approach, OPCNN-FAKE, based on CNN architecture. They assessed its performance against RNN, traditional ML approaches, and LSTM. The model includes an embedding layer that generates embedding vectors, a dropout layer for improved regularization, a convolution-pooling layer for extracting and reducing features, a flattened layer that produces a one-dimensional vector, and an output layer that decides whether the input text is fake or real based on the previous layer's output.

Yang et al. [365] designed another CNN-based approach (TI-CNN) that detects fake news by integrating obvious and hidden features from text and picture data. The authors evaluated their approach to various models such as LSTM, GRU, and CNN. Similarly, Raj

et al. [255] developed a CNN approach that uses text and image data to categorize internet news. However, these methods may fail to capture long-term contextual information. Furthermore, word embedding does not capture context-specific information inside the text.

Hashmi et al. [111] proposed a thorough approach to detecting fake news utilizing three publicly available datasets: WELFake, FakeNewsNet, and FakeNewsPrediction. They combined FastText word embeddings with a range of ML and DL methods, improving these algorithms through regularization and hyperparameter optimization to reduce overfitting and boost model generalization. Significantly, a hybrid model that integrated CNN and LSTM and was enhanced with FastText embeddings surpassed other methods in terms of classification performance across all datasets, achieving accuracy and F1-score of 99%, 97%, and 99% in WELFake, FakeNewsPrediction, and FakeNewsNet, respectively.

In addition, the BERT is a transformer-based DL model that is widely used for NLP tasks [79]. This model is an unsupervised pre-training of deep bidirectional representations within the text, which can represent the text from any direction (left or right) in all its layers [79]. This model can represent any token in its bidirectional property as it is an encoder-only transformer. It consists of two sizes when discussing its pre-trained model: the BERTBASE and the BERTLARGE. The difference between each is that the BERTLARGE uses more parameters than BERTBASE. This model utilizes an input representation, allowing it to display a single and a pair of sentences in one token sequence. The CLS is a classification token used for a classification task that is found at the beginning of each input sentence. When looking at single-sentence tasks, a Word Piece token and a separator token (SEP) are placed after the CLS token. In contrast, a pair of sentences is combined into a single sequence separated by a SEP, or an extra sentence is included in each token to indicate which sentence it belongs to. Furthermore, a positional embedding is included in each token to indicate the position of the sequence. BERT has been pre-trained in an unsupervised way using two tasks named Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) which operate slightly differently from one another [79]. When presented with a sequence of tokens, MLM aims to predict the masked tokens. On the other hand, NSP aims to predict if sentence number two is coherent when matched with sentence number one.

BERT has many variants, such as XLNet [366], Robustly Optimized BERT Pretraining Approach (RoBERTa) [190], and A lite version of BERT (ALBERT) [183] to deal with different problems. The XLNet architecture uses the bidirectional context of BERT without masking. It combines the bidirectional property of BERT with autoregressive language modelling of Transformer-XL. RoBERTa is trained on more data and has longer training than BERT. The masking pattern will be changed in this variant, and the NSP procedure of

BERT is removed here. The third variant of BERT is ALBERT, which uses parameter reduction techniques to increase the training speed. ALBERT uses sentence order prediction instead of NSP which solves the limitation of inter-sentence coherence in ALBERT.

2.6 Summary

In this chapter, we provided a comprehensive background on the concept of fake news and its implications across various domains. It begins by defining fake news and discussing its significant impact on society, including its effects on public perception, decision-making, and trust in media. The chapter then explores related concepts such as misinformation, disinformation, rumors, hoaxes, satire, hyperpartisan news, propaganda, spam, and clickbait, highlighting the differences and similarities among these terms.

The chapter also delves into the notion of TL, explaining how it can be used to improve the performance of ML models by leveraging knowledge from related domains. This section discussed the benefits of TL, particularly in scenarios where collecting training data is challenging or expensive.

Another critical topic covered is the class imbalance problem, which occurs when the number of instances in one class significantly outnumbers those in another. The chapter outlined various strategies to address this issue, including oversampling, undersampling, and cost-sensitive learning.

Finally, the chapter provided an overview of traditional ML and DL methods used for FND. It compares the advantages and limitations of these approaches, emphasizing the superior performance of DL models in handling large datasets and extracting complex features.

In the next chapter, we will build a SLR to comprehensively study fake news detection using DL. We will discuss the used DL models and datasets in the detection process. We also discuss the TL concept, its mechanisms, and challenges in FND. The SLR will finally detail the strategies for dealing with class imbalance and the challenges concerning FND.

Chapter 3

Fake News Detection Using Deep Learning: A Systematic Literature Review

This chapter provides a comprehensive SLR on the use of DL methods for the detection of fake news. With the increasing prevalence of fake news on social media platforms, researchers have turned to advanced DL algorithms to improve the accuracy and efficiency of detecting false information. The purpose of this chapter is to synthesize existing studies, explore the strengths and limitations of current methodologies, and identify gaps that require further research.

In this SLR, we investigate existing FND strategies that use DL. We focus on publicly available datasets used in FND and their NLP approaches. We aim to gather information about the TL techniques applied and the methods used for addressing class imbalance, to examine their effect on detection accuracy. Our survey aims to identify open issues and research gaps in current studies. To the best of our knowledge, we are the first to provide a comprehensive SLR that investigates the effects of TL and class imbalance treatment in the FND domain.

This chapter is based on the following published paper:

- M. Q. Alnabhan and P. Branco, "Fake News Detection Using Deep Learning: A Systematic Literature Review," in IEEE Access, vol. 12, pp. 114435-114459, 2024, doi: 10.1109/ACCESS.2024.3435497.

3.1 Introduction

Due to a greater interest in the use of the internet, the spread of fake news has become more common than ever before. Before the popularity of social media platforms, fake news was less common and much more difficult to spread to a vast amount of people, as it was achieved either through word of mouth or through printed media. Fake news can be defined as the phenomenon that occurs when incorrect information is purposefully spread throughout social media outlets with a significant ability to convince the reader of the written content [24]. Today, anyone can publish content without regulation or scrutiny. Several social media platforms, such as Facebook and Twitter, serve as a means of spreading fake news, as people and influencers use them to share their opinions, videos, and various activities [91, 130].

Fake news increased significantly in 2016 during the period preceding the USA presidential election [291]. As such, fake news on social media networks has captured the attention of many researchers. Recently, detecting fake news has become an emerging area of interest for many researchers, such as [291] and [357]. However, detecting fake news is a complicated task that requires the use of complex models to compare related or unrelated information with known truthful information [358]. Furthermore, fake news is perceived by researchers in several ways, leading to multiple ways of addressing and solving this issue. Some terms related to misinformation are used interchangeably in multiple cases. These terms include fake news, rumors, spam, and disinformation that usually contain numerical, categorical, textual, and image contents [44, 114, 196]. Unfortunately, many people have the urge to spread false information on social media, backed by professionally written, long, and referenced comments that allow the reader to more easily agree with the misinformation provided (e.g. [180, 299]). The purpose of the researchers is to eliminate the increased spread of misinformation by detecting the various ways in which misinformation can be spread. As such, researchers have resorted to the use of DL algorithms to detect fake news before it spreads (e.g. [148]). This is accomplished by collecting or creating a data set that contains both true and false information within articles. Then, a pattern is determined, creating a model that can predict whether a given article contains true or false information.

There are noticeable gaps in existing studies on FND that our chapter highlights. This includes (i) a lack of clear distinction between the definitions of misinformation, disinformation, and false information; (ii) a lack of systematic reviews based on DL on varying types of misinformation problems; (iii) a lack of generalizable DL models that allow achieving an acceptable detection accuracy of a base on different datasets, which introduces the scarce use of TL in this context; and (iv) a lack of models that deal with

different levels of imbalance datasets in a FND environment.

As technology progresses, the ability to detect misinformation becomes more complicated and thus more difficult to detect using standard ML techniques. This motivates our focus on DL techniques for the problem of FND.

Key Contributions

The main contributions of this chapter are as follows:

- We provide a detailed discussion of the main DL-based algorithms used to detect fake news, including their effectiveness.
- We discuss the main datasets available for FND as well as their respective characteristics, advantages, and disadvantages.
- We study TL techniques and strategies to deal with class imbalance in this application domain. We also investigate their effects on detecting fake news and the challenges of implementing these strategies.

Chapter Organization

This chapter is organized as follows. Section 3.2 presents the research methodology, including the search strategy, research questions, source databases, search query, inclusion and extraction criteria, and summary of data collection. In Section 3.3, we investigate the DL algorithms used to detect fake news. Section 3.4 describes the publicly available datasets in the field of fake news and the associated challenges. Section 3.5 discusses TL strategies and open challenges in the FND context. Section 3.6 analyzes the class imbalance problem in FND. Section 3.7 summarizes the data collected in this SLR and answers our research questions. Section 3.8 addresses the threats to the validity of the research, and Section 3.9 discusses the main gaps and open issues that still exist in FND. Lastly, Section 3.10 summarizes our chapter.

3.2 Research Methodology

3.2.1 Search Strategy Overview

Our SLR is generated based on the guidelines and a set of detailed steps described in [160]. We begin by defining our research questions, after which we build the keywords for the

search query to obtain the relevant articles for our study. Then, we select the most relevant databases to query and establish the inclusion and exclusion criteria. Finally, we define the fields to be extracted from the retrieved documents.

3.2.2 Research Questions

The key focus of our SLR is understanding how DL techniques have been used to address the FND problem. We are also interested in how TL has been applied in this field and how the problem of class imbalance has been tackled.

- **RQ1:** Which DL algorithms have been used for FND throughout time?
- **RQ2:** Which datasets are used in the FND domain?
- **RQ3:** How effective are the DL methods for FND?
- **RQ4:** Which solutions incorporate TL mechanisms, if any?
- **RQ5:** Which solutions deal with different levels of imbalanced datasets (if any)?

3.2.3 Source Databases and Search Query

For the purpose of collecting research articles, we selected four digital databases that are renowned for their comprehensive coverage and relevance to our field of study. These databases include:

- Google Scholar (we selected the articles that appeared in the first thirteen retrieved pages);
- Association for Computing Machinery (ACM) Digital Library database;
- IEEE Xplore database; and
- Scopus.

Based on the research questions established in Section 3.2.2, we collected a set of precise concepts that can cover the topic we are studying. We, therefore, formulate the search query as follows:

((fake **OR** *misinformation* **OR** *false* **OR** *unverified* **OR** *inaccurate* **OR** *rumor** **OR** *misleading)*
AND
(information **OR** *news* **OR** *article** **OR** *media)*
AND
*(detect** **OR** *classification)*
AND
("deep learning" **OR** *"machine learning"* **OR** *"neural"* **OR** *"artificial intelligence"))*

Figure 3.1: Search query used in our SLR.

The search statement displayed in Figure 3.1 addresses the research questions by focusing on the four key concepts in the studied topic: "fake", "information", "detect", and "deep learning". We searched both the title and the abstract for articles published between January 2018 and December 2023 inclusive. Limiting the search to this date range is motivated by the fact that FND has become more popular in recent years, especially during the COVID-19 pandemic that started around the beginning of 2020.

We defined the following set of restrictions on the results that limit the selection among the returned articles.

- The selected articles must be published in peer-reviewed journals or conferences. Thus, we excluded patents and any articles that did not conform to this condition.
- The language of the surveyed papers must be English. Any papers retrieved that were not written in English were excluded.
- Articles containing classification models that do not mention the performance evaluation of the methods (e.g. accuracy, precision, recall, F1-score, etc.) were excluded.
- We excluded the articles that have not mentioned the classifier/model used in the detection task in their methodology.
- We excluded the articles that only applied standard ML algorithms instead of DL ones.
- We excluded older articles when extensions and more recent editions were found.
- We excluded articles that were published in domains outside of Computer Science such as art, business, or other domains.

The total number of articles obtained from the search query, adhering to the extraction, duplicate removal, inclusion, and exclusion criteria, is 176. This includes 88 journal articles and 88 conference articles. The process of obtaining the articles is detailed in Section 3.2.6.

3.2.4 Inclusion Criteria

We considered the following inclusion criteria for our SLR:

- Peer-reviewed journals and conference articles retrieved from the search query defined in Figure 3.1.
- Articles from the Computer Science domain.
- Research articles that focus on detecting or classifying fake news.

We applied the backward snowballing technique [134] to gather relevant articles that might have been missed in our search by inspecting the reference sections of the retrieved papers. We identified two articles that were not picked up through our search query and were added to the set of manuscripts to analyze.

3.2.5 Data Extraction

We used Covidence [41], a special web-based software for supporting the data aggregation and extraction of SLR. The extracted data was organized in a spreadsheet that was exported from Covidence. The data that were extracted from the retrieved and selected articles are the following:

- Date: date of the publication;
- Publication Type: where the article has been published (conference/journal).
- Classifier/Model: algorithms used for FND in the article.
- Network Structure: the architecture of the network (details including the number/types of layers and any special setup in the network.
- Dataset: name of the fake news corpus or dataset(s) used.

- TL Techniques: the TL mechanism(s) used in the proposed solution.
- Imbalance Techniques: shows whether the imbalanced issue was treated in the proposed solution and how they dealt with it.
- Effectiveness: depicts the performance of the model in terms of accuracy, precision, recall, F-measure, and other evaluation metrics.

3.2.6 Data Collection Summary

Overall, our search query retrieved 1642 articles. We found 436 duplicate articles that were removed. After the first screening of the titles and abstracts, we ended up with 393 research papers, matching our research keywords and the inclusion and exclusion criteria. After a second full-text screening, we excluded 217 articles obtaining 176 research papers for analysis. Figure 3.2 shows the PRISMA chart demonstrating the retrieved papers' selection strategy.

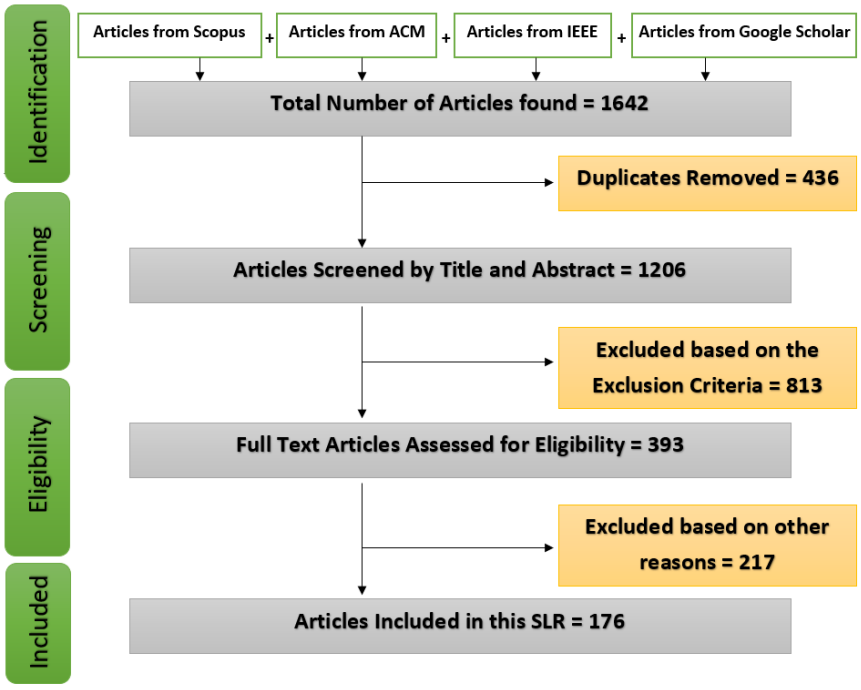

Figure 3.2: PRISMA Chart of selecting and retrieving the articles

3.3 Deep Learning algorithms used for fake news detection

The thorough examination of various models and techniques pointed out the significant role that DL plays in different classification tasks including detecting fake news. Building and improving such algorithms became a pressing necessity, especially during the COVID-19 pandemic when a large volume of fake news and rumors were being disseminated widely. Figure 3.3 demonstrates a clear increase in the use of DL models over the years.

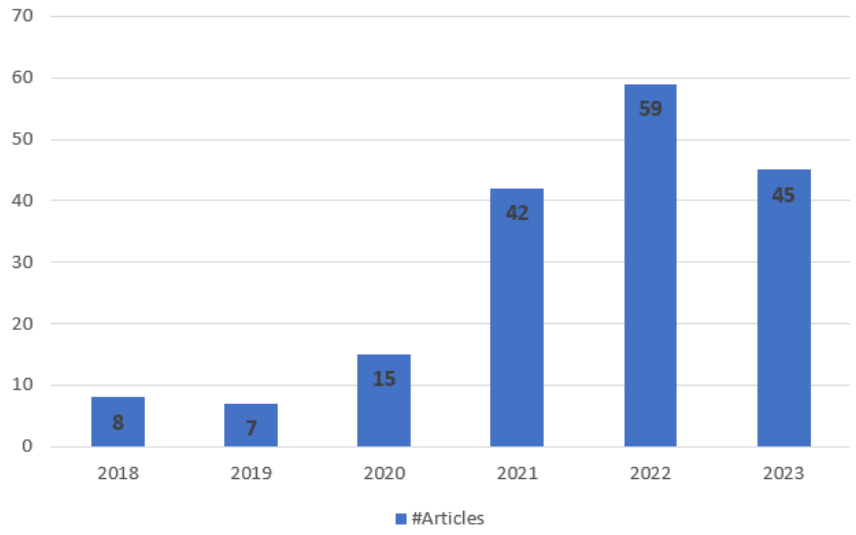

Figure 3.3: DL Models used for FND between the years of 2018 and 2023.

The extracted data shows that the FND task usually follows a generic framework as illustrated in Figure 3.4. Initially, the process involved acquiring or generating a dataset. The majority of studies have relied on news articles sourced from openly accessible datasets. After collecting the dataset, pre-processing techniques were employed to prepare the data for input into a NN. Prior investigations have mainly employed Word2vec and GloVe word embedding methods to transform words into vectors [99]. Finally, the NN model is trained and the predictions are obtained.

Neural networks for FND can be categorized into different types based on their architecture and how they process data. The first type is FNN, including single-layer and multi-layer perceptrons. CNN are another type, which is designed to process data with a grid-like topology, such as images. They include traditional CNN, residual networks, and

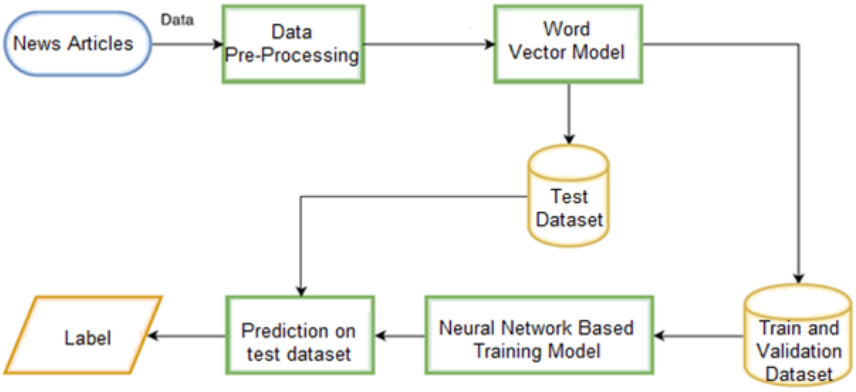

Figure 3.4: The General DL Framework that Used for FND

dense networks.

RNN are designed to handle sequential data, such as time series or language, and include basic recurrent neural networks and bi-directional RNN, LSTM, GRU and bi-directional GRU, and BiLSTM. we will refer to the model and its bi-directional version collectively as (Bi)X, where X is the model. Graph Neural Network (GNN), a newer type of NN, is designed to operate on graph-structured data, such as social networks, chemical molecules, or protein structures.

Recently, attention-based models have gained popularity due to their ability to focus on certain parts of the input data selectively. They include self-attention networks and multi-head attention networks. Hybrid models, which combine different types of neural networks, have also become popular. For example, convolutional recurrent neural networks combine the spatial processing capabilities of CNN with the temporal modelling capabilities of RNN. Transformer networks, like BERT, combine self-attention mechanisms with FNN to process sequences of data. Figure 3.5 shows a taxonomy of the various types of NN that are used for FND. It is worth mentioning that Figure 3.5 is intended to show a general categorization of the models regardless of the accurate position of the submodels in this illustration. We kept one level of models under each main category even though a model may be considered under a subcategory rather than a main category, such as the case of BiLSTM not being under LSTM.

Based on the data that we gathered from the surveyed articles, it is evident that the researchers extensively explored several DL algorithms for the detection task. Figure 3.6 shows the usage of different DL detection models for fake news. More precisely, this figure displays the percentage of articles in which a particular model was used. We observe that BiLSTM was the model most frequently included in 72% of the articles, and the CNN

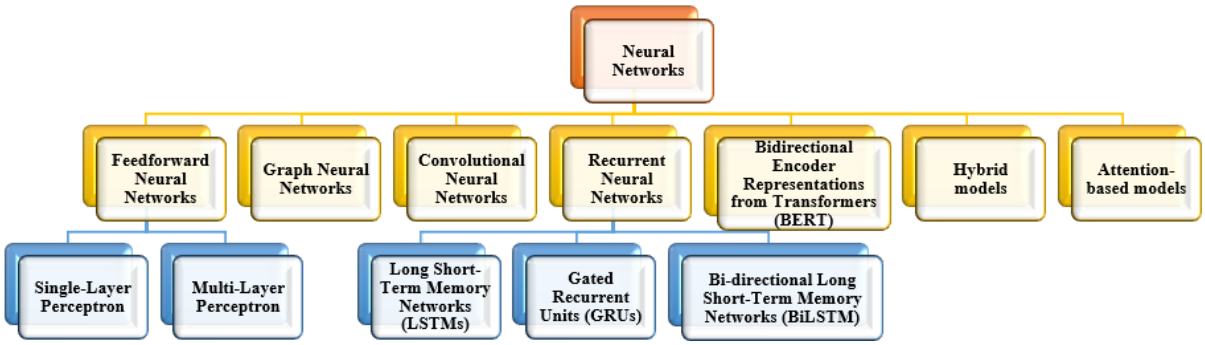

Figure 3.5: Taxonomy of the main NN categories used for FND.

model was the second most used model, utilized in 61% of the articles reviewed. The third architecture used is the hybrid architecture, which combines different types of NN in the detection process. Since multiple models may be used in the same research paper, summing up the percentages in Figure 3.6 exceeds 100%. The following sections provide a detailed discussion of the main architectures used for FND.

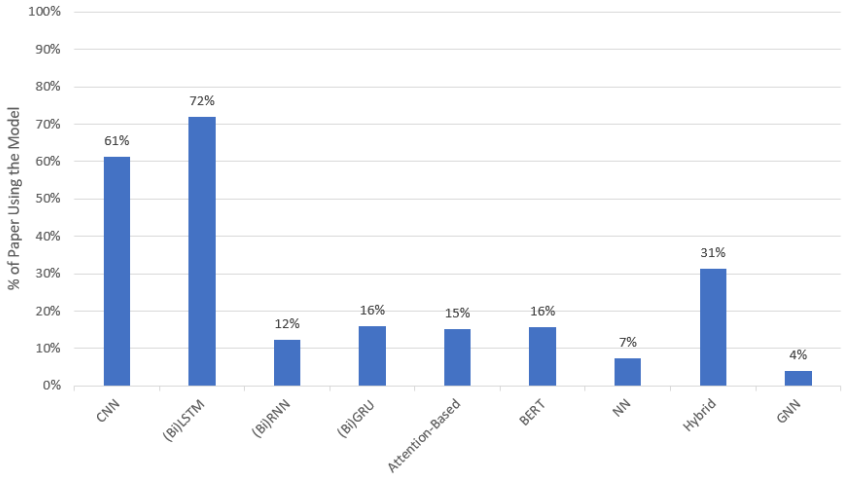

Figure 3.6: Deep Learning Models for FND

3.3.1 Architectures based on Convolutional Neural Networks

Our findings also show that 61% of the previous works used CNN to handle the detection process, attempting to boost the performance of the FND process through the use of this

DL algorithm [1–3, 14, 16, 18, 19, 28, 31, 36, 53, 58, 72, 78, 81, 83, 84, 89, 97, 100, 101, 103, 125, 128, 131, 133, 135, 136, 138, 145, 147, 148, 150, 151, 156, 161, 167, 170–172, 179, 192, 198, 205–208, 215, 223, 227, 230, 232, 245, 247, 249, 252, 255, 257, 261, 267, 269, 274, 276–278, 281, 282, 287, 292, 295, 305, 310, 314, 319, 322, 324, 327, 328, 331, 338, 346–349, 359, 361, 363, 369, 370, 376].

The detection effectiveness is the result of CNN's ability to carry out feature extraction [186]. It is worth noting here that CNN were trained on different fake news datasets. The CNN achieved notable effectiveness, with accuracy ranging between 95% and 98%, depending on whether it was used individually [179] or in conjunction with another model, such as the GRU [161], respectively. It is also worth mentioning that the CNN has fallen in some cases to about 47% detection accuracy [125] which leads to the conclusion that some key points may affect the effectiveness of the CNN in FND tasks. The first point is the degree of the deepness of the network being used. A deeper CNN is considered an advantage for solving the overfitting issue [145]. This is what we discovered using the data collected which contains a case of building a deeper CNN, called FNDNet, which solved the overfitting problem by learning the discriminatory features for FND using multiple hidden layers [148]. The second point affecting the CNN' effectiveness concerns the selected dataset that will be used in the detection task and its readability and cleanness before being fed into the model [100]. Finally, the overall architecture that will be used for the detection task may adopt either the CNN itself or a CNN in a hybrid approach as we can see in our extracted results [161].

Figure 3.7 illustrates an example of a CNN architecture used for FND, as proposed in [287]. The CNN architecture used in this study is composed of an input layer, an embedding layer, and three sets of convolutional and max pooling layers. The input layer resizes the input data to a uniform size of 1000, while the embedding layer reduces the size to 100 by embedding the data. The convolutional and max pooling pairs extract features from the input. To perform this task, filters are applied to each convolutional layer, each of which consists of 128 filters with a kernel size of 5 and a ReLU activation function. Additionally, the fully connected network includes both a flat and a dense layer. Lastly, the feature maps are classified using a dense layer with a softmax activation function.

3.3.2 Architectures based on Recurrent Neural Networks

Another popular FND algorithm examined in previous studies is RNN and its variations. The authors have investigated various RNN models to detect fake news in sequential data. They have proposed LSTM, GRU, unidirectional LSTM-RNN, vanilla RNN, and BiLSTM.

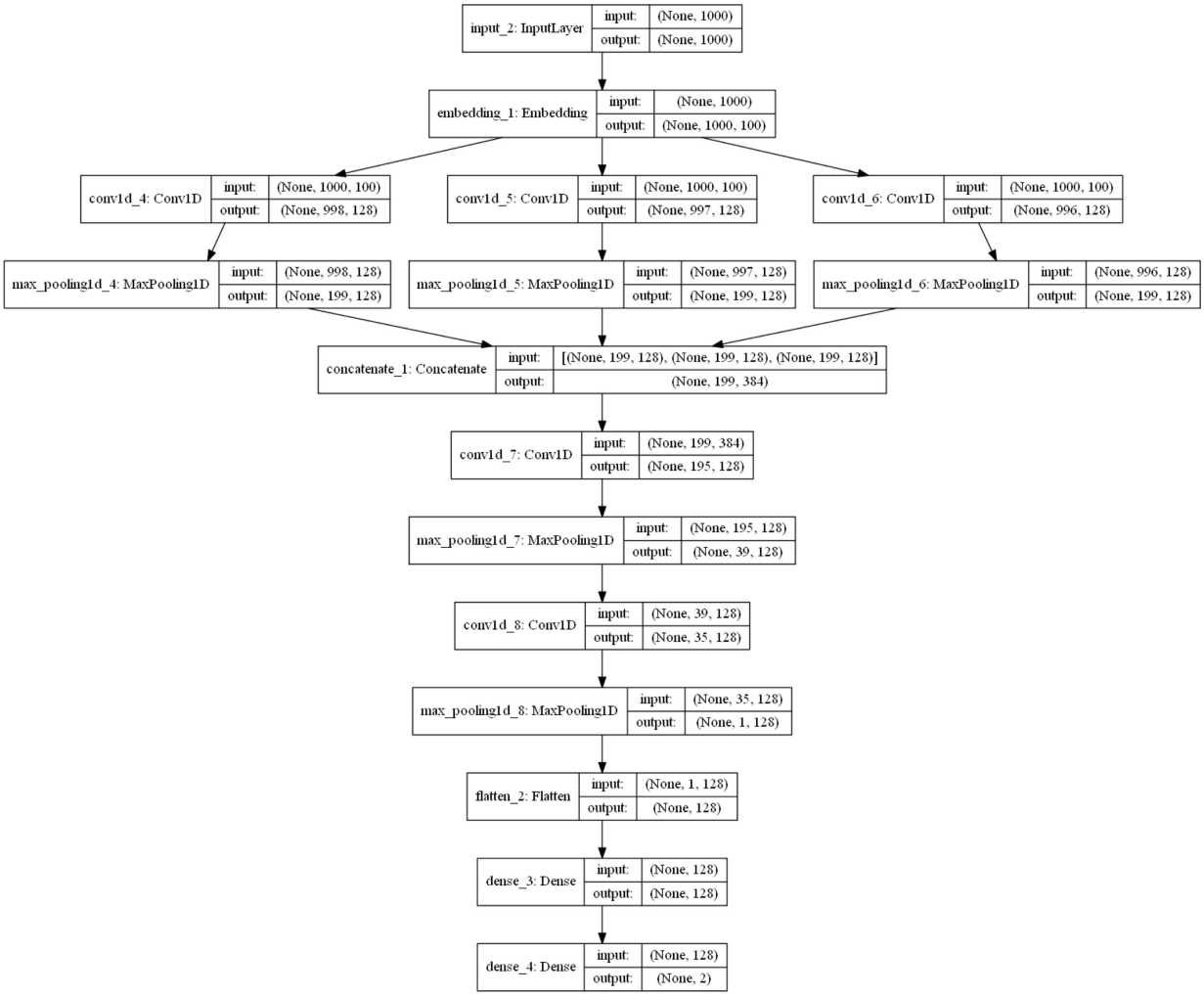

Figure 3.7: An example of CNN architecture used in FND [287]

Our findings show that the researchers' focus is highly shifted toward RNN and their variations in FND. Figure 3.8 shows the utilization of RNN in previous studies.

It is noticeable from our findings that the researchers examined FND using classic RNN only in 12% of the total number of articles surveyed [34, 72, 77, 99, 103, 128, 175, 198, 199, 203, 230, 274, 305, 307, 310, 314, 325, 361, 374]. Despite the importance of RNN in such domains, the research authors discussed the problem of RNN vanishing [21]. One solution to solve vanishing in RNN is to use other architectures such as LSTM and BiLSTM. The percentage of articles that examined both LSTM and BiLSTM was around 72% of the

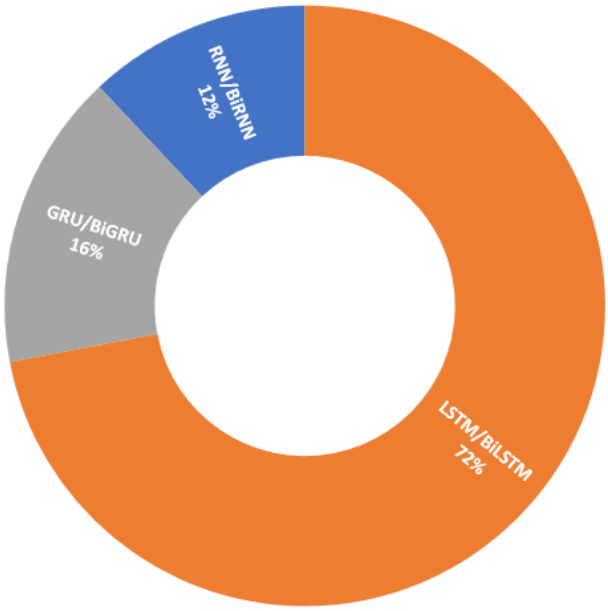

Figure 3.8: RNN utilization in FND.

total articles [1–3, 10, 14, 16, 18, 32, 33, 36, 38, 40, 43, 58, 63, 64, 67, 72, 77, 78, 80, 83, 84, 89, 93, 97, 99, 101, 103, 125, 128, 131–133, 137–139, 145, 147, 150, 151, 154, 156, 157, 164, 167, 171, 172, 175, 178, 182, 184, 188, 192, 197–200, 202–204, 206, 210, 211, 215, 217, 228, 229, 247, 250, 252, 254, 256, 257, 261, 264, 268, 269, 271, 272, 274, 281, 282, 287, 289, 292, 295, 297, 305, 306, 310, 322, 325, 327–329, 332, 334, 349, 359–361, 363, 369, 370, 374].

Other solutions were also adopted in previous studies that include the use of Bidirectional Gated Recurrent Unit (BiGRU) as a detection architecture. BiGRU has been examined in 16% of the total of articles surveyed [12, 19, 32, 33, 38, 58, 67, 82, 83, 99, 128, 161, 198, 203, 268, 280, 287, 310, 328, 331, 337, 347, 370].

Figure 3.9 shows the architecture based on RNN GRU for FND presented in [287]. In this proposed solution, the use of GRU RNN for FND is explored. The proposed model includes an input layer and an embedding layer with data sizes of 1000 and 100, respectively. The GRU layer is then implemented with hyperparameters identical to those of the LSTM layer to facilitate a reliable comparison between the two. Finally, fully connected networks are used, along with a batch normalization layer, and a dense layer with a softmax activation function is applied for classification.

The findings of our survey also show that RNN and their variations had a remarkable detection accuracy in the fake news domain compared to other detection models taking

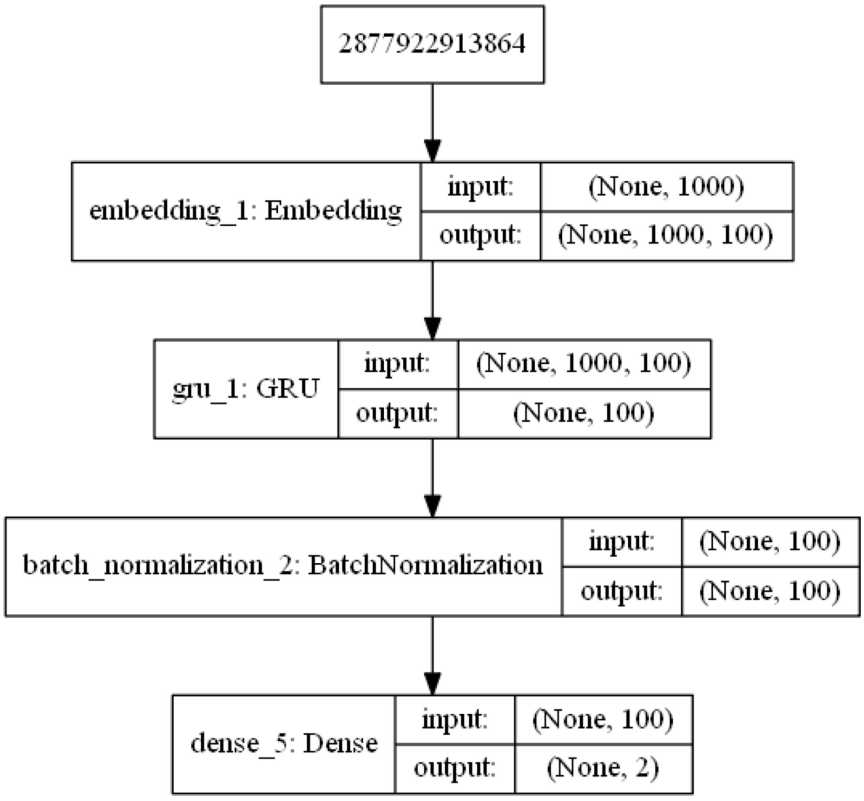

Figure 3.9: An example of GRU architecture used in FND [287].

into account the usage of different datasets. The detection effectiveness of RNN ranged from 48% [274] to around 92% [203] to around 99% [252] detection precision.

Using another architecture with RNN does not appear to increase the accuracy of the detection results as shown in the works of Ilie *et* al. [128] and Nasir *et* al. [230]. It is also noticeable in our findings that the GRU model had also participated in detecting fake news with an accuracy ranging between about 76% [280] and 97% [58]. These satisfactory results are not the case when using the BiGRU instead of the standard GRU architecture. In the latter case, the detection accuracy decreases to a range of 28% and 71% [268]. Finally, it was clear that the detection was more accurate when done by a second model besides the GRU in a hybrid mechanism. This was obvious when the researchers used the GRU with a CNN in [161] and [280] and when a GRU was used with BiLSTM in [38].

Despite the effectiveness of the above-mentioned architectures in FND, previous studies showed that the LSTM and the BiLSTM are the future key players in enhancing FND.

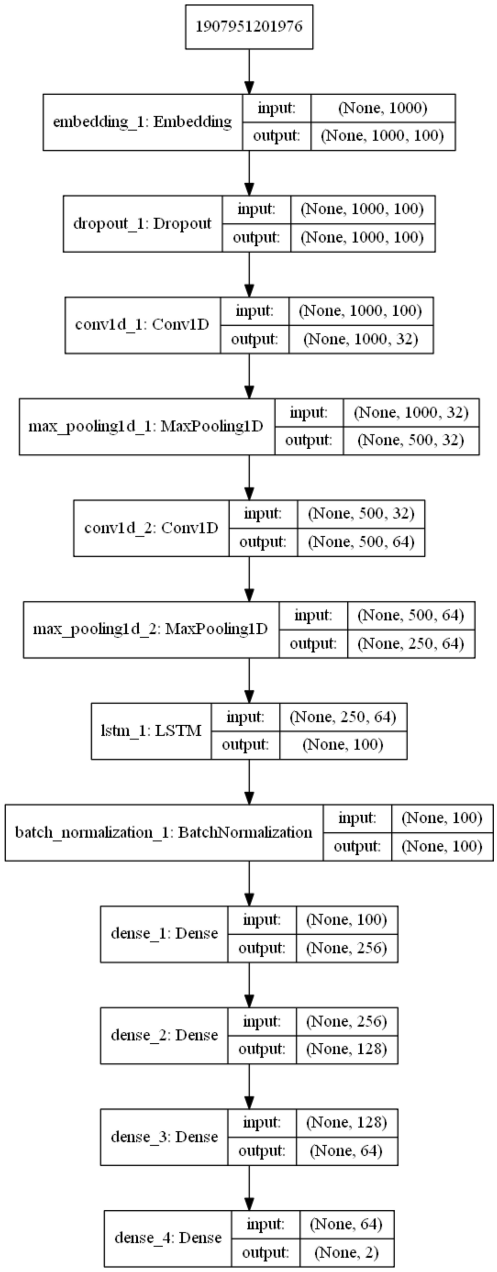

Figure 3.10: An example of CLSTM architecture used in FND [287]

The average accuracy of detecting fake news using LSTM architectures ranged between 79.03% and 81.21%. These models also reported a maximum of 99.9% detection accuracy in [83] and a minimum of 11% in [171].

In addition, the findings show that LSTM was used in a hybrid fashion with one or more architectures to determine the optimal FND system among the proposed systems. Figure 3.10 shows the architecture of the CNN-LSTM model proposed in [287]. This model utilizes both hybrid and recurrent models on collected news data. The proposed hybrid model incorporates both the CNN and LSTM models. The algorithm includes an input layer that resizes the input data frames to 1000 and an embedding layer that embeds the input tensor size from 1000 to 100. The embedded tensors are then processed through two sets of convolutional and max-pooling layers for feature extraction. The convolutional layers have 32 and 64 filters, respectively, and a kernel size of 3. The feature extraction process is then performed by the LSTM layer with 100 units, a dropout rate of 0.2, and a recurrent dropout rate of 0.2. Additionally, the fully connected network is designed with a batch normalization layer, followed by three dense layers with a ReLU activation function and several filters of 256, 128, and 64, respectively. The classification task is carried out using a dense layer with a softmax activation function.

Researchers focus more on testing the effects of developing hybrid models that adopt LSTM in the detection process. They tested the importance of BiLSTM over LSTM and reported that the BiLSTM+CNN achieved considerably higher accuracy than when they attempted to use the LSTM with the CNN. They reported a detection accuracy of about 99% detection accuracy when they attempted to use the BiLSTM instead of the LSTM [252].

When LSTM is combined with CNN, studies also reported an accuracy ranging between 97.8% in [200] and 47.06% in [125], with an average accuracy of 82.3%. In the case where LSTM is combined with a DNNs architecture, we observed an accuracy of 91.16% [228], while when it is combined with BERT the accuracy achieved was 84.10% [254].

BiLSTM is also getting popular in FND as our survey findings show. It recorded the highest detection accuracy of 99.52% [182] and the lowest of 28% [268] with an average of 75.22%. It also appeared connected to other detection architectures such as CNN and GRU. BiLSTM with CNN recorded the highest accuracy of 98.65% in [217] and the lowest accuracy of 35.13% in [328]. The average detection accuracy in such cases was about 77.6%. BiLSTM with GRU reached 89.8% detection accuracy [38].

3.3.3 Architectures based on Graph Neural Networks

Another popular model in fake news detection is the GNN and its variants such as Sequence Graph Transform (SGT) [318], Graph Attention Networks (GAT) [61], GraphSAGE [107], and Graph Convolutional Networks (GCN) [166]. GNN is a neural network directly operating on the graph structure. One of its popular applications is node classification in which every node in the network has a label. This network predicts the node's label without the ground truth [284]. In FND using GNN, news articles, and related information are represented as a graph. The nodes of the graph represent the individual entities, such as news articles, users, or social media posts, and the edges represent the relationships or interactions between them. To create the graph, the news articles are typically preprocessed to extract features such as the article content, metadata, and social media interactions. These features are used to construct the nodes and edges of the graph, with the nodes representing the articles and the edges representing the relationships between articles, users, or other entities. For example, edges could represent similarities between articles or social media interactions such as retweets or mentions. Once the graph is constructed, GNN are used to analyze the graph structure and extract useful features for FND. The GNN use graph algorithms to propagate information across the graph and learn representations of the nodes and edges that capture their relationships and interactions. These learned representations can then be used to classify the news articles as fake or real based on their similarity to other articles and the overall structure of the graph.

Our findings show that only 4% of the selected research articles adopted GNN architectures for FND [42, 119, 201, 244, 263, 313, 333, 347]. The claimed detection accuracy was incredibly low compared to the other DL models on different datasets used. The highest detection accuracy obtained when using the GraphSAGE was 89.7% accuracy without mentioning whether this was on the training or the testing dataset [244]. The accuracy went deeply down to 61.5% when they adopted the GNN. It also recorded a 73.12% [42] with GCN with a maximum of 88.6% [244]. The other variants such as SGT, GCN, and GAT had reached an average accuracy of about 83.1%.

3.3.4 Attention-based and BERT-based Architectures

Another notable advancement happened in FND with the use of attention-based approaches using different datasets. Our findings show that their use has been increasing since the year of 2018 and has reached the maximum in the year of 2022. In addition, this approach appeared in 15% of the surveyed articles mostly in the year 2022. Authors have applied it to the other detection models including RNN [128], GRU [12, 19, 82, 128, 234], LSTM, and

BiLSTM [3, 15, 132, 217, 247, 257, 329, 333, 349, 360], BERT [349], and CNN alone [247, 257] or with other models [3, 257, 329]. The detection accuracy ranged between 54% [257] and 98.65% [217].

Another deep learning model in our surveyed works that shows cutting-edge detection is the BERT [79] model. It is a sophisticated pre-trained word-embedding model built on a transformer-encoded architecture. It was used widely for different detection tasks such as FND, hate speech detection [309], depression detection [127], and others. The findings show that 16% of the surveyed studies adopted the BERT as a FND mechanism [31, 40, 78, 84, 129, 155, 163, 163, 171, 176, 227, 229, 238, 254, 269, 272, 276, 277, 293, 310, 325, 338, 349, 370, 374]. The findings also show that authors started using the BERT as a detection model for fake news in 2021 which makes it still a novel tool for the detection model and a future direction in the FND field. Our findings show that this model has reached a remarkable detection accuracy with the highest recorded accuracy of 98.5% [293] and an average accuracy of around 90%. It is also clear in our findings that researchers experimented with the effectiveness of applying the BERT with other models such as LSTM [254] and CNN [277] for the detection of fake news using different datasets.

An example of using BERT in the FND process, FakeBERT has been proposed in [146] which outperforms all other models with an accuracy of 98.9%. Figure 3.11 illustrates the proposed FakeBERT.

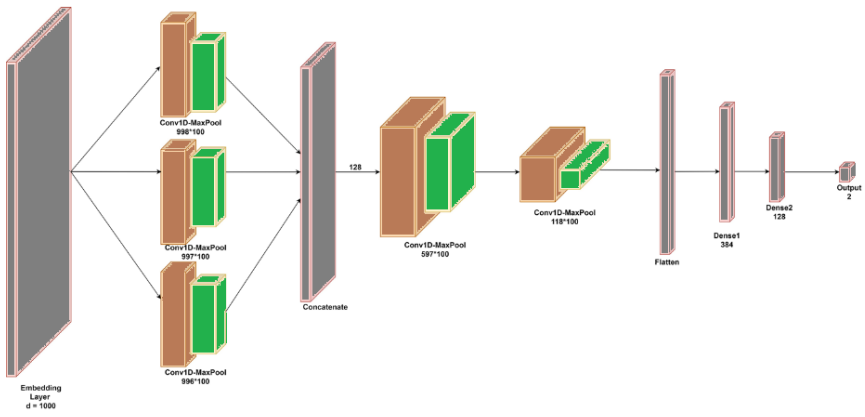

Figure 3.11: An example of BERT-based architecture used in FND, FakeBERT

As Figure 3.11 shows, this design employs three parallel blocks of 1D-CNN with 128 filters, each with one convolutional layer. The first layer has a kernel size of 3 and 128 filters, reducing the input embedding vector from 1000 to 998. The second layer has a kernel size of 4 and 128 filters, reducing the input vector from 1000 to 997. The third

layer has a kernel size of 5 and 128 filters, decreasing the input vector from 1000 to 996. Max-pooling layers are also included after each convolutional layer to reduce the dimension further. A max-pooling layer with a kernel size of 5 reduces the vector to 1/5th of 996, which is 199. After concatenating the three convolutional layers, another convolution layer with a kernel size of 5 and 128 filters is applied. This is followed by two hidden layers with 384 and 128 nodes, respectively. The number of trainable parameters for each layer is also provided in the "Param number" column for further details.

A recent study has thoroughly compared different DL models in FND using various datasets [25]. The authors studied the effect of the deployment of BiLSTM, CNN-RNN, C-LSTM, CNN, and BERT in the detection of fake news. They used seven FND datasets with each model to be able to draw a generalized conclusion. They figured out that the BiLSTM and BERT detection models achieved the best detection accuracies and F1-scores. The authors have also concluded that BERT performs better than BiLSTM when the model aims to detect fake news in different contexts from the one it was trained on [25].

3.3.5 Architecture based on Feedforward Neural Networks

Finally, other DL models have been used in the FND field with basic and standard FNN settings. Authors categorized these models under simple neural networks (NN; ANN, DNNs, and FNN). Although these models are referred to as simple detection techniques and were used in 7% of the total surveyed articles, they still reached a noticeable accuracy in detecting fake news. Our findings show that FNN has reached a detection accuracy of 89.8% [172] to about 95% when it was provided by solid support from a strong embedding technique [264] and a lower accuracy of 83.35% [77]. On the other hand, DNNs reached an accuracy of 94.68% [323] while it was less accurate when applying it with an LSTM by 2.8% [228]. The findings also show that using a multichannel ANN [323] has increased the detection accuracy by approximately 13% of the basic ANN which was 80.9% accurate in detecting fake news [149].

In conclusion, we observe a clear growing trend in the solutions proposed using BiLSTM, CNN, BERT, etc. throughout the years, as Figure 3.12 shows.

3.3.6 Challenges related to deep learning methods for fake news detection

Despite the promising results of DL methods for FND, several challenges remain to be addressed. These include issues related to dataset quality, model performance on imbalanced

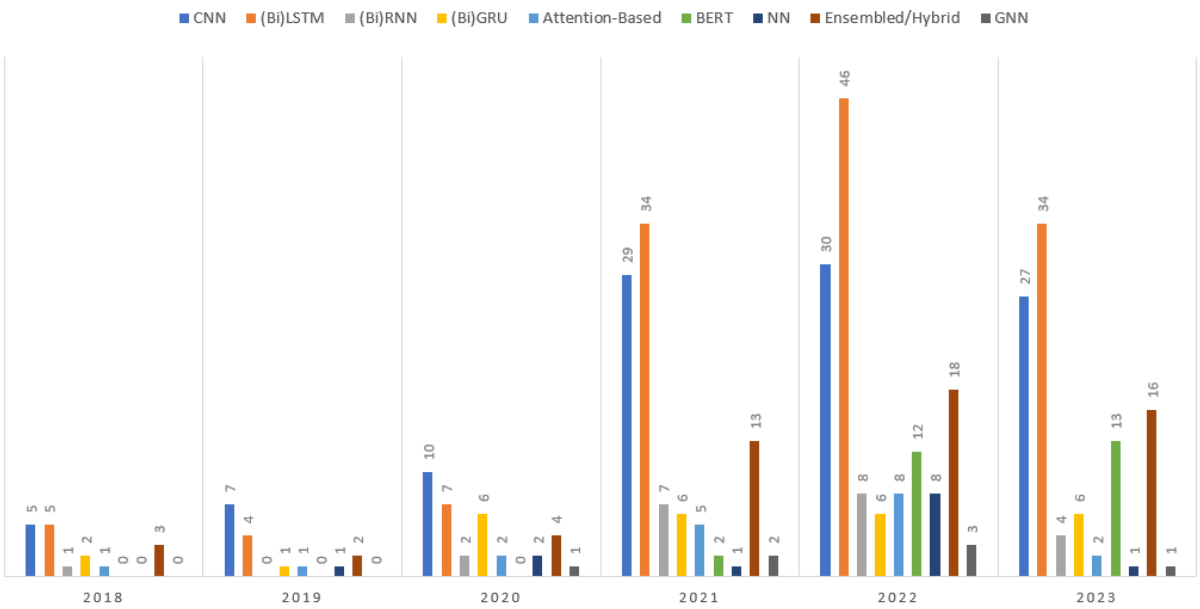

Figure 3.12: DL models in FND throughout the time.

datasets, and the generalizability of models across different datasets. In this section, we will explore these challenges in more detail and discuss potential solutions.

It is important to recognize that there are several challenges when it comes to achieving effective fake news detection using deep learning methods. One major issue is the potential for overfitting, where models achieve high accuracy on the training data but perform poorly on new unseen data [369]. Some previous research has reported extremely high accuracy results, but these were obtained by evaluating the model on the same data that was used for training. The performance of their models achieved a high accuracy of 99.9% [64, 83, 139, 182, 211, 217, 282], and [369]. This raises questions about the model's ability to generalize to new data. Another challenge is the use of accuracy as the sole evaluation measure for imbalanced datasets, where the number of fake news samples vastly outweighs the number of real news samples [1–3,16,64,132,145,151,178,211,271,289,297,346,349,369]. Accuracy can be misleading in these cases, as it can be skewed by the dominance of the majority class. A more appropriate measure, such as precision or recall, would provide a better understanding of the model's performance. Additionally, different datasets can have varying characteristics and biases, and models that perform well on one dataset may not generalize to other datasets. This is noted from our findings in [154, 171, 256, 268], and [274]. This was also proved by the extensive experiments carried out in a recent study

of cross-domain FND [25]. Finally, the quality and diversity of the training data can greatly impact the performance of the model [222, 372]. In some cases, models have been trained on datasets that are not representative of the full range of fake news content, leading to poor detection performance [106, 222].

3.4 Datasets used for fake news detection

In this section, we first discuss the main characteristics of the datasets used in the surveyed works. Then, we discuss some of the open challenges related to the datasets in this application domain.

3.4.1 Main datasets used for fake news detection

Researchers have used several datasets in the context of FND. However, we found that only a small part of these datasets is publicly available, while a considerable percentage is created by the researchers and/or is not disclosed publicly. A pie chart of the used datasets in the surveyed studies is presented in Figure 3.13.

We observe that ISOT [9], PHEME [377], Liar [345], and FakeNewsNet [300], with its three sub-datasets, GossipCop, PolitiFact, and BuzzFeedNews, are examples of publicly available fake news datasets. These are among the most popular and commonly used datasets.

The LIAR dataset includes short statements obtained from the Politifact fact-checking website. This dataset includes a total of 12.8 K labelled short statements. The annotation task has been done by the Politifact site, and the statements are classified into 6 classes: pants-fire, false, barely-true, half-true, mostly-true, and true. In addition, another fake news dataset was collected from real-world news articles called ISOT. The real news cases were collected by crawling news articles from Reuters.com, and the fake news examples were obtained from unreliable websites, which were annotated by the Politifact website. The PHEME dataset was collected from Twitter based on 9 newsworthy events classified by journalists. The annotation process was conducted by journalists (human annotators) and each tweet was annotated with one of the following labels: "proven to be false", "confirmed as true" or "unverified". The FakeNewsNet dataset consists of three subdatasets, which are GossipCop, PolitiFact, and BuzzFeedNews. In total, the FakeNewsNet dataset contains approximately 19,838 news articles labelled as either "fake" or "real". The news articles in the FakeNewsNet dataset were annotated by a team of human annotators. Annotators

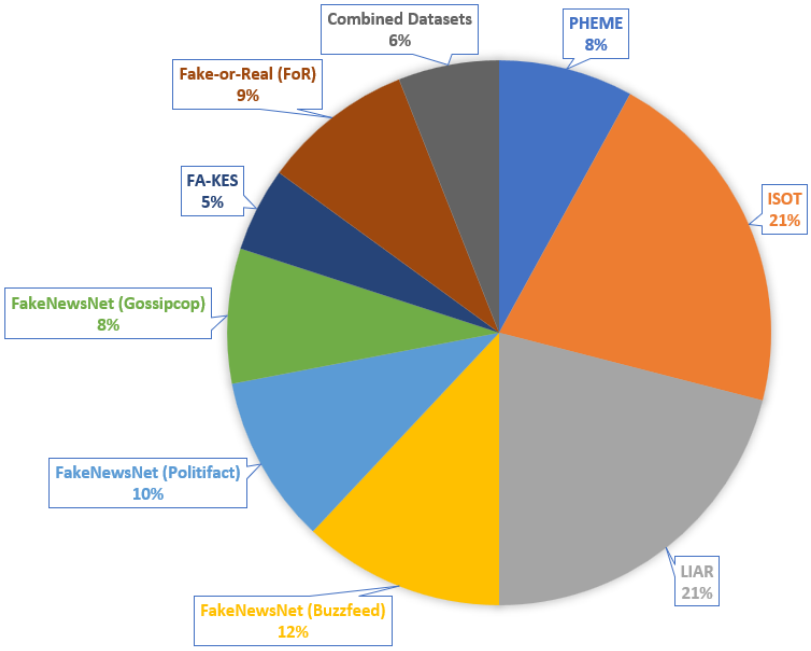

Figure 3.13: Datasets used in the surveyed studies

received guidelines for identifying fake news and were trained to identify various characteristics of fake news, such as misleading headlines, fabricated content, and misleading images. Table 3.1 summarizes the main characteristics related to these datasets.

Table 3.1: Main characteristics of the publicly available FND datasets.

Dataset	Size	Num. of Labels	Type
LIAR	12.8K	6	Political Statements
PHEME	5800	2	Social media (tweets)
ISOT	45k	2	News articles
BuzzFeedNews	5,835k	2	News articles
PolitiFact	12,835k	2	News articles
GossipCop	1168k	2	News articles

From our findings, LIAR achieved a maximum of 98.95% when a BiLSTM model was

used for training [154]. The same dataset was an option for training the BiGRU in [268] which recorded a low detection accuracy of 28.12%. ISOT recorded high detection effectiveness in many cases, especially when the trained model was a BiLSTM [282] with an accuracy of 99.95%. Still, the accuracy decreased when the models used for training were RNN and CNN, with the lowest performance recorded at 82.5% [230]. PHEME also exhibited high performance when the CLSTM model was used, achieving an accuracy of 91.88% and recording a minimum accuracy of 65.5% with the training of a CNN [324]. Lastly, FakeNewsNet sub-datasets used in training different models such as CNN [274], various RNN [254, 268, 274, 325, 374], GNN [201], and BERT [238, 254, 325, 374]. The best detection accuracy was achieved when training GAT with a 96.42% accuracy while recording a 71.16% accuracy when used to train BiGRU.

3.4.2 Challenges related to the datasets used for fake news detection

One of the main difficulties in the field of FND is the scarcity of labeled cases [59, 233, 335]. Even though multiple datasets with massive amounts of records exist, they are mostly unlabeled or have only a few records labeled. Researchers have collected datasets over the last few years for use with DL models in different contexts associated with FND. Datasets are massively diverse from each other due to having different research goals inside the fake news detection application domain [352]. For example, some datasets contain exclusively political statements, while other datasets only include news articles or posts on social media [25].

To collect appropriate datasets to serve in FND, we need fake articles and non-fake articles. Fake articles are collected from fraudulent websites that are designed to spread misinformation and false news. The fake news published on these websites will eventually be shared on social media to be read and circulated by innocent people who do not check the source of the news.

It is also clear from our findings that the datasets used in FND are insufficient for training models due to their characteristics, such as language features or size [115]. That leads us to the question of creating a dataset to serve as a benchmark in the detection process. However, this can be challenging due to several reasons, some of which are:

- Sources of fake and non-fake news: Identifying reliable sources of fake and non-fake news can be difficult, especially in today's world where there are numerous sources of information and not all of them are trustworthy [240]. It is crucial to ensure that

the dataset contains a diverse range of sources to ensure that the model is trained to detect fake news from a variety of sources.

- Bias in the data: Bias can be introduced in the data due to various reasons such as the sources of the data, the labelling process, or the selection criteria for the dataset. Bias can affect the accuracy of the model and can also lead to unfair predictions [158].
- Labeling issues: Labeling data for FND can be challenging, as there can be discrepancies in the definition of what constitutes fake news. Human labels may be subjective, and there may be inconsistencies in the labelling process. Automatic labels generated using ML techniques can also have their limitations [59, 339].
- Bots involvement: Bots can be used to generate large volumes of fake news and spread it rapidly across the internet, making it difficult to detect and remove. Bots can also be used to manipulate the labelling process by providing biased labels, leading to inaccuracies in the dataset [240].
- Rapid evolving nature of fake news: The nature of fake news is constantly evolving, and new techniques for creating and spreading it are being developed all the time. This makes it difficult to create a comprehensive dataset and up-to-date [315, 352].

To address these issues, it is crucial to have a well-designed and diverse dataset that is regularly updated to reflect the changing nature of fake news. It is also important to have robust labelling procedures in place, using a combination of human and machine labels, to ensure that the dataset is unbiased and accurate. Additionally, researchers should consider incorporating techniques such as adversarial training to improve the robustness of the model to adversarial attacks.

3.5 Strategies for transfer learning

3.5.1 Transfer learning strategies applied to fake news detection

Numerous real-world applications have made use of the ML and DL techniques. These learning methodologies assume that the input feature space and data distribution properties are maintained across the experiments carried out because the training data and testing data are drawn from the same domain [354]. This assumption, however, may not be accurate in some real-world ML situations. In fact, in some circumstances, gathering

training data can be costly and/or challenging. As a result, the research community has been considering the development of high-performance learners who are trained using data that could be more easily obtained from other various domains instead of the deployment domain.

TL is a powerful approach for improving learning efficiency by leveraging knowledge from a related domain. Consider two individuals learning Spanish: one has no prior knowledge of other languages, while the other is fluent in Italian. Due to the similarities in grammar, vocabulary, and linguistic structures, the Italian speaker can transfer their existing knowledge to Spanish, making the learning process faster and more effective [239]. Similarly, in ML, previously acquired knowledge from a related task or domain can accelerate and enhance the learning of a new, similar task.

The essence and necessity of TL appear when there is a dearth of target training data [354]. This can be the result of the data being rare, expensive to gather and label, or inaccessible. The use of other existing datasets that are related to, but not precisely the same as a given target domain of interest makes TL solutions an alluring strategy since big data repositories become more widespread. TL has been successfully used in many ML and DL applications, including text sentiment classification [343], image classification [35, 174, 375], classification of human activity [110], classification of software defects [224], and classification of multi-language text [248].

Different techniques can be utilized in TL to accomplish tasks as the following [367]:

- Training models in similar domains: This TL method trains models that belong to similar domains. For instance, if there is insufficient data to complete task X, but task Y is similar and has adequate data, a model can be trained on task Y and then used to create a new model for task X [30].
- Feature extraction: Feature extraction is another TL approach where deep NN are trained to extract features automatically. After training them on pre-existing models, the representations are exported to new models. This technique is commonly employed by data scientists [121].
- Utilizing pre-trained models: This approach involves developing pre-trained models that take TL variables into account. Companies experienced in model development often have access to a library of models that can be used to create future models. This means that when dealing with a new problem, a pre-trained model can be selected, optimized for the problem at hand, and then reused to train another model [30].

The first TL technique involves training models in similar domains by using a pre-trained model from a source domain (i.e. the original domain used to train this model) that is similar or related to the target domain (i.e. the new domain we are considering in our task). The idea is that the knowledge learned from the source domain can be leveraged to improve model performance on the target domain, even if the target domain has limited labelled data.

Training models in similar domains typically involve the following steps:

1. Selecting a source domain: The source domain should be chosen based on its similarity or relevance to the target domain. Ideally, the source domain should have similar data distribution, task, or domain characteristics as the target domain, so the knowledge learned from the source domain can be effectively transferred to the target domain.

2. Acquiring or creating a labelled dataset in the source domain: A labelled dataset in the source domain is needed for training the pre-trained model. This dataset should be representative of the data in the source domain and should cover the task or tasks of interest.

3. Pre-training the model on the source domain: The pre-trained model is trained on the labelled dataset in the source domain. This involves training the model using standard ML or DL techniques, such as supervised learning or unsupervised learning, depending on the availability of labelled data in the source domain.

4. Fine-tuning or adapting the pre-trained model to the target domain: After pre-training on the source domain, the pre-trained model is fine-tuned or adapted to the target domain. This typically involves further training the model using the limited labelled data available in the target domain, while retaining the knowledge learned from the source domain. Fine-tuning can be done by updating the weights of some or all of the layers of the pre-trained model, depending on the specific task and data.

5. Evaluating and validating the model performance: The fine-tuned model is evaluated and validated on the target domain dataset to assess its performance. This may involve measuring metrics such as accuracy, precision, recall, F1 score, or other relevant performance indicators to determine the effectiveness of the TL approach.

TL by training models in similar domains can be useful when the target domain has limited labelled data, but related or similar domains have abundant labelled data. By

leveraging the knowledge learned from the related source domain, the model can benefit from the additional data and potentially achieve better performance on the target domain task. However, it is important to carefully consider the similarity and relevance between the source and target domains to ensure that the knowledge transfer is effective and results in improved performance.

For the second TL technique, feature extraction is one of the common techniques used in TL, where a pre-trained model is used to extract features from data in one domain and these features are then used to train a new model for a different task or domain [209].

In TL with feature extraction, the pre-trained model is typically a deep NN trained on a large dataset from a source domain. This model has learned to extract relevant features from the source domain data, which can be representations or embeddings of the input data at different layers of the network. These learned features are then used as inputs to a new model, often referred to as the target model, which is trained on the limited labelled data available in the target domain.

The process of using feature extraction in TL typically involves the following steps:

1. Selecting a pre-trained model: The pre-trained model should be chosen based on its relevance to the target task or domain. Ideally, the pre-trained model should have been trained on a large dataset from a source domain that is similar or related to the target domain, so that the learned features are relevant to the target task.

2. Removing the last layers of the pre-trained model: The last layers of the pre-trained model, which are often responsible for task-specific predictions, are removed to retain the feature extraction capability of the model. These last layers are replaced with new layers that are specific to the target task.

3. Extracting features from the source data: The pre-trained model is used to extract features from the data in the source domain. This typically involves passing the data through the layers of the pre-trained model up to a certain layer and using the outputs of that layer as the learned features.

4. Training a new model on top of the extracted features: The extracted features are then used as inputs to a new model, which is trained on the limited labelled data available in the target domain. This new model, often referred to as the target model, is trained using standard machine learning or deep learning techniques, such as supervised learning or fine-tuning, depending on the availability of data in the target domain.

5. Evaluating and validating the target model performance: The trained target model is evaluated and validated on the target domain dataset to assess its performance. This may involve measuring metrics such as accuracy, precision, recall, F1 score, or other relevant performance indicators to determine the effectiveness of the TL approach.

Feature extraction in TL allows leveraging the knowledge learned from the source domain to extract relevant features from the data in the target domain, even if the target domain has limited labelled data. By using the learned features as inputs to a new model, the target model can potentially benefit from the representations or embeddings learned from the source domain. This can help improve the performance of the target model on the target domain task. However, it is important to carefully consider the similarity and relevance between the source and target domains to ensure that the features extracted from the source domain are relevant to the target task.

For the third TL technique, utilizing pre-trained models is a common approach in TL where a pre-trained model, typically trained on a large dataset, is used as a starting point for training a new model on a smaller target dataset. The idea is that the knowledge learned from the source domain can be transferred to the target domain, even if the two domains are different, to improve the performance of the target model [143].

Here are some key steps involved in utilizing pre-trained models for TL:

1. Select a pre-trained model: Choose a pre-trained model that is trained on a large dataset and relevant to your target task. For example, suppose you are working on an image classification task. In that case, you can choose a pre-trained CNN such as VGG, ResNet, or Inception, which are trained on large image datasets like ImageNet.

2. Remove or freeze some layers: Depending on the architecture of the pre-trained model, it may be necessary to remove or freeze some layers. For example, you can remove the output layer(s) of the pre-trained model and replace them with new layers that are suitable for your target task. Alternatively, you can freeze the weights of some of the layers in the pre-trained model and only fine-tune the remaining layers during the training process.

3. Add new layers: Add new layers on top of the pre-trained model to adapt it to your target task. These new layers are typically randomly initialized and are trained using the target dataset. The output of these new layers serves as the final prediction layer for your target task.

4. Fine-tune the model: Train the entire model, including the pre-trained layers and the newly added layers, on your target dataset. During the fine-tuning process, the weights of the pre-trained layers and the new layers are updated using the gradients computed from the target dataset. Fine-tuning allows the model to learn task-specific representations while leveraging the knowledge from the pre-trained model.

5. Evaluate and tune: After training, evaluate the performance of the transferred model on your target task. You may need to tune the hyperparameters and architecture of the transferred model to optimize its performance.

Utilizing pre-trained models can be an effective TL approach as it allows leveraging the knowledge learned from large datasets, reducing the need for extensive training data in the target domain, and potentially improving the performance of the target model. However, it is important to carefully choose the pre-trained model, architecture, and fine-tuning strategy to ensure that the transferred knowledge is relevant and beneficial for the target task.

There are various pre-trained ML models available in the market, such as Google's Inception model [317], Microsoft's MicrosoftML R package [213] and Microsoftml Python package [212], and others like AlexNet [173], Oxford's VGG model [304], and Microsoft's ResNet [113]. In addition, some of the well-known pre-trained models used for NLP-related data problems are Google's word2vec model [214], Stanford's GloVe model [242] and BERT [79].

BERT is a pre-trained language model that was initially introduced by Google in 2018. The model is trained on a large corpus of unlabeled text data to learn the underlying structure of the language. It utilizes a transformer-based architecture that allows it to capture long-term dependencies and contextual relationships between words. After pre-training, the model is fine-tuned on a specific downstream NLP task, such as sentiment analysis, question answering, or named entity recognition. This fine-tuning step enables the model to adapt to the specific requirements of the downstream task. In summary, BERT is a TL technique that leverages pre-training on unlabeled text data and fine-tuning on specific NLP tasks to achieve state-of-the-art performance on a variety of NLP benchmarks [79].

In particular, the most common TL strategy in FND is fine-tuning pre-trained models. Models like BERT, Llama, and Generative Pre-trained Transformer (GPT) have been pre-trained on extensive text corpora and can be fine-tuned for FND [84]. By adjusting the weights of these models on a specific fake news dataset, researchers can achieve high detection accuracy with relatively low computational resources.

Other TL strategies used for FND include the adaptation of CNN, traditionally used for image recognition, to text classification tasks, including FND [28]. Models like VGG16, which previously trained on large image datasets, may be reused by replacing the last layers and retraining on textual data [87]. This strategy takes advantage of CNN' hierarchical feature extraction capabilities, which enable them to detect detailed patterns in textual data that indicate fake news. In addition, pre-training hybrid models on large datasets and then fine-tuning them on specific fake news datasets used in fake news to exploit the strengths of both architectures.

Recently, the researchers shifted the whole focus to transformer-based models, particularly those like BERT, GPT-3, and Llama [237, 241]. These models are pre-trained on massive datasets using self-supervised learning techniques, which enable them to understand and generate human-like text. For FND, these models can be fine-tuned on labelled datasets specific to fake news, enabling them to distinguish between fake and real news with high precision.

Based on the collected data from the surveyed articles, along with their corresponding FND effectiveness, the following conclusions can be drawn related to the use of TL techniques in this domain:

1. CNN with AlexNet as a TL technique achieved an accuracy of 93.2%. In comparison, not applying TL recorded an accuracy of 70.1% in [28].

2. In [129], pre-trained BERT as a TL technique achieved an accuracy of 94.66%. Similarly, a pre-trained BERT has also helped in the detection of fake news using the ISOT dataset [84]. In another case of BERT variations, RoBERTa achieved an accuracy of 92.77% and 91.7% on Politifact and Gossipcop respectively [237] which outperform the state-of-the-art, without TL, techniques by achieving an average accuracy of 10.49% and 14.53% improvements on Politifact and Gossipcop, respectively.

3. CNN with various TL techniques, such as AlexNet, ResNet50, MobileNet, DenseNet, XceptionNet, InceptionV3, VGG16, and VGG19, achieved high accuracy on the EMERGENT dataset [87]. The detection accuracy was ranging between 91.22% and 97.68% in [255]. VGG16 was also used as a pre-trained model with freezing some layers and trained on a self-created dataset, achieving about 98% detection accuracy [205].

4. The Universal Language Model Fine-tuning TL technique has achieved over 80% for all the evaluation metrics (Accuracy, Precision, Recall, F1) on PHEME dataset [324].

3.5.2 Transfer learning challenges for fake news detection

TL has been used in various NLP applications, including FND. However, there are several challenges associated with applying TL in this domain.

One initial aspect that we must highlight is that TL has not been extensively explored in FND. we consider that this is partly due to the complexity of the task, which requires identifying subtle linguistic cues and context-specific information.

Another challenge concerns the difficulty in finding related domains and publicly available datasets that can be useful for training the models. The success of TL relies on the availability of large and diverse datasets that share some commonality with the target task. However, in the case of FND, relevant datasets are often limited, and it can be challenging to find related domains that can be used for TL.

The rarity of data is another significant challenge in FND. Since the detection of fake news is a relatively new area of research, there are limited annotated datasets available for training and testing models. This scarcity of data makes it difficult to apply TL techniques, which rely on large amounts of labelled data for pre-training.

In addition to these challenges, other issues need to be addressed to effectively apply TL in FND. For instance, the choice of pre-trained models and their adaptation to specific tasks can significantly impact the performance of the models. Furthermore, the transferability of pre-trained models across different languages, domains, and cultures is still an active area of research. Considering TL strategies is a relevant area for further research that can lead to improved solutions for FND.

3.6 Strategies for dealing with imbalance

DL and ML algorithms presuppose that the target classes of the training data have similar prior probabilities. This assumption, however, is flagrantly violated in a variety of real-world applications, including FND. In this section, we start by summarizing the main techniques used to deal with the class imbalance problem and describe our main findings from the surveyed articles in the context of FND. Then we summarize the main open challenges related to the class imbalance problem that are still open in this domain.

3.6.1 The class imbalance problem in the context of fake news detection

In many real-world domains, the majority of the available examples belong to one class (the majority or negative class). In contrast, a much smaller number belongs to the other class (the minority or positive class), which is typically the most important class [288]. This situation is known as the class imbalance problem. The dominant class tends to overpower classifiers in this situation, causing them to overlook the minority class. The significance of the imbalance problem grew as more researchers discovered that it leads to inadequate classification performance and that most algorithms perform poorly when datasets are highly imbalanced [191]. From the standpoint of applications, the nature of the imbalance can be divided into two categories: data that is naturally imbalanced (e.g., credit card frauds, earthquakes, shuttle failure and rare diseases) or data for which it is too expensive to obtain data on the minority class for learning such as natural disasters prediction, or uncommon events prediction such as volcanic eruptions or tsunamis, may require historical data or expert knowledge, which could be sparse or expensive to obtain [191]. This is also the case for FND where the number of fake news available is much less represented in the available data.

Several techniques have been proposed to address the issues associated with class imbalance. The three main techniques that can be applied are resampling techniques (or data pre-processing), algorithmic level techniques, and data post-processing techniques [51]. The solutions most commonly used are the data pre-processing or algorithm-level techniques.

In data preprocessing techniques, sampling is applied to the training data to add new samples or remove existing ones. These techniques aim to change the training data distribution to force the learning algorithm to focus on the most relevant class. This change in the training data can be accomplished through over- and/or under- sampling. Oversampling is the process of adding new samples to the training data while under-sampling is the process of removing samples. Figure 3.14 and Figure 3.15 illustrate the random under-sampling and random over-sampling techniques. These techniques act by randomly removing cases or adding copies of existing cases.

In Random Undersampling, examples from the majority class are randomly removed from the training dataset until the class distribution becomes more balanced. This can be achieved by randomly selecting examples from the majority class and removing them from the training dataset. Random under-sampling can be a simple and quick technique to address class imbalance, but it may result in the loss of valuable information from the majority class, leading to a potential loss of predictive performance.

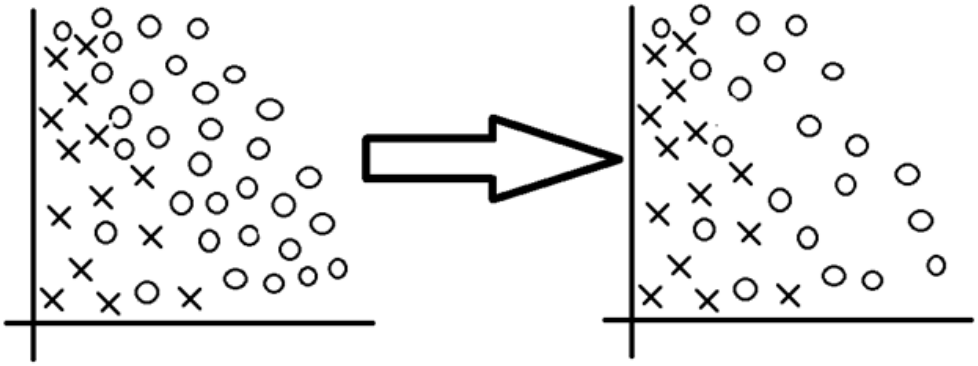

Figure 3.14: Random undersampling

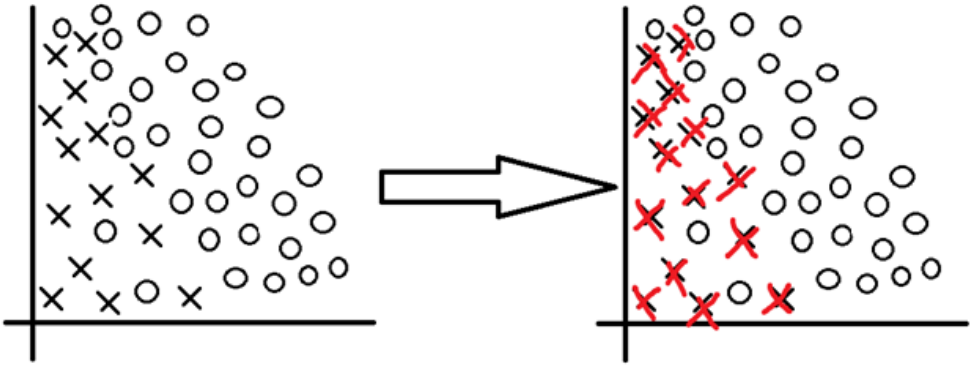

Figure 3.15: Random oversampling

In Random oversampling, examples from the minority class are randomly duplicated or synthetically generated to increase their representation in the training dataset. This can be achieved by randomly selecting examples from the minority class and duplicating them or generating synthetic examples using techniques such as Synthetic Minority Oversampling Technique (SMOTE) [65] or Adaptive Synthetic Sampling (ADASYN) [112]. Random oversampling can help in increasing the representation of the minority class, but it may also result in overfitting or amplification of noise if not done carefully.

The second method for resolving class imbalance is to create or modify an existing algorithm. Instead of changing the distribution of the training data, the change is applied to the learning and the decision process by increasing the importance of the positive class. The cost-sensitive method and recognition-based approaches, kernel-based learning, such as SVM and radial basis function [66], are among the algorithms that have been adapted to

address the class imbalance problem. Typically, specially developed algorithms for dealing with the class imbalance issue will work very well for a specific domain for which they were thought. However, they will fail under other domains and they require a thorough understanding of the algorithm to implement the modifications [51].

Our findings show that the use of various imbalance techniques, such as oversampling and downsampling, has shown promising results in improving the performance of different classifiers, including RNN variations, CNN, and hybrid models like CNN+LSTM. The results indicate that oversampling has been effective in improving the accuracy of LSTM and CNN models in [18], achieving accuracies of 95.51% and 98.96%, respectively. Similarly, oversampling has also been beneficial for BERT, achieving an accuracy of 94.66% in [129].

Moreover, the use of SMOTE oversampling has demonstrated effectiveness in dealing with class imbalance. For instance, in [45], SMOTE was used to improve the performance of a DNNs model, achieving an accuracy of 98% on the Politifact dataset.

Additionally, another study [15] utilized the focal loss function to prevent classification bias towards the majority class, which significantly improved the performance of their models on imbalanced datasets.

Downsampling has shown effectiveness in improving the training accuracy of hybrid models like CNN+LSTM on PHEME and FN-COV datasets in [147], achieving accuracy rates of 91.88% and 98.62% respectively. Additionally, downsampling has improved the accuracy of CNN and LSTM models in [281], achieving accuracies of 92.38% and 93.56% respectively.

It should be noted that despite the effectiveness of class imbalance techniques in improving the accuracy of FND models, a significant portion of the literature has not thoroughly investigated or addressed this issue. From our findings, only five research articles investigated the class imbalance effect on FND. This highlights the need for further research and exploration of various imbalance techniques to better understand their impact on model performance and generalizability in the context of FND.

3.6.2 Challenges related to the class imbalance problem in fake new detection

Class imbalance is a common problem in multiple application domains, and FND is not an exception. However, it has not received as much attention as it deserves in the context of FND, which we consider a big challenge to be addressed. The imbalance between real and fake news samples in the dataset can lead to biased classification, where the model

performs well on the majority class but poorly on the minority class [51]. Even when considering the usage of DL models, it was shown that the class imbalance problem will still affect the performance of the models [98].

One of the challenges related to the class imbalance problem in FND is the issue of using adequate performance assessment metrics to evaluate the model's performance. Traditional metrics such as accuracy can be misleading, as the model may perform well on the majority class but miss out on correctly identifying the minority class. This issue emphasizes the need for specialized metrics such as F1 score, precision, and recall [95].

In the FND field, there is a lack of systematic studies that evaluate the impact of known techniques for dealing with class imbalance. Techniques such as oversampling, undersampling, and ensemble methods have been widely used in many other domains. However, their effectiveness in FND remains understudied. Therefore, more research is needed to explore the effectiveness of these techniques in the FND field. An important challenge with the application of these techniques for FND is related to the generation of fake news texts. In this case, it is necessary to generate complete texts that look like real news, but it is also necessary to generate texts that correspond to fake news. This leads to another challenge connected to the need to carefully craft the synthetic text generation so that it corresponds to either fake news or real news. In particular, the generated fake news articles should be realistic and representative of the actual fake news articles to ensure the effectiveness of the model.

The context is also a challenge when considering the generation of fake news. Since fake news is often generated in response to specific events or situations, it can be difficult to apply generic techniques for dealing with a class imbalance that does not consider the specific context in which the fake news was generated.

Lastly, special-purpose algorithms that can deal with the class imbalance problem have not been explored or evaluated for FND. These algorithms include cost-sensitive learning, manipulating the loss functions, or building ensembles that are specially developed to address the class imbalance problem [51]. These techniques have shown promising results in multiple domains, and their effectiveness in FND requires further investigation.

In conclusion, addressing the class imbalance problem in FND is crucial for developing accurate and reliable models. Still, not much research has been done to address this problem. Researchers and practitioners need to pay more attention to this problem and explore various techniques to overcome it. This is a possible area where future researchers should focus on that may lead to improved solutions for FND.

3.7 Answers to Research Questions

In this section, we attempt to answer the research questions presented in Section 3.2.2 based on our findings. The detailed answers are described below.

RQ1: Which algorithms are used for fake news detection throughout time?

Given our findings, DL models are considered effective models in FND. There is a notable increase in the number of articles that address the different models and architectures for this task. We also noticed that the research focus shifted towards DL models for FND during the global COVID-19 Pandemic in 2021 which forms about 83% of the research effort that was conducted on FND. The remaining 17% of the FND research was conducted before this year.

Our findings also show that fake news can be detected by CNN, RNN, GRU, LSTM, and BERT models in many variations and with different architectures. We noticed that LSTM/BiLSTM were the models that appeared more frequently in the surveyed articles. The detection was also examined using hybrid models, which increased the detection effectiveness at some points. It is also noticeable that using the BERT model in the detection of fake news exhibits a huge positive impact on the detection effectiveness.

RQ2: Which datasets are used in the fake news detection domain?

The most difficult part of detecting fake news is the absence of a labeled dataset with trustworthy ground-truth labels with accepted size [300]. For several usages in DL, researchers have attempted to collect datasets over the past few years. The collected datasets are massively varied from each other due to the purpose of the study. For instance, some of these datasets are political and consist of political statements, as is the case in PolitiFact. Other datasets are built with news articles collected in a specific time frame, while other datasets include social media posts such as Twitter. Moreover, fake news is frequently collected from duplicitous websites intended to disseminate misinformation. This fake news will end up being shared on social media platforms by its creator. This fake news will also be shared by other individuals unintentionally without checking the news source or by other malicious users and bots.

Our findings show that Liar, ISOT, PHEM, and FakeNewsNet (with their three variations) are the most popular datasets used in fake news detection. These six datasets have been used in about 80% of the surveyed articles. We also noticed that researchers frequently attempted to create their own dataset to reach the required size and the domain which is obvious in about 45% of the surveyed studies. Other researchers combined two or more datasets to have an acceptable-sized dataset.

It is also worth mentioning that selecting a suitable dataset is crucial in FND since it impacts detection effectiveness. It is noticeable from our findings that the application of the same detection model in different datasets has an enormous difference in detection accuracy [25, 82, 93, 100, 125, 149, 154, 161, 170, 171, 201, 230, 256, 263, 268, 274].

RQ3: How effective are deep learning methods for FND

The researchers studied various DL algorithms to detect and classify fake news as mentioned above. These algorithms include CNN, RNN (with variations), GNN, BERT, Attention-based mechanisms, and various hybrid approaches. The detection effectiveness of these algorithms is influenced by the datasets used and the combination of different architectures for detection.

CNN and BiLSTM have been the most used detection models and have achieved the highest detection accuracy compared to other approaches. RNN, including their variations such as LSTM/BiLSTM and GRU, are utilized with considerable effectiveness in approximately 70%. Their ability to maintain information over sequences allows them to better understand context, which is essential for identifying fake news. CNN on the other hand have proven to be effective for FND tasks, appearing in 61% of the research articles we surveyed. BERT and hybrid detection models have also made noticeable detection effectiveness appearing in about 47% of the surveyed articles. FNNs and GNN were also used in the detection process even though not in many studies.

It is worth mentioning that in one research article, many DL models were developed to draw comprehensive conclusions. Hence, the total percentage of all the models that appeared in the surveyed articles is more than 100%.

The detailed effectiveness of DL detection models in the fake news field is in Section 3.3.

RQ4: Which solutions incorporate transfer learning mechanisms, if any?

TL is the process of exploiting what has been learned in one task to improve the generalization in another task [354]. The goal of TL is to improve learning the target task by leveraging knowledge from the source task.

TL is not applied in many FND studies as our findings show. Only seven research studies examined the effect of TL on the detection accuracy [28, 84, 129, 163, 205, 237, 314]. However, utilizing TL strategies enhanced the detection accuracy. The TL that was utilized in the FND field can be categorized under fine-tuning pre-trained models, using CNN-based architectures, employing pre-trained hybrid models, and leveraging transformer-based models. The highest improvement presented by using TL was reaching an accuracy

of 93.2% when applying the Alexnet pre-trained model which represents an improvement of 23.1% compared to the baseline case which is performed without applying TL.

It is worth mentioning that the application of the same detection model on different datasets recorded enormous differences in detection accuracy [25, 82, 93, 100, 101, 125, 149, 154, 161, 171, 201, 230, 256, 263, 268, 274]. This issue might be tackled by including a TL approach so that the detection model can report an approximate accuracy.

RQ5: Which solutions deal with different levels of an imbalanced dataset?

A dataset with a skewed class distribution where the end-user preferences are biased towards the least represented class(es) suffers from a class imbalance problem. A model learned under these conditions will focus on the majority class and will not learn correctly the minority and important classes [51].

Most of the available fake news datasets are imbalanced. From the articles we surveyed, only seven articles specifically treated class imbalance and studied its effect on FND utilizing various strategies to handle this issue. These strategies were random and advanced oversampling in four articles, random undersampling in two articles, and using a different loss function in one article. Oversampling has been used by increasing the number of instances in the minority class to match the majority class, improving the detection effectiveness [18, 128, 129]. In addition, advanced oversampling techniques such as SMOTE generate synthetic examples of the minority class by interpolating between existing instances rather than duplicating existing ones. SMOTE helped the trained model to achieve 95% detection accuracy with a noticeable improvement compared to the baseline case without treating the class imbalance [45]. This helped mitigate the risk of overfitting and enhanced the model's generalization ability.

Another strategy presented to balance the dataset was the random undersampling which involves reducing the number of instances in the majority class to match the minority class [147, 281].

Finally, the focal loss function is designed to address the class imbalance by down-weighting the loss assigned to well-classified examples and focusing more on hard-to-classify instances [15]. This approach helps to prevent the model from becoming biased towards the majority class and ensures that the minority class instances are given appropriate attention during training.

Handling the imbalanced dataset achieved a better accuracy result compared to the baseline cases that do not deal with the class imbalance. Thus, this is a relevant area for further research that can lead to improved solutions for FND.

3.8 Threats to Validity

SLRs are prone to several threats to validity that can lead to bias in review outcomes. These threats are publication bias and errors in data collection, study exclusion, and data extraction. Regarding publication bias, studies with positive results are more expected to be selected over negative studies. This issue is alleviated by attempting to determine whether the studies discuss their results and limitations. Moreover, the sole purpose of this SLR is to report the effectiveness of DL models rather than present new results. In addition, there is no motivation from our SLR to select studies reporting only positive results.

Regarding filtering out studies based on the search criteria, our objective was to have a broad search query as mentioned in Section 3.2.3 to alleviate this threat. We could also expand the survey date range to include studies published before 2018. However, fake news became more popular from 2018 onward, and our objective was to provide an updated review of the most recent trends in this application domain. This motivation is supported by Figure 3.3, which shows a notable increase in publications on the detection of fake news over time.

Regarding the issue of incorrectly excluded articles and data extraction, we alleviated this problem by asking another researcher to review some random studies. There is no rule for determining the number of articles for the random check task, but about half of the surveyed articles were selected for this special check.

3.9 Main gaps and open issues

From our investigation, we collected a list of the main gaps and open challenges that still deserve the attention of the research community for FND problem. We must highlight that this is a challenging task, involving several difficulties which we describe to allow future researchers to focus on the most important open issues.

- ***Lack of labelled data:*** One of the major challenges in training DL models for FND is the limited availability of labelled data [59, 339]. Fake news datasets are often small, and obtaining accurate and comprehensive annotations for training can be challenging. This can impact the performance and generalization of DL models, as they heavily rely on large amounts of labelled data for effective training.

- ***Potentially biased datasets:*** Another issue in FND is the potential bias in the datasets used for training and evaluation [158]. Fake news datasets may contain inherent biases, such as political or cultural biases, that can affect the performance and fairness of DL models. It is essential to carefully curate and preprocess datasets to mitigate these biases and ensure the reliability and generalizability of the models.

- ***Lack of benchmarks:*** There is a lack of standardized benchmarks for evaluating the performance of DL models in FND. The absence of benchmark datasets, evaluation metrics, and protocols makes it challenging to compare the performance of different models and assess their effectiveness [25, 364]. The development of standardized benchmarks can facilitate fair and rigorous comparisons and foster advancements in the field.

- ***Long-term performance of FND models:*** One critical area that requires further attention is the long-term performance of FND models. This issue was not fully studied in the literature. It reveals that the performance of the detection models can fluctuate significantly over time due to changes in the nature and topics of the data. Guimarães at al. [102] studied this topic and stressed the necessity of evaluating models based on their immediate accuracy, robustness and stability over extended periods. Current research often focuses on short-term performance metrics, neglecting the potential degradation of model effectiveness as the data landscape evolves. Addressing this gap is crucial for developing more reliable and adaptable fake news detection systems. This should incorporate longitudinal evaluations to better understand the dynamics of FND and ensure that models maintain high performance in the face of changing data trends.

- ***Transfer learning solutions not sufficiently explored:*** TL, which leverages pre-trained models for feature extraction or model initialization, has shown promise in improving the performance of DL models for various tasks [298, 321]. However, in the context of FND, the exploration of TL solutions is still limited. There is a need to further investigate and optimize TL approaches for FND to leverage knowledge from related tasks and domains.

- ***Class imbalance not adequately addressed:*** Class imbalance, where the number of samples in different classes is significantly imbalanced, is a common issue in FND. DL models trained on imbalanced datasets may result in biased and inaccurate predictions, as they tend to be biased towards the majority class [6, 162]. Although some studies have explored imbalance techniques such as oversampling or undersampling, the effectiveness of these techniques in DL for FND needs further investigation.

- ***Limited understanding of fake news dynamics:*** Despite extensive research on fake news, there is still a limited understanding of the complex dynamics and mechanisms underlying the spread and impact of misinformation [221]. DL models for FND may be limited by the lack of a comprehensive understanding of how fake news is created, disseminated, and received, which can impact the models' accuracy and effectiveness. Further research is needed to better understand the underlying dynamics of fake news and inform the development of more effective solutions.

- ***Real-world applicability:*** While DL models for FND show promising results in controlled research settings, their real-world applicability, and effectiveness in detecting fake news in diverse and dynamic environments, such as social media or online news platforms, is still a challenge. Real-world factors, such as varying levels of information quality, diverse sources of misinformation, and rapid information spread, can impact the performance and reliability of DL models in practical scenarios [315,352].

3.10 Summary

This chapter provided a comprehensive SLR on the use of DL methods for FND. With the increasing prevalence of fake news on social media platforms, researchers have turned to advanced DL algorithms to improve the accuracy and efficiency of detecting false information. The purpose of this chapter was to synthesize existing studies, explore the strengths and limitations of current methodologies, and identify gaps that require further research.

The chapter began by investigating existing FND strategies that use DL, focusing on publicly available datasets and their NLP approaches. It gathered information about the TL techniques applied and the methods used for addressing class imbalance, examining their effect on detection accuracy.

Key findings from the SLR included the extensive use of various DL models such as CNN, RNN, LSTM, and BERT for FND. The review highlighted the superior performance of DL models in handling large datasets and extracting complex features compared to traditional ML methods. Several commonly used datasets for FND were identified, including LIAR, ISOT, PHEME, and FakeNewsNet. However, the scarcity of large, well-annotated datasets was noted as a limitation, which affects the generalizability of the models. TL was found to be a promising approach for improving FND, particularly in scenarios where collecting training data is challenging or expensive. Various TL strategies and their effectiveness in enhancing model performance were discussed. The class imbalance problem, where the number of instances in one class significantly outnumbers those in another, was

identified as a critical issue in FND. The chapter outlined various strategies to address this issue, including oversampling, undersampling, and cost-sensitive learning.

Despite the advancements in DL for FND, several gaps were identified. Many existing models perform well within specific datasets but struggle to generalize when applied to news from different domains, highlighting a lack of robustness in real-world applications. The potential bias in datasets was also noted, which can affect the performance and fairness of DL models, indicating a need for more diverse and representative datasets to ensure reliable and unbiased FND. While various strategies for handling class imbalance were discussed, their effectiveness in FND needs further investigation, and more research is needed to develop robust methods for dealing with imbalanced datasets.

In the next chapter, we will build on these findings by evaluating DL models for cross-domain FND, addressing the issues of dataset generalization and class imbalance. We will conduct extensive experiments to assess the performance of various DL models across multiple datasets, providing insights into their robustness and effectiveness in real-world scenarios. This evaluation will lay the groundwork for developing novel approaches to enhance the accuracy and reliability of FND systems.

Chapter 4

Evaluating Deep Learning for Cross-Domains Fake News Detection

Building on the SLR presented in the previous chapter, which highlighted key DL models and challenges in FND, this chapter shifts focus to evaluating the performance of these models across multiple domains. In this chapter, we evaluate various DL models for cross-domain FND. Although many models perform well when trained and tested on the same dataset, their performance often degrades significantly when applied to news from different domains. This is known as the cross-domain generalization problem. Cross-domain generalization is crucial to creating robust FND systems that can be deployed in real-world scenarios where fake news spreads across diverse platforms, regions, and topics.

The chapter begins by reviewing the relevant literature on cross-domain learning, outlining the challenges of domain adaptation, and introducing the DL models selected for our study. We then discuss the datasets used in the experiments and present the results of model evaluations both within the same domain and across different domains. In this chapter, we demonstrate that while the selected models show strong detection capabilities on the same dataset, their performance significantly declines when applied to different datasets and domains, underscoring a critical issue with generalizability. The findings of this chapter lay the groundwork for the development of novel approaches, which will be discussed in later chapters, particularly BERTGuard and ProFineLlama.

This chapter is based on the following published paper:

- Alnabhan, M.Q., Branco, P. (2024). Evaluating Deep Learning for Cross-Domains Fake News Detection. In: Foundations and Practice of Security. FPS 2023. Lecture

Notes in Computer Science, vol 14552. Springer, Cham. https://doi.org/10.1007/978-3-031-57540-2_4

4.1 Introduction

With the growing usage of the internet, the prevalence of fake news has reached unprecedented levels. In the past, before the advent of social media platforms, the dissemination of false information was less common and more challenging due to limited channels of communication, relying mostly on word of mouth or printed media [24]. Fake news refers to the deliberate spread of inaccurate information through social media platforms, often aimed at persuading readers to believe the fabricated content. It is commonly motivated by political gain or financial profit through advertising [24]. Nowadays, the ease of publishing content on social media without regulation or scrutiny has contributed to the proliferation of fake news [56]. Platforms like Facebook and X (formerly Twitter) have become popular vehicles for spreading misinformation, as they are extensively used by individuals and influencers to share opinions, videos, and various activities [91, 130].

The surge in fake news gained significant attention in 2016, particularly during the lead-up to the US presidential election [291]. Consequently, detecting and combating fake news on social media networks has become a subject of great interest among researchers. However, detecting fake news is an extremely complex task that requires sophisticated models to compare and analyze related or unrelated information against known truthful data [358]. The challenges are not restricted to the models used for the detection. The concept of fake news also encompasses various forms, which brings an added difficulty, potentially leading to different approaches and solutions in addressing this issue. Terms like misinformation, rumours, spam, and disinformation are often used interchangeably, encompassing numerical, categorical, textual, and image-based content [44, 114, 196]. Moreover, the recent developments in NLP for text generation do not make the detection of fake news easier [114].

The goal of researchers is to curb the spread of misinformation by identifying the diverse methods employed in its dissemination. To achieve this, researchers have turned to DL algorithms to detect and mitigate the spread of fake news [148]. This involves curating or generating datasets comprising both true and false articles and developing models that can predict whether a given article contains genuine or fabricated information.

There are noticeable gaps in the existing studies that were conducted on DL for FND. A core gap is a lack of understanding of the generalizability of DL models that allow them to

achieve an acceptable detection accuracy over different fake news datasets. To investigate this matter, we conducted an extensive amount of experiments using different DL models across multiple publicly available datasets for FND. We aimed to study the effectiveness of these models in a particular dataset (i.e., training and testing them on the same dataset) and across datasets (i.e., training them on one dataset and testing them on a different one).

Key Contributions

The main contributions of this chapter are as follows:

- we provide an extensive analysis of the performance of DL models in detecting fake news when they are trained and tested on the same dataset;
- we provide an extensive analysis of the performance of DL models when they are trained on one dataset and tested in a different one;
- our experiments are the first to compare the performance on a large number of datasets and DL algorithms for this particular problem.

Chapter Organization

This chapter is organized as follows. Section 4.2, presents a background and literature review. In Section 4.3, we discuss the steps that we follow in our study to build a comprehensive study of the generalizability of FND models. Section 4.4 analyses the results obtained and answers our research questions. Lastly, Section 4.5 summarizes the chapter.

4.2 Background and Literature Review

Fake news is fabricated news in the form of articles, or videos distributed throughout social media which aim to imitate real news media content [18]. Such fake news is deliberately intended to mislead and persuade the readers into believing the content presented [218]. Fake news can be obtained from a variety of tools which include cyborgs, trolls, and social robots [218]. Social robots are what exploit the fake news, and the exchange of discussions among the readers. These bots are initially generated as computer algorithms to contribute to online discussions and continuously spread fake information. As such, fake news is a danger due to its ease of spreading while potentially covering a high diversity of topics and reaching a large volume of sites. This can have a direct impact on readers by touching

on topics related to health, democracy, journalism, finance, and social issues among many others [165].

To detect fake news, researchers tend to develop DL models. Recently, more attention has shifted towards the usage of DL as opposed to other traditional ML methods. In traditional ML, features are manually constructed, which is time-consuming and can lead to biases [250]. However, DL requires a larger dataset to train the model [250]. On the other hand, DL has shown a high detection performance for fake news by automatically extracting useful features. A variety of DL models have been used for FND including CNN [171], RNN [77], BERT [293], Attention Mechanism [257], and GNNs [244]. We previously detailed these models in Chapter 3, Section 3.3.

Researchers claimed that the main challenge in the FND task is the absence of a comprehensive benchmark dataset containing reliable ground truth labels and the limited availability of large-scale datasets [300]. Moreover, only a few datasets are available online for FND, having varying labels, sizes, and application domains [300]. Some datasets consist solely of political statements, others are from social media posts, and there are cases where the source is news articles. This diversity in the fake news domain is a serious challenge. These datasets are collected over a few years and are utilized by DL and ML for different purposes. Some of the well-known publicly available datasets for FND are LIAR [345], ISOT [9], PHEME [377], FakeNewsNet dataset [300], FakeNewsChallenge dataset [350], FA-KES [275], Fake-or-Real (FoR) [90], Fake News Detection [142], etc. We detailed the datasets used for FND comprehensively in Chapter 3, Section 3.4.

4.3 Research Methodology

4.3.1 Research Questions

To achieve our goals, we propose to answer the following research questions:

RQ1. What is the effectiveness of DL models in detecting fake news when they are trained and tested on the same dataset?

RQ2. How do these models perform when they are trained on one dataset and tested on a different one?

To answer these research questions, we selected five general NN architectures from the literature and seven datasets for FND. We provide the experimental details in the next sections.

4.3.2 Selected Datasets

We selected the following seven most popular and publicly available datasets related to fake news:

1. **LIAR [345]:** This dataset is made up of carefully labeled brief assertions, containing 12.8K labeled short phrases. The Politifact site handled the annotation process for this dataset, assuring reliable and accurate labeling. It has six classes: "pants-fire", "false", "barely-true", "half-true", "mostly-true", and "true." We binarized this dataset using the strategy in [268]. The labels "true", "mostly true", and "half-true" were merged under the "real" label. We also considered "fake" instead of "pants-fire", "false", and "barely-true".

2. **ISOT [9]:** This dataset's real news cases were collected between 2016 and 2017 by crawling news articles from Reuters.com. The fake news examples were obtained from unreliable websites, which were annotated by the Politifact site. It includes over 44k articles in each CSV file; "True.csv" and "Fake.csv", with articles primarily focusing on political news. The dataset contains subjects related to one of six types of information (world-news, politics-news, government-news, middle-east, US news, left-news).

3. **PHEME [377]:** This dataset was collected from X (formerly Twitter) based on nine newsworthy events classified by journalists. The data is structured in directories. Each event has a directory with two subfolders for true and fake claims. The annotation process was conducted by journalists (human annotators), and each tweet was annotated with one of the following labels: "proven to be false", "confirmed as true", or "unverified". We dropped the unverified data from this dataset.

4. **Fake News [144]:** This dataset contains verified news articles from various news resources. For each article, a label of 0 or 1 is assigned, with 0 corresponding to authentic news articles and 1 to fake ones. There were almost 10k articles in each category. For each news article, this dataset included the article ID, title, author's name, text, and the label.

5. **Fake News Detection [142]:** This datasets has about 4k records in a Kaggle repository. It has the news URLs, Headline, Body, and the Label; true or fake news.

6. **FA-KES [275]:** In 2019, a dataset was created specifically to showcase fake news related to the Syrian conflict. This dataset includes a collection of articles labeled

as either 0 (fake) or 1 (credible). The credibility assessment of the articles was determined by comparing them to ground truth information sourced from the Syrian Violations Documentation Center (VDC). The authors employed a crowdsourcing platform to extract information from each article, such as the date, location, and number of casualties. Subsequently, the authors cross-referenced these articles with the VDC database to ascertain their authenticity and determine if they were fake or genuine.

7. **Fake-or-Real News (FoR) [90]:** This dataset includes features such as the article title, text content, publication date, and source. The data contains almost 11000 articles tagged as either real or fake.

4.3.3 Deep Learning Models

We selected five representatives DL architectures based on their diversity and the recent proposal for news detection where they showed good performance.

4.3.3.1 Bidirectional long short term memory (BiLSTM)

A hybrid BiLSTM model with GloVe embeddings [242] and self-attention, combined with a dense layer, was proposed in [217]. This model consists of the following layers: Input Layer, Embedding Layer, 4 Dropout Layers, 2 BiLSTM Layers, 3 Batch Normalization Layers, Attention Layer, and 2 Dense Layers. The choice of BiLSTM over LSTM is based on the model's ability to process data in both forward and backward directions, allowing for effective correlation between words in text data [217]. The inclusion of self-attention further enhances the model's capacity to learn the relationships among different words. In addition, a fully connected dense layer is employed, adding complexity to the model. This was shown to be advantageous, enabling the model to better understand and distinguish between different classes.

4.3.3.2 Hybrid Convolutional and Recurrent Neural Networks (CNN + RNN):

This algorithm combines a CNN and an RNN [230]. Initially, a CNN layer with Conv1D architecture is employed to process the input vectors and extract local features at the text level. The output of the CNN layer, which comprises the feature maps, is then passed as input to an LSTM layer consisting of LSTM cells. This LSTM layer effectively captures

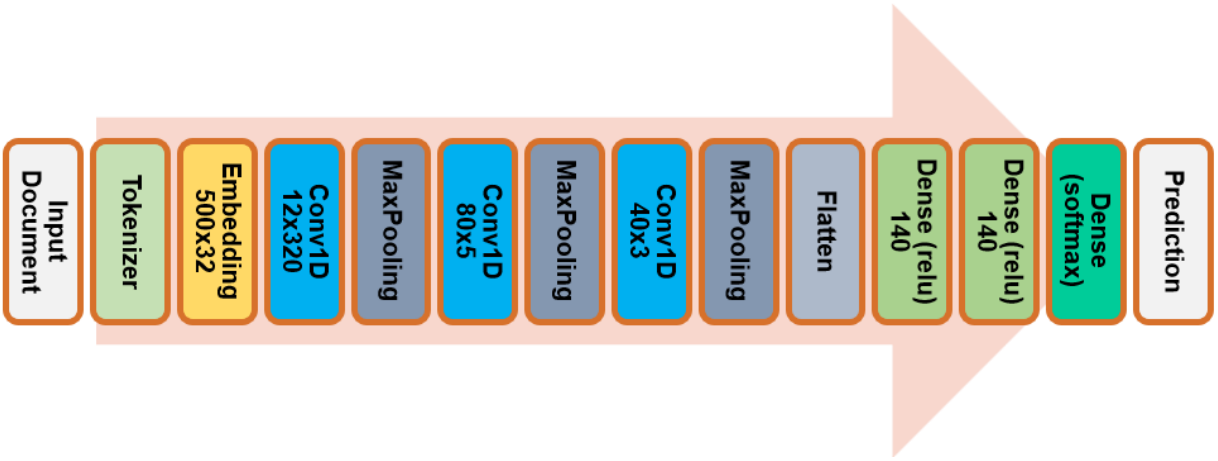

Figure 4.1: A CNN-based architecture for FND.

the long-term dependencies among local features obtained from the news articles, enabling accurate classification based on their authenticity.

4.3.3.3 Convolutional Neural Networks (CNNs):

A CNN consisting of multiple layers of neurons, forming a complete stack, and incorporating an embedding layer is proposed in [171]. Initially, each document is encoded as a collection of one-hot vectors, with the number of vectors set to 500, determined through experiments. This approach ensures a fixed length of 500 vectors. The classification head uses a dense layer with softmax activation. This layer receives a vector representation resulting from the input of a single document, as Figure 4.1 shows.

4.3.3.4 Hybrid CNN with LSTM (C-LSTM):

Kaliyar *et* al. [147] proposed to stack the CNN and LSTM layers in a unified structure named C-LSTM. Because multiple-branch convolution contains varied filters and kernel sizes for successful feature mapping and because the LSTM network provides a batch normalization process, this combination is beneficial. For effective detection of fake news, the LSTM layer can extract useful information from the convolutional layer. Different filter sizes across each dense layer were also considered. Figure 4.2 depicts the layered architecture of the proposed C-LSTM deep neural network.

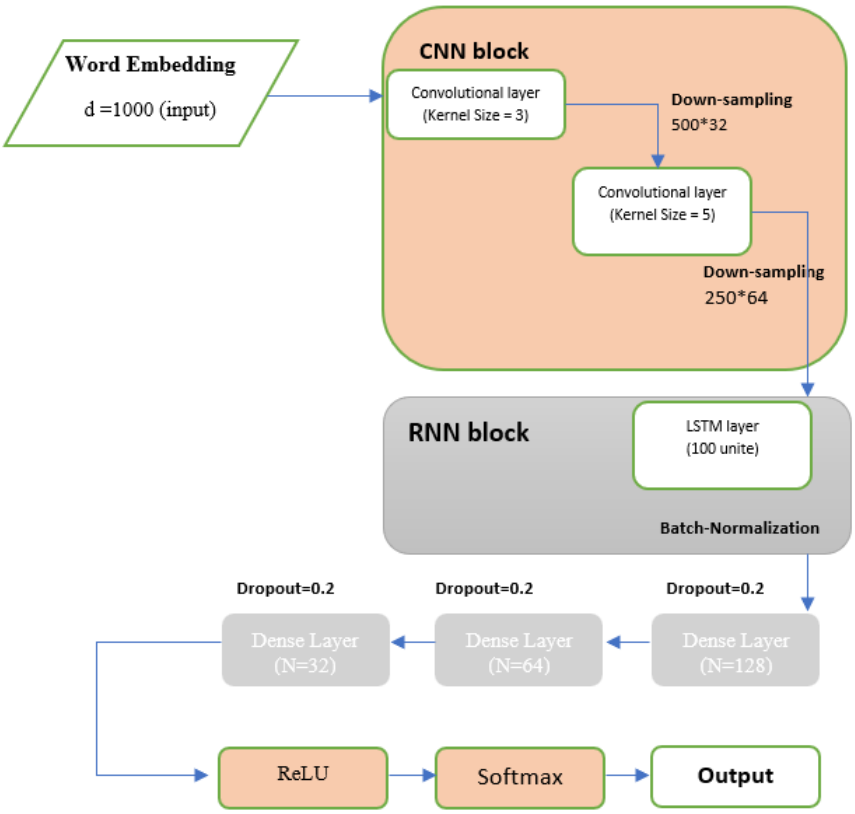

Figure 4.2: A C-LSTM architecture for FND.

4.3.3.5 Bidirectional Encoder Representations from Transformers (BERT)

To address the limitations of earlier FND approaches, which failed to capture semantic correlations between words, a language-based transformer model was introduced in [238]. The model incorporates the powerful bidirectional encoder representations from BERT model [79] used to extract textual features, allowing for the preservation of semantic relationships between words. We evaluate the usefulness of the BERT in the FND context.

4.3.4 Experimental Settings

Our methodology consisted of two main phases. In the first phase (intra-domain), each model was trained and tested on a specific dataset. We ensured a systematic approach by splitting the data into appropriate training and testing sets, incorporating the necessary

preprocessing steps, and monitoring the duration of training. The data split ratio was 7:1:2 for the training, validation, and testing sets, respectively. The hyperparameter tuning process involves evaluating and comparing the learning rate, number of epochs, optimizer, and batch size. The second phase (cross-domain) involved training each model on one dataset and subsequently testing it on all other datasets to assess its performance across multiple datasets and diverse domains.

The evaluation of the model performance was conducted using multiple metrics, including accuracy, precision, recall, and F1-score. These metrics were selected to provide a comprehensive assessment of the effectiveness of the model in detecting fake news.

4.3.5 Computational Environment

The experiments in this chapter were conducted on an ROG Strix G733ZW laptop powered by a 12th Gen Intel Core i9-12900H processor, an NVIDIA GeForce RTX 3070 Ti GPU, and 32GB of DDR5 RAM. Development and experimentation were carried out primarily using Python, with Jupyter Notebooks and VSCode serving as the main programming interfaces. A range of libraries, including NumPy, transformers, pandas, torch, tqdm, scikit-learn, pyarrow, accelerate, sentencepiece, ipywidgets, datetime, logging, imblearn, papermill, and XlsxWriter, facilitated various aspects of training and evaluating the model in addition to data preprocessing.

The computational setup was well-suited for training RNN-based models such as LSTMs, BiLSTMs, RNNs, and CNNs. However, BERT posed additional memory challenges due to its larger model size and parameter complexity. To optimize memory usage, gradient accumulation and batch size adjustments were implemented, ensuring efficient execution within the available GPU capacity. For particularly resource-intensive tasks, cloud-based computing resources were leveraged to enhance processing efficiency without compromising performance.

4.3.6 Evaluation Metrics

The performance of the system was evaluated using various metrics including Accuracy, Precision, Recall, F1-score, and Geometric-mean (G-Mean), as detailed in Table 4.1. These metrics were chosen to provide a detailed assessment of the effectiveness of the model in detecting fake news. We include accuracy as well as several metrics suitable for imbalanced domains. This allowed us to obtain a better understanding of the performance of the model in an imbalanced setting [95].

Table 4.1: Performance evaluation metrics.

Metric	Formula	Evaluation Focus
Accuracy	$\frac{tp+tn}{tp+fp+tn+fn}$	The proportion of correctly predicted positive and negative instances out of the total instances.
Precision (pre)	$\frac{tp}{tp+fp}$	Measures the positive patterns that are correctly predicted from the total predicted patterns in a positive class.
Recall (rec)	$\frac{tp}{tp+fn}$	The proportion of actual positive instances that were correctly predicted by the model.
F1-Score	$\frac{2*pre*rec}{pre+rec}$	The harmonic mean of precision and recall, balancing both metrics.
G-Mean	$\sqrt{Sn \times Sp}$	The G-Mean evaluates the balance between correctly identifying positive cases (sensitivity) and correctly identifying negative cases (specificity).

Note: *tp*—true positive; *tn*—true negative; *fp*—false positive; *fn*—false negative; *Sn*—Sensitivity; *Sp*—Specificity,

The tp value represents the number of correctly identified fake news instances. The fp represents the number of instances incorrectly identified as fake news when they are actually real. The tn represents the number of correctly identified real news instances, while the fn is the number of instances incorrectly identified as real news when they are actually fake.

The G-Mean is a crucial metric for evaluating performance in imbalanced classification tasks. Unlike accuracy, which can be misleading in imbalanced datasets, G-Mean ensures a balance between correctly classified positive and negative instances. It is particularly useful when the dataset has a class imbalance issue, as it prevents the model from favoring the majority class. In this thesis, G-Mean is mainly used when addressing class imbalance.

4.4 Results and Discussions

4.4.1 Models' Performance Within the Same Dataset

This section presents the results of training and testing each model in each dataset, i.e., one dataset is used for training a model and testing it. Figures 4.3, 4.4, 4.5, 4.6, and 4.7

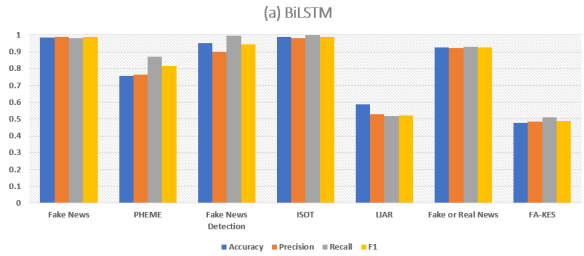

Figure 4.3: BiLSTM Performance

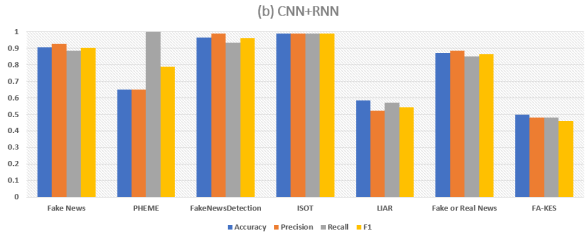

Figure 4.4: Hybrid CNN+RNN Performance

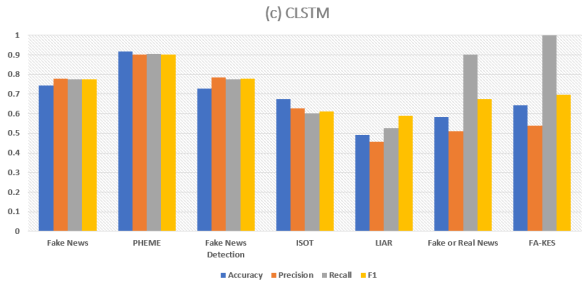

Figure 4.5: CLSTM Performance

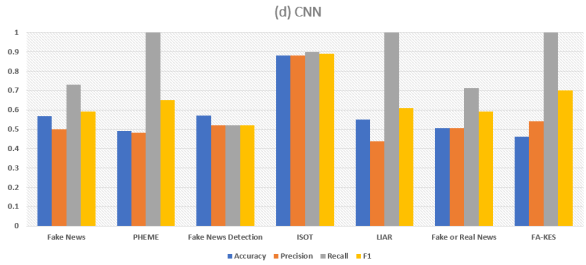

Figure 4.6: CNN Performance

show each model's performance when trained and tested in a particular dataset by showing the Accuracy, Precision, Recall, and F1 results for the selected models and datasets.

Our experiments show that the detection models have varying performances, working well on some datasets while performing not so well on other datasets.

BiLSTM achieved a detection accuracy of 98.6% on the Fake News dataset, the original dataset that was used to train this model [242]. It also performed well on other datasets such as Fake News Detection and ISOT achieving an accuracy of 95.1% and 98.9%, respectively. However, it was not successful in the other datasets. It achieved a low performance of 47.8% on the FA-KES dataset and 58.7% on the LIAR dataset.

Our experiments also show that the hybrid CNN+RNN model that was trained on ISOT achieved 99% detection accuracy while achieving a lower accuracy of about 50% in FA-KES and LIAR datasets. The same pattern appeared for the CNN model on different datasets other than the dataset ISOT that was originally used to train this model in [171]. It achieved 88% of detection accuracy on ISOT but its performance fell for all other datasets used.

BERT demonstrated a strong performance in the detection of fake news across multiple datasets, surpassing the other models previously examined. When originally trained

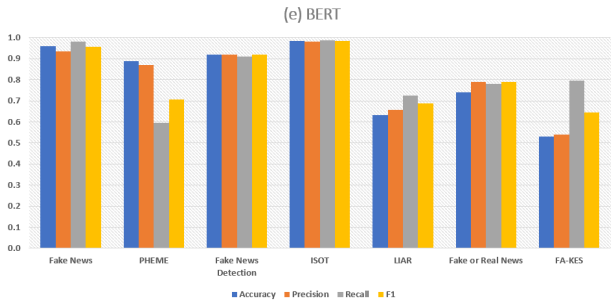

Figure 4.7: BERT Performance

on the FakeNewsNet dataset in [238], BERT achieved an accuracy of 92%. In our experiments, when we applied BERT to various datasets such as Fake-or-Real, Fake News, PHEME and ISOT, it consistently exhibited high accuracy rates of 74%, 95.8%, 88.8%, 98.3%, respectively. These results highlight BERT's ability to effectively capture contextual information and semantic relationships between words, enabling accurate detection of fake news instances. However, it is important to acknowledge that BERT's performance can be influenced by the characteristics of the dataset it is trained on. When trained on the LIAR dataset, BERT achieved a detection accuracy of 63.1%, while on the FA-KES dataset, the accuracy was 52.9%. These results emphasize the need for further research to better understand the factors that contribute to BERT's performance variation across different datasets and to enhance its effectiveness on challenging datasets.

4.4.2 Models' Performance Across Datasets

This section presents the performance of the selected models when trained on a particular dataset and tested on each of the remaining datasets. Our goal is to observe their performance across multiple datasets and diverse domains. Figures 4.8, 4.9, 4.10, 4.11, 4.12, 4.13, and 4.14 present the cross-domain testing results (i.e., training on one dataset and testing on the remaining ones (x-axis)). The accuracy is used to reflect these models' performance.

From the results we obtained, all models achieved superior results when the same dataset was used to train and test the model. The BiLSTM, trained on the Fake News dataset, achieved high accuracy (98.7%) when tested on the same dataset. However, its performance decreased with all other datasets, indicating limitations in its generalizability. In addition, BiLSTM struggled to detect fake news accurately in the ISOT and LIAR datasets that achieved an accuracy of 51.7% and 43.7%, respectively, suggesting a potential

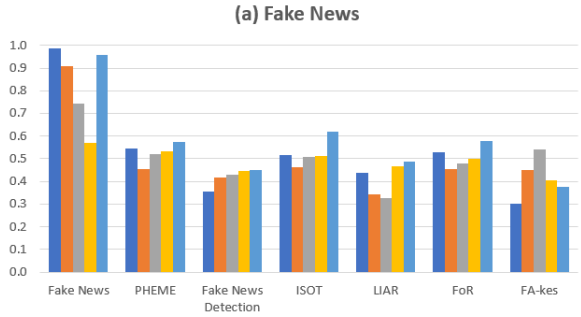

Figure 4.8: Models Trained on Fake News

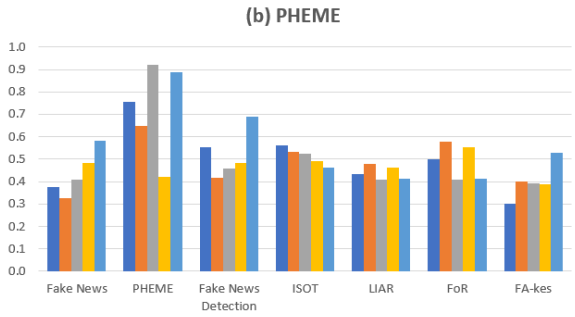

Figure 4.9: Models Trained on PHEME

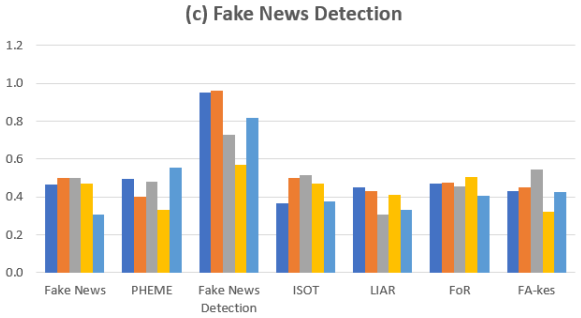

Figure 4.10: Models Trained on Fake News Detection

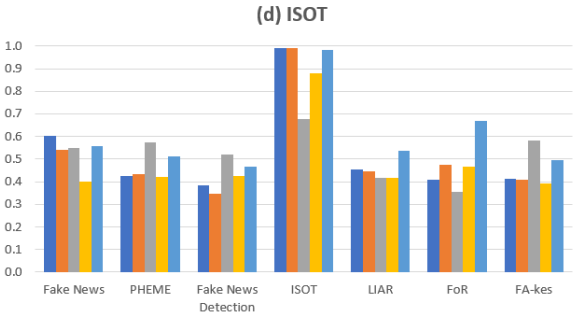

Figure 4.11: Models Trained on ISOT

challenge in the handling of various sources of deceptive information. The accuracy also decreased when the FND and FA-KES datasets were used to test the model.

The hybrid CNN+RNN trained on the Fake News dataset showed a competitive accuracy of 90.8% when tested on the same dataset. However, its performance varied across other datasets. Notably, the CNN+RNN model encountered difficulties in accurately identifying fake news in both the LIAR and the Fake News Detection datasets. This pattern persisted even when the model was trained on other datasets, such as the Fake-or-Real News, where it achieved an accuracy of 87.3% when evaluated on the same dataset. However, when tested on the FA-KES dataset, the accuracy dropped significantly to 25.1%. This pattern also appears in the C-LSTM case.

Notably, BERT's performance varied across datasets, suggesting that dataset characteristics can influence its effectiveness. Despite this variation, BERT consistently outperformed other models, demonstrating its robustness and effectiveness in FND. BERT achieved competitive accuracy rates, including in the Fake News (95.8%) and ISOT (98.3%)

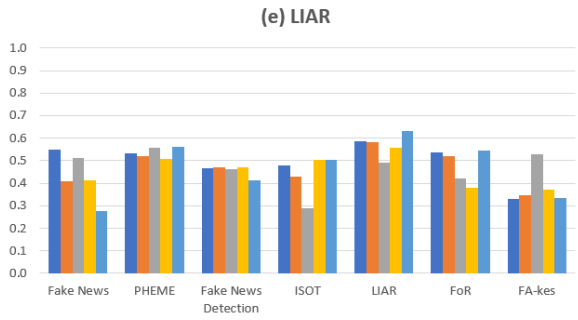

Figure 4.12: Models Trained on LIAR

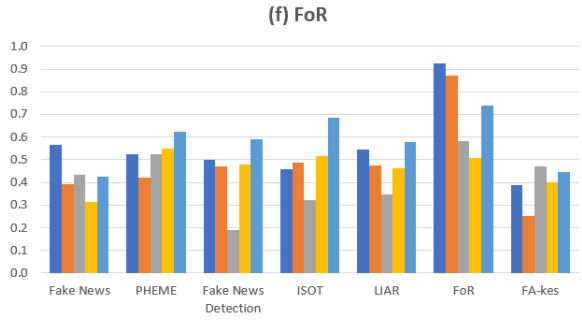

Figure 4.13: Models Trained on FoR

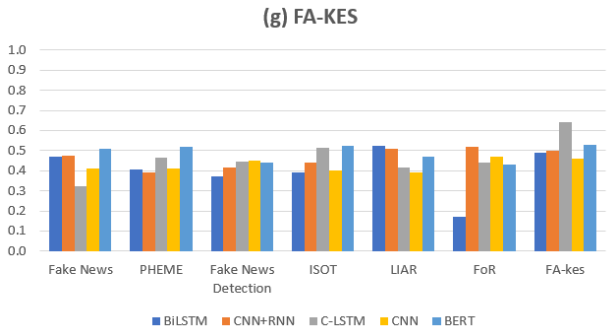

Figure 4.14: Models Trained on FA-KES

datasets. However, it achieved an accuracy ranging between 38% and 62.2% when the Fake News dataset was used for training the model and it was tested on the other datasets. These findings showcase the importance of utilizing advanced language representation models like BERT to enhance the accuracy and reliability of FND systems.

4.4.3 Research Questions' Answers

From the extensive experiments we run, the defined research questions in Section 4.3.1 can be answered as the following:

RQ1. *What is the effectiveness of DL models in detecting fake news when they are trained and tested on the same dataset?*

The selected detection models exhibit a notable pattern in their performance: they tend to work well on specific datasets but perform poorly on others. This indicates that

the effectiveness of these models in detecting fake news is highly dependent on the characteristics of the dataset they are trained on. For example, BiLSTM and hybrid CNN+RNN models achieve high detection accuracies when trained and tested on the same dataset, suggesting their ability to capture patterns specific to that particular dataset. However, their performance significantly declines when different datasets are used, indicating a lack of generalizability. In addition, BERT stands out as a more robust model in terms of its ability to adapt to different datasets, although further research is needed to understand the factors that contribute to its varying performance and improve its efficacy on challenging datasets.

RQ2. *How do these models perform when they are trained on one dataset and tested on a different one?*

Based on our experiments, we observed that the models' performance tends to decline when tested on datasets different from the ones they were trained on. For example, the BiLSTM and hybrid CNN+RNN models, which achieved high accuracies when trained and tested on the same dataset, experienced a huge performance drop when tested on different datasets. This suggests that these models struggle to generalize their learned patterns and fail to capture the specific characteristics of new datasets. Similarly, the C-LSTM model, despite performing well on the dataset it was trained on, faced challenges when applied to other datasets.

In contrast, BERT demonstrated relatively better and more stable performance when trained on one dataset and tested on other datasets. It consistently achieved high accuracy rates across multiple datasets, indicating its ability to capture contextual information and semantic relationships between words effectively. BERT's strong performance suggests its potential for generalization and adaptability to different sources of fake news.

Overall, our results confirm that specialized techniques such as TL should be applied to enhance the model's performance so that they can be used for FND in a real-world scenario.

4.5 Summary

This chapter sheds light on the efficacy of DL for intra- and cross-domain FND. To investigate the models' detection generalizability and robustness, we conducted an extensive set of experiments. We studied the models' performance when they were trained and tested on several datasets from different domains in the context of FND. Moreover, we also studied the performance of these models when were trained in one dataset and were tested on

other different datasets. As far as we know, this is the first study that carries out such an extensive set of experiments including a vast number of fake news datasets and DL algorithms. Our findings show that the selected models exhibit strong detection capabilities when trained and tested on the same dataset, highlighting their ability to capture dataset-specific patterns effectively. However, their performance notably deteriorates when applied to different datasets and on different domains, indicating a lack of generalizability. Overall, this chapter provides valuable insights into the performance of DL models for cross-domain FND, highlighting the need for continued research and development to advance the field and address the challenges associated with detecting fake news in various contexts.

In the following chapter, we will introduce BERTGuard, a novel two-tiered multi-domain FND system. This approach is designed to enhance the effectiveness of FND across multiple domains while addressing the class imbalance problem.

Chapter 5

BERTGuard: Two-Tiered Multi-Domain Fake News Detection with Class Imbalance Mitigation

In the previous chapter, we evaluated the performance of several DL models in cross-domain FND, revealing the challenges posed by class imbalance and generalization across datasets in addition to the effect of utilizing BERT for the detection of fake news. In this chapter, we introduce a novel two-tiered approach designed to enhance multi-domain FND while specifically addressing the challenges posed by class imbalance. As previous chapters highlighted, the effectiveness of DL models can be significantly hampered by the uneven distribution of fake news instances across various datasets. This two-tiered framework leverages a combination of model architectures to effectively capture different aspects of fake news, while also implementing techniques to mitigate the impact of class imbalance. The goal is to develop a more robust detection system that can operate effectively in real-world scenarios where data variability is commonplace. This chapter is based on the following published paper:

- Alnabhan, Mohammad Q., and Paula Branco. 2024. "BERTGuard: Two-Tiered Multi-Domain FND with Class Imbalance Mitigation" Big Data and Cognitive Computing 8, no. 8: 93. https://doi.org/10.3390/bdcc8080093

5.1 Introduction

Social media has become a primary source of news that is readily accessible to people worldwide. Today, sharing news content on social media platforms is as simple as clicking a button, leading to a constant stream of news from various domains being uploaded daily [302]. Unfortunately, the ease of uploading news content coincides with the increased spread of fake news. This poses a major problem in terms of reliability amongst consumers and can have negative impacts, not only on individuals but also on social mobility, potentially inciting widespread social panic [68]. In 2020, it was reported that one in every five USA adults relied on social media as their primary source for news about the 2020 USA presidential election [291]. Therefore, detecting fake news is essential for assessing the truthfulness of the content being consumed.

While human judgment remains effective for FND, ML models are increasingly employed by researchers for rapid and efficient FND [285, 290]. News gathered from social media has been categorized by topic into multiple domains, including health, social, politics, and entertainment [226]. While FND is important in all domains, some domains have extremely limited fake news [225]. Further, fake news from specific domains has an impact on society that is deemed more significant [226]. For instance, the dissemination of fake news within the political domain during the 2016 USA presidential election may have had a significant effect on the election outcome [24, 342]. Another example that led to a worldwide social panic [68] is linked to the influx of various information related to the COVID-19 pandemic. The impact of spreading fake news via various social media platforms led to a reduction in pandemic prevention measures [57]. As such, fake news within specific domains like political or health domains tends to be more significant than other fake news domains because of its ability to delve deeper, spread widely and more rapidly, and reach a larger audience.

To address the issue revolving around detecting fake news, a wide range of ML detection models have been suggested. Typically, these approaches concentrate on Single-Domain Fake News Detection (SFND) in areas such as politics or health, where an algorithm is specifically trained to identify fake news within a particular domain. However, models trained in this manner often produce inadequate results when applied to other domains [25]. Although it has been found that news pieces from various domains can be interconnected, which can potentially enhance the detection of fake news for the target domain, the use of SFND is currently viewed as a disadvantage [225, 302]. As such, multi-domain FND has been used to address the limitations of SFND [146]. Multi-domain FND allows for the utilization of various domains at once within a given dataset, leading to an improvement in the performance in all domains. However, this comes with great complexity, creating a

flaw within multi-domain FND such that the detection of fake news in some domains will be higher at the expense of other domains performing worse [320].

In addition, the majority of the available real-world domains' instances belong to the real news class (the majority or negative class), while a significantly smaller number corresponds to the opposite fake news class (the minority or positive class), which is the most important class [27,288]. This is known as the class imbalance problem, in which the dominant class has an advantage when training predictive models, causing them to disregard the minority class. The imbalance problem leads to inadequate classification performance, and most algorithms perform badly when datasets are substantially imbalanced [191]. The class imbalance problem also affects FND, as there is typically a lower representation of fake news instances compared to true news instances in the available data, which makes the model predict the true class (true news) almost all the time [27].

A critical gap remains from previous works, which is the lack of the details needed to train models for each domain comprehensively. Moreover, existing solutions typically train models on one or two datasets at most, which limits the scope of their analyses. Recognizing the lack of systematic and comprehensive work in multi-domain FND, this chapter introduced our newly invented BERTGuard, a holistic approach to multi-domain FND. BERTGuard combines datasets from a diverse set of news domains to build the groundwork for a complex detection system that functions on two levels. The first level of our detection system performs domain classification, which is a critical phase in which a BERT model is trained on a merged multi-domain dataset. This initial classification attempts to determine the exact domain to which a given news story belongs. The second level uses domain-specific BERT models to predict the news inside each designated domain. This hierarchical structure ensures that the detection process is tailored to the unique characteristics of each domain, increasing the overall system's precision and reliability.

To assess the performance and generalizability of our technique, we rigorously tested the invented detection system in five distinct domains: politics, health, science, crime, and social. We considered as baselines a single-level BERT model (Baseline 1) and the model proposed by [146] (Baseline 2). Finally, we assessed the impact of class imbalance on detection accuracy and explored three different handling approaches; random downsampling, random upsampling, and class weight adjustment.

Key Contributions

The main contributions of this chapter are as follows:

- **Introduction of BERTGuard**: We propose BERTGuard, a novel two-tier solution

for multi-domain FND utilizing BERT-based models, which addresses limitations in existing methods by improving the semantic understanding of textual data.

- **State-of-the-art performance**: We carry out an extensive evaluation of our proposed approach by comparing it against existing alternative solutions and demonstrate that BERTGuard outperforms existing methods by achieving a 27% improvement in accuracy and a 29% increase in F1-score across multiple benchmark datasets, establishing a new state-of-the-art in FND.
- **Class imbalance analysis**: We analyze the impact of three methods for dealing with the class imbalance inherent to these domains, providing valuable insights for future research and practical applications in FND.

Chapter Organization:

This chapter is organized as follows. Section 5.2 provides the background and a review of the literature. In Section 5.3, we present our proposed multi-stage solution for FND, BERTGuard. In Section 5.4, we outline the experimental methodology employed in our study, providing details on the datasets and the evaluation metrics utilized. Section 5.5 presents the experimental results and analyzes the findings. Lastly, Section 5.6 summarizes the chapter.

5.2 Background and Related Work

Transformer-based models have been explored for FND. BERT is a DL model designed to generate text, analyze sentiment, and understand natural language. It leverages transformer encoders to perform various NLP tasks [219]. BERT has been employed in previous studies to achieve superior results in various sentiment analysis tasks [187]. Despite being pre-trained on a large set of textual data, BERT needs to be modified to perform efficiently on a given task [362].

Several versions of BERT exist in the literature, each tailored to specific domains or tasks. Some notable BERT versions include:

- **RoBERTa**: This is a variant of BERT designed to improve the training process. It was developed by extending the training duration, using larger datasets with longer sequences, and employing larger mini-batches. The researchers achieved significant performance improvements by adjusting several hyperparameters of the original BERT model [177].

- **Distilled version of BERT (DistilBERT)**: This is a more compact, faster, and cost-effective version of BERT, which is derived from the original model with certain architectural features removed to improve efficiency [279]. Similarly to BERT, DistilBERT can be fine-tuned to achieve strong performance in various NLP tasks [195].
- **BERT-large-uncased**: This is a BERT model with 12 more transformer layers, 230 million more parameters, and a larger embedding dimension, allowing the model to encapsulate much more information than BERT-base-uncased [326].
- **BERT-cased**: This model was trained with the same hyperparameters as first published by Google [195].
- **ALBERT**: This is a streamlined version of BERT designed to reduce memory usage and accelerate training. It employs two primary parameter reduction techniques: dividing the embedding matrix into two smaller matrices and using shared layers grouped together [183]. ALBERT is recognized for its reduced number of parameters compared to BERT, which enhances its memory efficiency while still delivering strong performance. These variations highlight BERT's adaptability across various domains and its potential for tailored customization to specific tasks.

Table 5.1 summarizes the latest key advances utilizing transformer models in FND.

Table 5.1: The main transformer-based FND models.

Ref.	Year	Model	Covered Multi-Domain?
[25]	2023	BiLSTM, Hybrid CNN+RNN, CNN, C-LSTM, and BERT	No
[194]	2023	BERT+LSTM model for text content analysis GAT for modeling social network features	No
[48]	2023	BERT	No
[85]	2023	BERT + LightGBM	No
[301]	2023	SBERT, RoBERTa, and mBERT	No
[311]	2023	Hybrid ensemble learning model: BERT for text classification tasks Ensemble learning models, including Voting Regressor and Boosting Ensemble	No
[20]	2023	CT-BERT with BiGRU	No
[283]	2023	BERT, LSTM, BiLSTM, and CNN-BiLSTM	No
[168]	2023	DistilBERT and RoBERTa	No
[272]	2023	hybrid model combining LSTM and BERT with GloVe embeddings	No
[86]	2023	TF-IDF N-gram, BERT, GloVe, and CNN	No
[231]	2022	Arabic-BERT, ARBERT, and QaribBert	No
[258]	2022	BERT, XLNet, RoBERTa, and Longformer	No
[219]	2022	BERT with Reinforcement Learning (RL)	Yes Gossip (Social) and Political
[262]	2022	BERT	No
[330]	2022	BART and RoBERTa	No
[146]	2021	FakeBERT: Combination of BERT and 1d-CNN	Yes Social and Political
[286]	2021	ALBERT, BERT, RoBERTa, XLNet, and DistilBERT	No
[126]	2021	BERT	Yes Political, Social, and Health

A comparative study was carried out that involved five transformer models: BERT, ALBERT, RoBERTa, XLNet, and DistilBERT [286]. The authors performed a comparison with different combinations of hyperparameters and yielded equivalent results.

Another comparative study investigated the effectiveness of cross-domain models and found that models based on BERT achieved the best detection accuracy [25].

Subsequently, a DL model was developed combining BART and RoBERTa to differentiate between true and fake news articles [330]. The embeddings from both BART and RoBERTa were first processed through LSTM and CNN architectures, then concatenated and further processed through additional LSTM and CNN layers.

In addition, Kaliyar et al. [146] introduced FakeBERT, a DL model for the detection

of fake news that uses BERT in conjunction with 1D CNNs of various kernel sizes and filter configurations to improve classification performance. BERT is used to generate word embeddings, which are subsequently processed through three parallel convolutional layers with different kernel sizes. The dataset used encompasses both social and political news, with a focus on the USA presidential election of 2016. This approach based on BERT achieved an impressive detection accuracy of 98.9%, outperforming other ML techniques.

An attention-based transformer model has also been used to detect fake news [251]. The authors compared their approach with a hybrid CNN model that integrates both text and metadata. The transformer model demonstrated superior accuracy compared to the hybrid CNN model. However, transformer-based methods often involve significant computational costs and require large amounts of training data.

Using single datasets for the detection of fake news has several limitations. First, they provide limited coverage: single datasets can focus on specific topics or domains, such as political statements or social media posts, leading to a lack of diversity in the types of fake news covered [233]. Second, there is dataset bias: Datasets constructed only with specific types of news, such as political or e-commerce news, can lead to biased models that perform poorly when detecting news related to other topics, resulting in dataset bias [159, 233]. Third, there is a lack of labeled data: the shortage of labeled data for training detection models impedes the development of effective FND systems [262]. Fourth, there are problems with overfitting and generalizability: limited and imbalanced datasets can cause ML and DL models to overfit or underfit, affecting their ability to generalize and perform well [98, 106].

To overcome these limitations, it is crucial to utilize a diverse set of publicly available evaluation datasets for the detection of fake news and incorporate multiple domain-specific datasets. Implementing cross-domain, cross-language or cross-topic analyzes provides a comprehensive approach by incorporating datasets across various domains or topics, thereby enhancing the detection process and improving the generalizability of FND models [260].

Many studies on automated detection of fake news have relied on datasets confined to a single domain, such as political, social, or health domains, for model training and evaluation. This focus on a single domain is driven by the performance decline observed in these ML and DL techniques when they encounter unseen data from other domains. Domain-specific features, particularly style-based attributes, can vary significantly across different domains [260]; consequently, features must be carefully selected to distinguish between fake and real news within the specific domain being examined.

The current state-of-the-art highlights the need for further research across various do-

mains. As a result, comprehensive cross-domain techniques for FND are essential, despite some previous studies that attempt to tackle this issue using cross-domain data, such as the recent work by Alnabhan and Branco [25].

Other works exist that address the multiple-domain problem in detecting fake news. Han et al. [108] proposed a continuous learning approach for domain-agnostic FND that utilized a GNN to sequentially learn across multiple domains. However, this method has two drawbacks: (1) it assumes that new domains will arrive in a specific sequence, and (2) it assumes that these domains are already known, which is not the case in real-world data streams. In contrast, Cardoso et al. [60] introduced a method that retains knowledge across different domains by leveraging a robust, optimized BERT model to select informative instances for manual annotation from a large unlabeled dataset. Consequently, earlier studies explored the integration of information from multiple domains to develop a model to detect fake news across domains. For example, Castelo [62] trained a model on the Celebrity dataset and tested it on the US-Election2016 dataset to determine the generalizability of their approach.

Huang et al. [126] proposed a novel framework for detecting fake news in new domains called DAFD: Domain Adaptation Framework for FND. The framework combines domain adaptation and adversarial training strategies to align the data distribution of the source and target domains and enhance the model's generalization and robustness. The framework consists of a pre-training phase, where the data distribution alignment is performed, and a fine-tuning phase, where adversarial examples are generated to further improve the model's performance. Experiments conducted on real datasets, including PolitiFact, GossipCop, and COVID show that DAFD outperforms state-of-the-art methods for detecting fake news in new domains with a small amount of labeled data. The framework components were analyzed, showing that both domain adaptation and adversarial training are crucial to improving detection performance.

Mosallanezhad et al. [219] proposed a domain-adaptive model called Reinforced Adaptive Learning FND (REAL-FND). REAL-FND leverages generalized and domain-independent features to differentiate between fake and true news. This approach is based on previous findings that suggest that domain-invariant features can improve the robustness and generalizability of FND techniques. For example, fake news publishers have frequently used clickbait writing styles to capture the attention of targeted audiences, highlighting a domain-invariant characteristic. In addition, patterns derived from social contexts, such as a user's comment disputing a news article or interactions between users and identified fake news disseminators, offer critical additional information for classifying fake news within a specific domain. In REAL-FND, the approach departs from the conventional method of using adversarial learning to train cross-domain models. Instead, it employs

a RL component to transform the learned representation from the source domain to the target domain. Unlike other RL-based methods that modify model parameters, in REAL-FND, the RL agent adjusts the learned representations to obscure domain-specific features while preserving domain-invariant components. This method offers greater flexibility than adversarial training, as the RL agent can directly optimize the confidence values of any classifier without the need for a differentiable objective function.

5.2.1 Fake News Datasets

Researchers identify the primary challenge in the detection of fake news as the scarcity of large-scale datasets and the absence of a comprehensive benchmark dataset with reliable ground truth labels [300]. Furthermore, there are not many datasets with different labels, sizes, and application domains that can be found online for the detection of fake news. Certain datasets are sourced exclusively from political statements, while others come from posts on social media and even news articles. This diversity presents a significant obstacle in the field of fake news.

Additionally, acquiring datasets for academic research is challenging due to privacy restrictions on extracting data from online sources. One solution is to purchase data from these platforms or crowd-sourcing websites. Another approach is to utilize existing datasets from the literature that align with the study's requirements, such as ISOT [9], PHEME [377], Liar [345], GossipCop [300], FakeNews [144], Snopes [303], COVID FN [312], Politifact [216], Climate [54], and COVID-Claims [270].

5.2.2 Dealing with Class Imbalance

The class imbalance problem has received a lot of attention [51]. However, when observing the particular domain of FND, we verify that this issue still has received very little attention. Existing research has shown the prevalence of the imbalance problem as cases of fake news proliferate, sometimes exceeding those of true news [6]. This imbalance complicates the development of accurate and reliable methods for detecting fake news. The overfitting of NNs due to class imbalance is a key challenge, highlighting the need for investigation and improvements in this domain [106].

Although some studies have explored specific elements of FND models, including dataset division, features, and classifiers, there has been an insufficient examination of the limitations of datasets and features and their impact on detection models, particularly with respect to class imbalance [27, 106]. Furthermore, the precision of the detection models

is significantly unsatisfactory, with a low detection rate and a long detection processing duration [6, 233, 260].

In addition, research has been performed to minimize domain biases and increase the accuracy of cross-domain FND, which is connected to the difficulty of imbalanced data across different domains [159]. In addition, a comprehensive review highlights the constraints of current FND models due to imbalanced and limited datasets, highlighting the necessity for comprehensive cross-domain techniques to address these difficulties [233].

Furthermore, learning techniques such as resampling methods (e.g., oversampling and undersampling), data augmentation, and cost-sensitive learning have been utilized to balance the class distribution and increase the accuracy of FND models [273]. These approaches have yielded promising results in addressing the class imbalance problem in FND.

Keya et al. [162] created 'AugFake-BERT' to control imbalances by data augmentation to boost the effectiveness of fake news categorization, specifically addressing the influence of imbalanced datasets on detection models. Similarly, [282] used back-translation as data augmentation, applying pre-trained word embeddings (Word2Vec, GloVe, and fastText) in CNN, BiLSTM, and ResNet models for FND.

Mouratidis et al. [220] addressed the class imbalance issue by implementing SMOTE to oversample the minority class. SMOTE is a widely used method for generating synthetic samples to effectively reduce the class imbalance problem in machine learning models. This approach resulted in an improved accuracy and F1-score, improving from 95% and 99% to 98% and 100%, respectively.

Additionally, other methods such as the focal loss function and specific learning techniques have demonstrated the ability to achieve high accuracy and satisfactory recall, further mitigating the effects of class imbalance [15, 37].

The number of research studies conducted highlights ongoing efforts to tackle the imbalance in detecting fake news.

5.3 BERTGuard: Two-Tiered Multi-Domain Fake News Detection with Class Imbalance Mitigation

In this section, we outline our BERTGuard approach FND, which utilizes a two-stage detection method based on BERT, and explore the impact of domain-specific classification. The first stage of our solution entails classifying the news domains with BERT to capture the nuances associated with various sources of information. Then, in the second stage, we

use domain-specific BERT models to determine the validity of news within their respective areas. Figure 5.1 presents the BERTGuard detection approach.

We chose to base our approach on BERT due to its proven ability to enhance FND by capturing complex linguistic patterns and contextual nuances. Numerous studies have validated the effectiveness of BERT in this domain [25, 129, 171, 238, 254, 316]. In addition, BERT's contextualized word representations enable it to capture the nuanced meaning and context of words and phrases within news articles, which is crucial for distinguishing between real and fake news. Moreover, BERT handles the issue of missing information in FND through its advanced language understanding capabilities and the ability to capture contextual information. By analyzing the language used in the news articles and comparing it to a database of known fake news, BERT can identify patterns and inconsistencies that suggest that a news article may be fake, even in the presence of missing information [146]. Furthermore, BERT's fine-tuning capability allows it to adapt to the nuances of the task and the dataset, which can help mitigate the impact of missing information on the overall accuracy of FND [146].

As depicted in Figure 5.1, the initial stage of the pipeline employs a multiclass classification model based on BERT. This model ingests diverse text formats, such as news articles, and analyzes them to determine and assign a relevant domain.

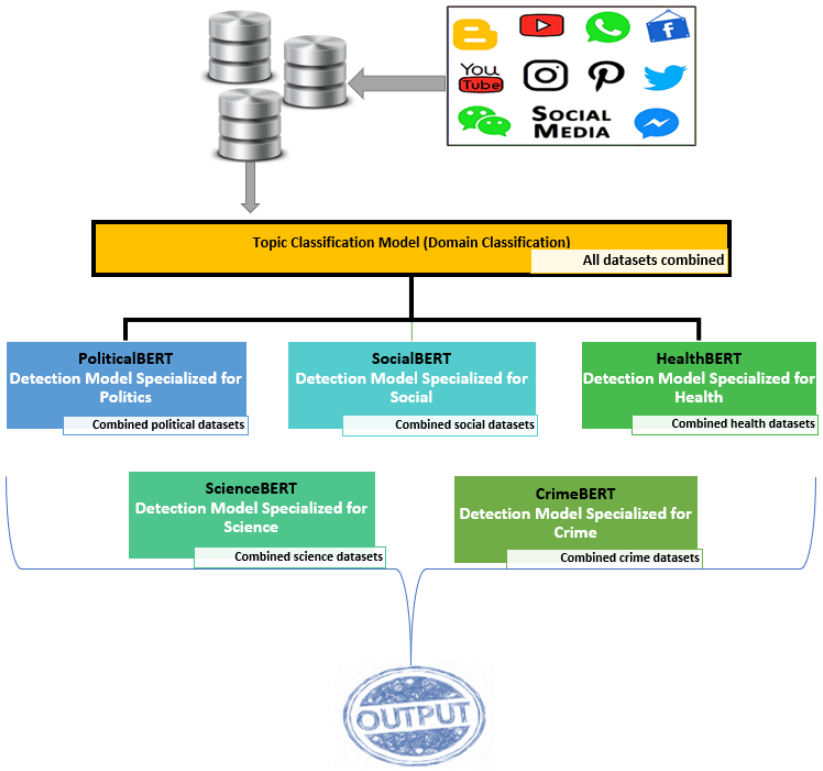

Figure 5.1: BERTGuard FND architecture.

Following initial testing, we discovered that the base version of DistilBERT, a distilled variant of BERT, outperformed other BERT models and was significantly faster in both the training and testing phases. Therefore, it was decided to use DistilBERT for classification.

We trained this model on all the training datasets allocated for this purpose. These training datasets were chosen from five different news domains. We merged the datasets FA-KES, COVID-FN, COVID-FNIR, FakeNews, ISOT, LIAR, Climate, and GossipCop into a single dataset for training DistilBERT. The data were preprocessed for classification by encoding the categories into integer values (crime: 0, health: 1, politics: 2, science: 3, and social media: 4).

The DistilBERT model was therefore set up to have five labels, one for each category. The training, validation, and testing datasets were encoded with the DistilBERT tokenizer and converted into PyTorch tensors. The motivation for domain classification arises from the fact that a news article often belongs to multiple domains or topics, i.e., news articles

may overlap across different topics. In addition, the domain classification model was tuned by adjusting the number of epochs, and several iterations of a given epoch were performed to assess the performance at a given epoch.

Following the domain classification stage, the pipeline directs the input data to a domain-specific detection model. We conduct an initial testing phase to determine the most suitable BERT for each domain. The models in this stage are trained on datasets with data belonging to that specific domain, making them specialized at detecting fake content within that domain. The model outputs a boolean value denoting whether the input is detected as fake or not. Fine-tuning of the models was done to maximize the F1-scores. The fine-tuning that was done on all of the models above for each category includes epochs, optimizer, and learning rate.

Five domain-specific models were trained on each of the five domains considered (crime, health, politics, science, and social media). The pre-trained model selection consists of the tokenizer, the sequence classifier, and the model itself. Below are all the options used:

1. Tokenizer: BertTokenizer, classifier: BertForSequenceClassification, and model: **BERT (base uncased)**;
2. Tokenizer: BertTokenizer, classifier: BertForSequenceClassification, and model: **BERT (base cased)**;
3. Tokenizer: DistilBertTokenizer, classifier: DistilBertForSequenceClassification, and model: **DistilBERT (base uncased, finetuned SST-2 English)**;
4. Tokenizer: DistilBertTokenizer, classifier: DistilBertForSequenceClassification, and model: **DistilBERT (base uncased)**;
5. Tokenizer: RobertaTokenizer, classifier: RobertaForSequenceClassification, and model: **RoBERTa (base)**;
6. Tokenizer: AlbertTokenizer, classifier: AlbertForSequenceClassification, and model: **ALBERT (base v2)**.

Overall, BERTGuard incorporates an initial BERT model for domain classification, followed by five specialized BERT-based models tailored to the detected domain.

In addition, BERTGuard addresses the critical issue of class imbalance in FND. Our approach includes balancing the training datasets by oversampling the minority class, undersampling the majority class, and adjusting the class weights.

5.4 Research Methodology

In our study, we developed BERTGuard: a two-stage FND approach utilizing BERT and focusing on domain-specific classification. The first stage involves classifying the news domains with BERT to capture the nuances of various sources of information. In the second stage, domain-specific BERT models are used to assess the validity of news within their respective domains. To establish baselines for comparison, we developed a model that disregards news domains to serve as the first baseline. We also used a well-known model using a base version of BERT called FakeBERT [146] as another baseline.

5.4.1 Selected Datasets

We chose the most widely used and publicly available datasets related to the following fake news domains shown in Table 5.2. This table provides an overview of the datasets used in our experiments, organized by news domain to highlight the diversity of the data. Each dataset is categorized into specific domains, such as politics, health, and social media, which helps understand the context and type of content each dataset contains. The table also indicates which datasets were used for training and which were used for testing. This distinction is crucial for evaluating the model's performance across different domains and ensuring that the model is exposed to a variety of data during both the training and testing phases. The primary goal of this organization is to provide comprehensive coverage of each domain. By having at least one dataset from each domain for training and another for testing, we can better assess the model's generalization capabilities and its effectiveness in handling domain-specific nuances. This structured approach allows us to systematically evaluate ProFineLlama's performance and robustness across multiple domains, providing a thorough assessment of its fake news detection capabilities.

It is worth mentioning that the datasets used previously in Chapter 4 are different from the ones we use in this chapter, Chapter 5, and in the next chapter, Chapter 6. The datasets used in Chapter 4 were selected to provide a broad evaluation of different architectures, including LSTM, BiLSTM, RNN, CNN, and BERT, allowing for an initial comparison of their effectiveness in FND. These experiments helped draw the conclusion that a generalized cross-domain detection model was needed, which required datasets covering various news domains to improve model adaptability. While Chapters 5 and 6 focus on advanced approaches, they incorporate most of the datasets used in Chapter 4 along with additional datasets from other domains to better address cross-domain generalization and improve the robustness of the models. The dataset selection evolved to align with the

research objectives in each stage, ensuring that the later experiments were conducted on a more diverse and representative set of news domains.

Table 5.2: Characteristics of the datasets used in our experiments.

Dataset	**Ref**	**Domain**	**True**	**Fake**	**Purpose**
FA-KES	[4]	Crime	426	378	Training
Snopes	[303]	Crime	195	120	Testing
COVID-FNIR	[270]	Health	3795	3793	Training
COVID-FN	[312]	Health	2061	1058	Training
COVID-claims	[270]	Health	1591	1230	Testing
Liar	[345]	Politics	7176	5669	Training
ISOT	[11]	Politics	23,481	21,417	Training
FakeNews	[144]	Politics	10,413	10,387	Training
Pheme	[377]	Politics	5089	1335	Testing
Politifact	[216]	Politics	11,760	9392	Testing
Climate	[54]	Science	654	253	Training
ISOT-small (ISOT-Science)	[11]	Science	1000	1000	Testing
GossipCop	[300]	Social	16,817	5323	Training
ISOT-small (ISOT-Social)	[11]	Social	1000	1000	Testing

Regarding the training data and the preprocessing tasks, the model in the domain classification stage (the first stage of BERTGuard) is trained on a dataset composed of FA-KES, COVID-FN, COVID-FNIR, FakeNews, ISOT, LIAR, Climate, and GossipCop datasets. These datasets represent the five news domains we selected. The data are processed into 'text' and 'label' columns. The 'text' column contains all the news article content (title and text) used for detection. The 'label' column will be either '1' or '0', corresponding to 'fake' or 'not fake', which serves as the target for the detection model.

Combining these datasets to train the domain classification model required careful consideration as some datasets include more textual columns such as 'title' and 'text'. We concatenated these fields into a single 'text' column containing all relevant content from the news articles. All other features were stored in a JSON dictionary and kept as supplementary metadata. Retaining the additional features in a metadata field can make them ready to be accessed if needed for future refinements or other aspects of the project.

Our datasets (news articles) were already labeled by their respective sources. We did not perform additional labeling. Each article was assigned a label in the 'label' column, where '1' indicates 'fake' and '0' indicates 'not fake'. Some datasets had more than two

possible truth values, so articles that were labeled half-true or mostly-true were labeled 0 (True), while articles labeled barely-true or pants-fire were labeled 1 (Fake). This label was used as the target variable for our detection models.

For domain labeling and categorization, we added a new 'domain' column to each dataset, indicating the specific category (domain) of each article based on its content and source. The domain label is consistent across all news articles within the same dataset. The domains were categorized into five primary areas: Politics, health, crime, social and science. These categories were encoded into integer values for consistency and ease of processing as follows: Crime: 0, Health: 1, Politics: 2, Science: 3, and Social: 4. We followed a systematic approach, splitting the data into training, validation, and testing sets with a ratio of 7:1:2. We also included essential preprocessing steps and tracked the duration of the training to maintain consistency and optimize effectiveness. We tuned the models by adjusting the number of epochs and running multiple training iterations at a given epoch.

To achieve a unified format for all the collected datasets, a comprehensive preprocessing was performed. This standardization ensures minimal additional processing in later stages and maintains data consistency across the project. The following steps detail the main preprocessing tasks:

1. Label standardization: The label values were standardized to ensure uniformity. The datasets that we utilized in this study were already labeled by their sources. In this project, '1' was assigned to indicate 'fake' news, and '0' indicates 'not fake' news. Some datasets originally had multiple classes representing varying degrees of 'fakeness'. These were binarized into a fake / non-fake format based on criteria that balanced the dataset as effectively as possible.

2. Text processing:

    - Cleaning: Raw data were cleaned to remove irrelevant content such as HTML tags, scripts, special characters, and advertisements.
    - Standardization: Standardized formats were applied, such as converting dates to a consistent format and removing duplicates.
    - Normalization: text normalization included converting text to lowercase and expanding contractions.
    - Tokenization: The text was tokenized to break it down into individual words or tokens using BERT's WordPiece tokenizer.

3. Handling missing data: entries with missing titles, bodies, or descriptions were removed to ensure the integrity of the dataset.

4. Multi-language data: for datasets containing multi-language data, articles were filtered to include only those written in English, ensuring uniformity in text processing and model training.

5. Metadata management: all other features not directly used in the primary analysis were stored in a JSON dictionary as supplementary metadata. This approach allows these features to be easily utilized if needed for further refinement of the results.

It is worth mentioning that no other text preprocessing techniques, such as stopwords removal and lemmatization, were applied so as not to affect BERT performance.

Regarding the detection stage (second stage of BERTGuard), each domain-specific model was trained using the training datasets that represent that domain. For example, we used the FakeNews, ISOT, and LIAR datasets to train the political BERT model. The remaining political datasets (Pheme and Politifact) were used later for final complete detection system testing.

Data were handled by loading them as a whole category to train and test models. This process included detecting all datasets files, loading them as dataframes, and then concatenating all of these dataframes to create a combined dataframe.

5.4.2 Experimental Settings

Various versions of BERT were utilized after carefully considering the performance and time required by these models. The selection of these BERT versions has been made based on the experimental results that we will detail later in Table 5.4 in Section 5.5.1. This dynamic selection process improves the adaptability and effectiveness of our solution. Based on these experimental results, we chose the following BERT versions for the following news domains:

- Crime: distilbert-base-uncased-finetuned-sst-2-english;
- Health: distilbert-base-uncased;
- Politics: bert-base-uncased;
- Science: roberta-base;

- Social media: distilbert-base-uncased.

For the first stage of our BERTGuard, we trained the DistilBERT model on a merged dataset comprising the FA-KES, COVID-FN, COVID-FNIR, FakeNews, LIAR, Climate, and GossipCop datasets. The dataset was preprocessed by encoding categories into integer values (crime: 0, health: 1, politics: 2, science: 3, and social media: 4). The DistilBERT model was configured with five labels: one for each category. In the second stage of BERTGuard, each domain model was trained on the training datasets from the respective domain. For evaluation, we compared the performance of BERTGuard against the baselines.

Our BERTGuard also attempts to mitigate the class imbalance since this is a critical issue in the FND context. These attempts include balancing the training datasets by oversampling the minority class, undersampling the majority class, and adjusting the class weights.

Oversampling entails boosting the number of instances in the minority class by randomly replicating them. This technique helps mitigate bias towards the majority class, enhancing the model's ability to learn from minority class examples.

Conversely, undersampling involves decreasing the number of instances in the majority class by randomly discarding some samples. This approach prevents the model from being dominated by the majority class, allowing it to concentrate more on learning from the minority class.

In addition, adjusting class weights is another technique used to address class imbalance. The idea is to assign higher weights to the minority class (e.g., fake news) and lower weights to the majority class (e.g., true news) during training to ensure that the model pays more attention to the minority class. The common approach to determining the optimal class weights is manual tuning. The purpose is to manually tune the class weights based on domain knowledge or prior experience with the dataset. The choice of class weights is often a balance between overfitting and underfitting. Assigning excessively high weights to the minority class can lead to overfitting, where the model becomes overly biased towards classifying instances as the minority class, potentially reducing its overall accuracy and effectiveness. Conversely, setting the weights too low can result in underfitting, where the model does not adequately learn from the minority class, leading to poor performance in those examples. This approach ensures that the loss function penalizes misclassification of underrepresented classes more heavily, thereby improving their recognition. The weights were fine-tuned iteratively, evaluating the model's performance on validation data to achieve an optimal balance between majority and minority class predictions. By dynamically adjusting these weights, the model was able to learn from the minority class

without overfitting to noise, improving classification robustness in the presence of skewed distributions.

Regarding the main baseline model, we developed a BERT (base uncased) model for one-stage FND (Baseline 1) without paying attention to the news domains. The implementation is based on what we did in the domain classification model, with the following changes: the label column will be either '1' or '0', corresponding to 'fake' or 'not fake', which will be the target for the detection model.

The other baseline (Baseline 2) that we used from the literature, FakeBERT [146], was trained on the same dataset used to train our main baseline.

5.4.3 Evaluation Metrics

The performance of the proposed models was evaluated using various metrics, as detailed in Chapter 4, Section 4.3.6. These metrics were selected to offer a complete evaluation of the effectiveness of the detection model. These metrics include recall, precision, accuracy, F1-score, and G-mean.

5.4.4 Computational Environment

The experiments in this chapter were carried out in high-performance local and cloud-based computing environments to ensure efficient model training, evaluation, and data processing. The primary setup utilized for these experiments included a ROG Strix G733ZW laptop equipped with a 12th Gen Intel Core i9-12900H processor, NVIDIA GeForce RTX 3070 Ti GPU, and 32GB DDR5 RAM. Additionally, cloud-based computing resources were employed, specifically Vast.ai, which provided access to an AMD EPYC 7302 16-Core processor, an NVIDIA RTX 4090 GPU, and 32GB RAM, offering peak performance of 81.8 TFLOPS.

The experiments were executed using CUDA 12.3 for GPU acceleration, with Python, Jupyter Notebooks, and VSCode as the primary development environments. The software stack included NumPy, transformers, pandas, torch, tqdm, scikit-learn, pyarrow, accelerate, sentencepiece, ipywidgets, datetime, logging, imblearn, papermill, and XlsxWriter, ensuring seamless handling of data preprocessing, model training, and evaluation.

The setups used in this chapter provided greater flexibility for iterative experimentation and rapid prototyping, balancing computational efficiency with accessibility. This approach enabled extensive testing while maintaining high computational performance, ensuring reliable and scalable experimentation across multiple FND tasks.

5.5 Results and Discussion

This section presents the results of our experimental evaluation. We begin by presenting the ad hoc testing results for selecting the best model for each case and domain. Afterward, we observe the performance of the baseline models considered and BERTGuard. Then we explore the effect of handling the imbalance problem by adjusting the class weights and applying upsampling and downsampling.

5.5.1 Model Selection

In this section, we present the initial ad hoc tests to pick the model for classification and detection. This is not meant to be an exhaustive test, but rather a quick comparison between the models when deciding the best course of action.

To select the domain classification model, we used a balanced dataset with 850 samples per category in the training/validation data and 160 samples/category in the testing data. The test/validation split was 80% to 20%. Table 5.3 shows the accuracy for each model used.

Table 5.3: Domain classification model results.

Model	**Average Accuracy**	**Training Time** (seconds)
albert-base-v2	95.8%	724.1
bert-base-uncased	92.9%	1933.9
distilbert-base-uncased	96.8%	96.2
distilbert-base-uncased-finetuned-sst-2-english	91.8%	220.6
roberta-base	98.6%	1546.0

While roberta-base achieved a slightly higher average accuracy (98.6%) compared to distilbert-base-uncased (96.8%), the training time for distilbert-base-uncased was significantly shorter. The distilbert-base-uncased model trained in 96.21 s, whereas the roberta-base model required 1545.98 s.

This substantial difference in training time demonstrates that distilbert-base-uncased is much more efficient, making it a more practical choice, especially when computational resources and time are constrained. Moreover, the marginal difference in accuracy (1.8%)

does not justify the significantly higher training time and resource consumption of the roberta-base model. Therefore, we opted for the distilbert-base-uncased model to achieve a balance between high accuracy and efficient resource utilization.

For the domain-specialized detection model, another quick testing task took place for each news domain. Table 5.4 presents the average accuracies of different models tested for each domain.

Table 5.4: The average accuracies of various models on news domains.

Model	Crime	Health	Politics	Science	Social
albert-base-v2	61.7%	84.8%	67.8%	75.3%	84.2%
bert-base-uncased	61.3%	87.5%	**87.8%**	74.7%	84.7%
distilbert-base-uncased	60.1%	**88.3%**	68.5%	77.7%	**95.6%**
distilbert-base-uncased-finetuned-sst-2-english	**68.8%**	86.2%	61.3%	71.6%	84.5%
roberta-base	56.9%	87.6%	71.8%	**80.2%**	88.7%
bert-base-cased	61.9%	87.1%	69.6%	76.4%	84.7%

For deciding the best model for each news domain, we prioritized the highest average accuracy for each specific domain:

1. Crime: The distilbert-base-uncased-finetuned-sst-2-english model performed the best, with an average accuracy of 68.8%. This suggests that the fine-tuned version of DistilBERT is better at capturing the nuances in the crime domain.

2. Health: The distilbert-base-uncased model achieved the highest accuracy of 88.3%, making it the most suitable for health-related news detection.

3. Politics: The bert-base-uncased model had the highest accuracy at 87.8%, indicating its effectiveness in detecting political news.

4. Science: The roberta-base model excelled in the science domain, with an accuracy of 80.2%, suggesting it handles the specific language and context of science-related news more effectively.

5. Social: The distilbert-base-uncased model stood out, with a remarkable accuracy of 95.6%, making it the best choice for the social domain.

These selections balance accuracy across various domains, ensuring that each model is optimized for the specific characteristics of the content it will encounter. Although the Roberta-base model showed strong performance in certain areas, the distilbert-base-uncased variants, including the fine-tuned version, offered competitive accuracies with the added benefit of being more resource-efficient, as noted in previous experiments.

5.5.2 Single-Stage Fake News Detection Results

In this section, we present the results of testing the first baseline model (Baseline 1) that we used in our work. This model is a BERT(base uncased) model trained on the merged datasets from the domains considered and tested on the set of unseen datasets. Domain categories were not taken into account when training this model. Table 5.5 presents the precision, recall, accuracy, F1-score, and G-Mean when testing the Baseline 1 model, which is the main baseline model we developed.

Table 5.5: Single-stage FND results—Baseline 1.

Testing Dataset	Precision	Recall	F1-Score	Accuracy	G-Mean
Snope	**0.583**	0.774	0.665	0.517	0.672
COVID-Claims	0.224	0.135	0.169	0.247	0.174
Pheme	0.313	0.193	0.238	**0.744**	0.245
Politifact	0.486	0.182	0.265	0.495	0.298
Climate	0.349	0.435	0.387	0.616	0.390
GossipCop	0.379	0.744	0.502	0.645	0.531
ISOT-small	0.581	**0.790**	**0.670**	0.557	**0.677**

The evaluation of Baseline 1 reveals the drawbacks of a non-domain-specific approach. Across several datasets from various domains, distinct patterns emerge. Starting with the accuracy and recall measures, the model's performance varies significantly between datasets. Notably, in the 'Snope' dataset, the model shows higher precision but lower recall, demonstrating a tendency to properly identify true news while missing a significant proportion of fake news cases. In contrast, in the 'GossipCop' dataset, the model achieves a greater recall, indicating a stronger capacity to detect fake news but with a trade-off in precision. On the other hand, specialized datasets such as 'COVID-Claims' pose distinct difficulties for the baseline model, as evident in the reduced precision, recall, and F1-score. This underscores the importance of tailoring models to specific content domains, particularly those with distinct characteristics such as health-related information. Additionally,

the G-Mean, which measures the balance between recall and specificity, is notably low for this dataset, further highlighting the challenge in achieving a balanced performance in certain domains.

The baseline model has modest precision and recall in the 'Politifact' dataset, indicating some adaptation across platforms. However, the findings highlight the need for a more complex and domain-aware strategy to detect fake news across several platforms. However, an unanticipated anomaly appears in the 'Pheme' dataset, where the model achieves significantly higher precision and accuracy than other performance measures. This anomaly necessitates more analyses to identify potential biases or unique dataset factors that could have influenced the model's performance. Understanding such anomalies is critical for improving the model's adaptability to different datasets.

The discovered disparities highlight the limitations of creating a one-size-fits-all FND methodology. While the baseline approach shows promise in some domains, its performance limits emphasize the need for domain-specific modifications.

5.5.3 Multi-Domain Fake News Approach

In this section, we present the results of testing our BERTGuard, a multi-tier detection system, by using unseen datasets that we kept for this purpose. We compared Baseline 1, the existing Baseline 2 solution (FakeBERT), and our proposed BERTGuard.

Table 5.6 shows the overall results of these three solutions averaged over all the testing datasets.

Table 5.6: Overall results of Baseline 1, Baseline 2 (FakeBERT), and BERTGuard.

Architecture	**Precision**	**Recall**	**F1-Score**	**Accuracy**	**G-Mean**
Baseline 1 (No Domains)	42%	46%	41%	55%	43%
Baseline 2 (FakeBERT)	32%	34%	31%	38%	32%
BERTGuard (Domain-Aware)	**82%**	**68%**	**70%**	**82%**	**72%**

Note: These results reflect the average testing values for various unseen datasets.

The results demonstrate that our domain-specific strategy significantly improves detection performance by leveraging domain-specific information. Moving into more detailed results of our experiments, Table 5.7 presents the precision, recall, accuracy, F1-score, and G-mean when testing the proposed BERTGuard for each testing dataset.

Table 5.7: The proposed BERTGuard testing results.

Testing Dataset	Precision	Recall	F1-Score	Accuracy	G-Mean
Snopes	0.838	**0.979**	0.903	0.870	0.906
COVID-Claims	**0.957**	0.976	**0.966**	**0.962**	**0.966**
Pheme	0.848	0.625	0.719	0.899	0.728
Politifact	0.800	0.083	0.151	0.479	0.258
Climate	0.535	0.577	0.555	0.742	0.556
GossipCop	0.836	0.663	0.739	0.888	0.744
ISOT-small	0.899	0.868	0.883	0.884	0.883

BERTGuard had significant improvements across multiple datasets that reflect multiple news domains, as evidenced by the precision, recall, F1-score, accuracy, and G-Mean metrics.

The findings from the 'Snopes' and 'COVID-Claims' datasets demonstrate the success of our domain-specific approach in these crucial domains. With precision reaching 83% and recall exceeding 97% in 'Snopes' and precision and recall both exceeding 95% in 'COVID-Claims', the model exhibits an impressive ability to properly classify both true and fake news instances, demonstrating its durability in areas where accuracy is critical. Moreover, the G-Mean values of 90.58% for 'Snopes' and 96.64% for 'COVID-Claims' reflect a well-balanced performance, indicating the model's effectiveness in detecting both true and fake news without favoring one over the other.

In the 'Pheme' dataset, the model performs superbly, with a precision of 84.76%, suggesting a high proportion of accurately identified fake news events. However, the recall is rather low at 62.47%, indicating possible difficulties in collecting all instances of fake news in this particular domain. This intricate balance requires further investigation to fully comprehend the complexities of information transmission within the 'Pheme' dataset. The G-Mean of 72.77% for 'Pheme' shows that while precision is strong, the recall limitation affects the overall balanced performance. Class imbalance could be a significant factor here, as the model might be biased towards the majority class (true news), leading to poor recall for the minority class (fake news). To improve this, techniques such as re-sampling the dataset could help address the imbalance.

The model encounters significant obstacles in the 'Politifact' and 'Climate' datasets. 'Politifact' has a lower precision of 80.05% and a recall of 8.32%, indicating a risk of false positives. The 'Climate' dataset, despite obtaining moderate precision and recall, reveals complexity in differentiating between authentic and fake news in the science-related

domain. The G-Mean values of 25.81% for Politifact' and 55.55% for 'Climate' further emphasize the challenges in achieving a balanced results in these domains. The very low G-Mean for 'Politifact' is particularly due to the extremely low recall (8.32%), which suggests that the model misses a large proportion of fake news, resulting in a significant imbalance between recall and precision. This low recall could be due to several factors, including class imbalance, where the fake news instances might be underrepresented in the dataset making it difficult for the model to identify fake news. To address this, re-sampling techniques could help balance the dataset. In contrast, the model performs well on the 'GossipCop' and 'ISOT-small' datasets, with excellent precision and recall values. In 'ISOT-small', the model achieves precision and recall levels that exceed 89%, demonstrating its adaptability and effectiveness in dealing with a wide range of information across domains.

Across all datasets, our BERTGuard maintains a balanced F1-score, demonstrating its capacity to perform consistently across domains. Taking domains into consideration improved the detection performance compared to the baseline strategies, which did not consider news domains in their detection approach. Figure 5.2 shows the F1-score comparison between both baselines and the proposed BERTGuard solution with domain-specific knowledge.

The use of different models for each news domain highlights the challenge of cross-domain generalization in FND. This suggests that a single model cannot generalize equally well across all domains, making domain-aware fine-tuning a necessary step to maximize performance. Initial experiments showed that when tested on a dataset from an unseen domain, BERTGuard successfully selected the most relevant trained domain, demonstrating its adaptability. However, if new data from a completely new domain becomes available, an ablation study could be conducted to evaluate its impact, followed by updating the classifier to incorporate the new domain. Additionally, even within the same topic, variations in data distribution could lead to different models being more effective, reinforcing the need for continuous domain adaptation and model selection based on evolving data patterns.

Finally, the domain-specific technique appears to be a promising step toward a system that can detect fake news with greater adaptability and accuracy. The nuanced performance across domains emphasizes the significance of designing models to fit the complexities of unique information sets and domains.

It is worth mentioning that statistical significance tests were not conducted because the experimental results across different models and domains were closely aligned across different runs, reducing the necessity for such tests. Given the similarity in performance metrics, statistical testing would likely not yield meaningful distinctions. Instead, the evaluation focused on accuracy, F1-score, and domain-specific performance trends, which

provided sufficient insights into model effectiveness.

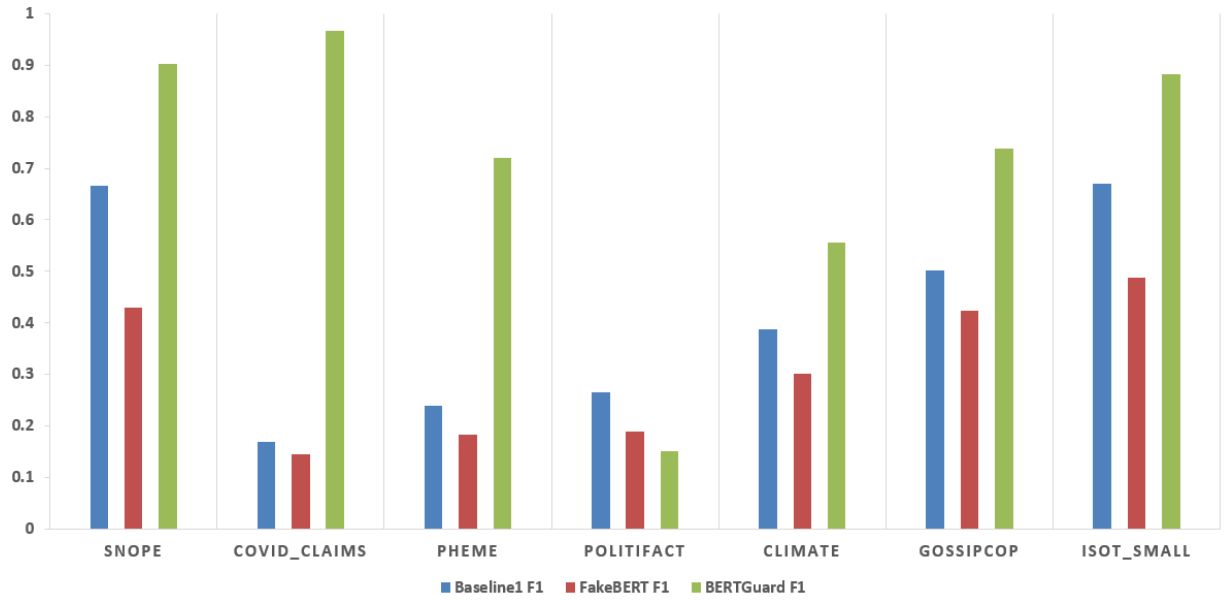

Figure 5.2: F1-scores for both baselines against BERTGuard.

5.5.4 Overall Impact of Class Imbalance Handling Strategies

This section details the findings from the experiments aimed at assessing the effectiveness of various strategies for addressing class imbalance, including random upsampling, random downsampling, and adjustments to class weights. We tested this impact on Baseline 1, FakeBERT (Baseline 2), and our proposed BERTGuard. We trained each model on balanced datasets by applying each imbalance strategy individually. The performance obtained was compared against the model's performance without handling the class imbalance. The results shown in Table 5.8 show the results of this experiment averaged across all datasets tested. We observe that modifying class weights consistently outperforms the other strategies, indicating its effectiveness in mitigating the class imbalance in the context of FND.

Table 5.8: The impact of class imbalance handling in BERTGuard.

Strategy	Precision	Recall	F1-score	Accuracy
BERTGuard				
No balancing	82%	68%	70%	82%
Random oversampling	81%	69%	71%	82%
Random undersampling	79%	61%	65%	68%
Adjusting class weights	**87%**	**68%**	**73%**	**78%**
Baseline 1				
No balancing	42%	46%	41%	55%
Random oversampling	46%	48%	45%	52%
Random undersampling	36%	41%	37%	46%
Adjusting class weights	**62%**	**51%**	**55%**	**56%**
FakeBERT (Baseline 2)				
No balancing	32%	34%	31%	38%
Random oversampling	46%	48%	45%	52%
Random undersampling	36%	41%	37%	46%
Adjusting class weights	**62%**	**51%**	**55%**	**56%**

Note: These results reflect the average testing values for testing using various unseen datasets to include Snope, COVID-Claims, Pheme, Politifact, Climate, GossipCop, and ISOT-Science.

These findings underscore the importance of treating class imbalances in FND, with class weighting being the most effective strategy among those tested, while downsampling did not consistently improve the model's performance across all testing datasets.

5.5.5 Impact of Balancing Datasets by Adjusting Weights

The effort to address class imbalance through adjusting weights in the dataset has yielded distinctive outcomes for the baselines and the proposed domain-specific FND approaches. Figures 5.3 and 5.4 show the effect of handling the class imbalance issue by adjusting the class weights on both the main baseline (Baseline 1) we developed and the BERTGuard approaches.

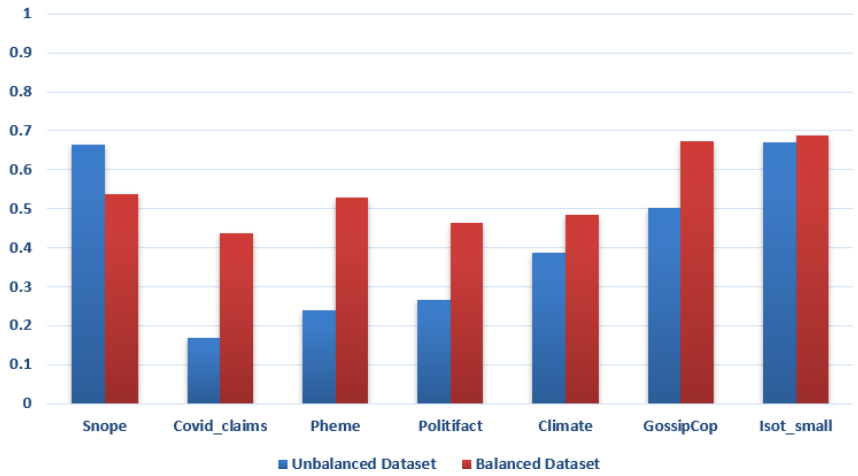

Figure 5.3: The effect of handling the imbalance on the main baseline (Baseline 1) F1-scores by adjusting the class weights.

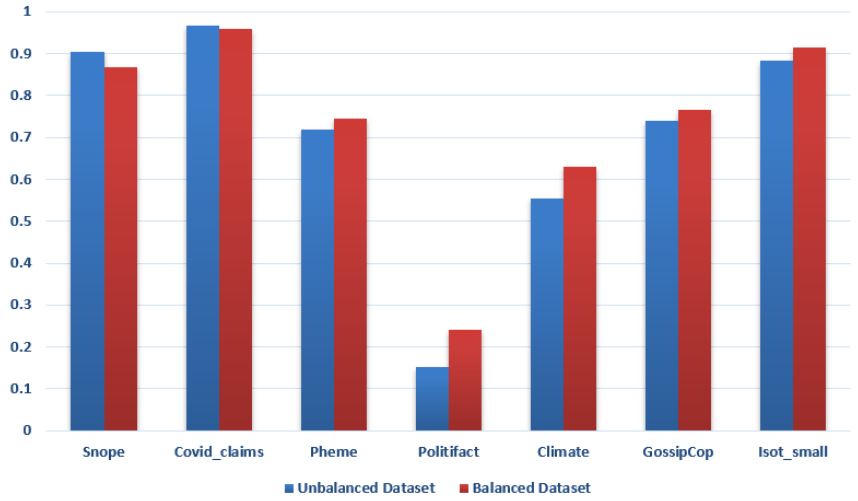

Figure 5.4: The effect of handling the imbalance on the BERTGuard F1-scores by adjusting the class weights.

The class weight adjustment approach results in significant improvements in some domains. Notably, 'Snopes' and 'COVID-Claims' show high precision and recall rates, demonstrating the model's expertise in these critical domains. The targeted changes in class weights show the potential efficacy of domain-specific techniques for resolving class

imbalance. Significant advances in 'Pheme' and 'GossipCop' were demonstrated by gaining higher precisions, recalls, and F1-scores. The model's adaptability in capturing complex information transmission within these domains demonstrates the advantages of altering the class weight in the detection task as well as the possibility for domain-specific techniques to efficiently handle class imbalance over the baseline approach with no domains.

Despite efforts to address the imbalance, difficulties exist in some domains, as seen in the 'Politifact' and 'Climate' datasets. The models struggle to achieve a harmonious balance between precision and recall in several domains, implying that domain-specific complications may necessitate more specialized tactics than simple class weight modifications. Figure 5.5 presents a comparison between the baselines and the BERTGuard F1-scores after handling the class imbalance by adjusting the class weights.

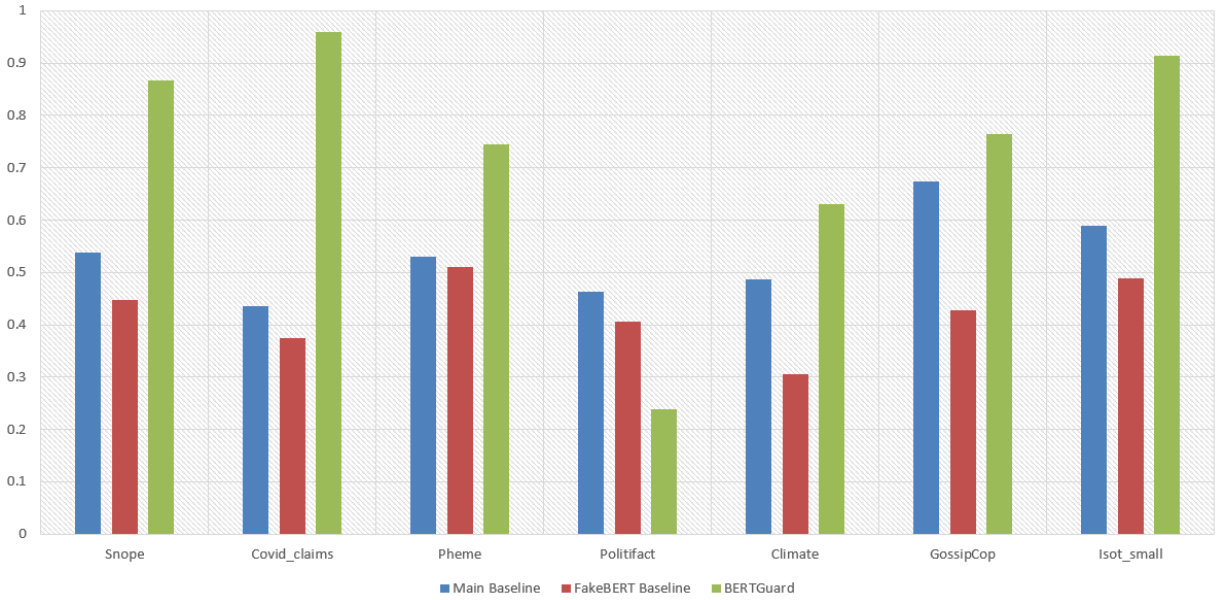

Figure 5.5: F1-scores of both baselines and BERTGuard after treating the imbalance by adjusting class weights.

In conclusion, the findings highlight the approach's versatility across a variety of datasets while also noting ongoing problems in specific domains. Balancing datasets by altering class weights is a complicated process. The domain-specific methodology demonstrates promising advances in the majority of the domains, emphasizing the importance of continuing refinement and adaptation. The findings also highlight the dynamic nature of fake news

identification, emphasizing the need for specific imbalance handling solutions to navigate the intricacies of various domains.

5.5.6 Impact of Balancing Datasets by Upsampling

In this section, we investigate handling class imbalance by using the random oversampling technique. Although this technique shows enhancements over unbalanced cases, upsampling performed more poorly compared to adjusting the class weights. Figure 5.6 shows the effect of the random oversampling technique on our proposed BERTGuard approach.

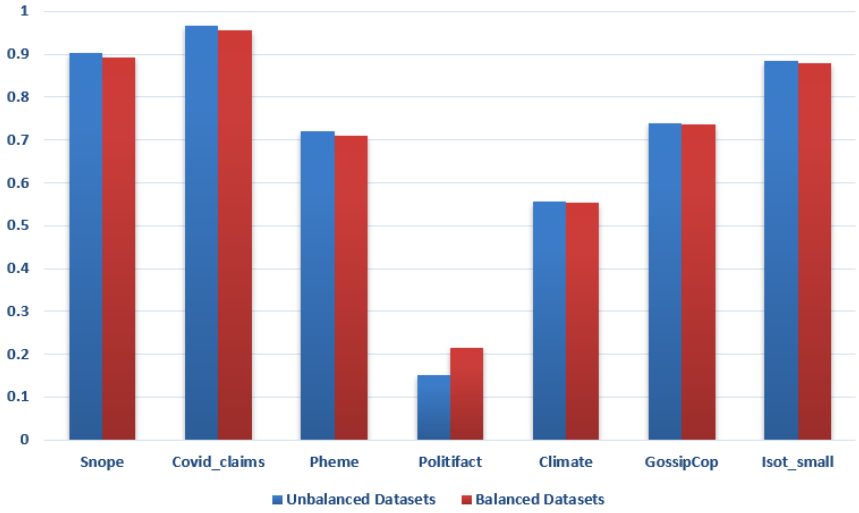

Figure 5.6: The effect of random oversampling on BERTGuard F1-scores.

The F1-scores improved notably across multiple datasets as compared to the baseline. The F1-scores topped 89% in 'Snope' and 'COVID-Claims', showing a significant improvement in both precision and recall. The approach achieved balanced precision and recall in both 'Gossipcop' and 'ISOT-small', indicating an effective trade-off between false positives and false negatives in both the baseline and domain-specific approaches.

While some datasets' F1-scores improved when compared to the imbalanced baseline, others declined. Notably, 'COVID-Claims' and 'Pheme' had lower F1-scores for the baseline approach, showing the limitations of the baseline model even in a balanced setting. In the domain-specific approach, several datasets, such as 'Politifact' and 'Climate', nevertheless faced difficulty, with lower F1-scores. This indicates that upsampling alone might not fully

address the complexities in certain domains, and other techniques may perform better, such as adjusting the class weights.

In conclusion, while both upsampling and class weight adjustment strategies help to improve FND in imbalanced datasets, the class weight adjustment approach displayed superior adaptability and competitive or superior F1-scores across many datasets, as presented in Figure 5.7.

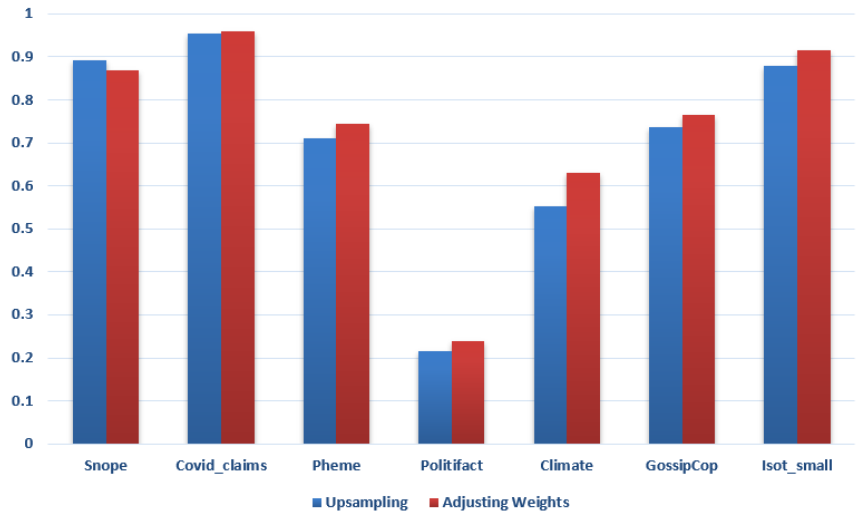

Figure 5.7: F1-score comparison between upsampling and class weight adjustments on BERTGuard.

5.5.7 Impact of Balancing Datasets by Downsampling

In this section, we investigate handling the class imbalance by using the downsampling technique. Although this technique has shown enhancement over the unbalanced case in different classification tasks, it did not help in our context.

Balancing the class distribution in the used dataset using random undersampling had a noticeable effect on the performance of our FND approach. Looking at the F1-scores before and after treating the class imbalance using random undersampling, we can see that for most of the datasets, the F1-scores decreased slightly with training the models on balanced datasets by randomly undersampling the majority class. This could be due to the loss of information that occurs when randomly removing instances from the majority class during undersampling. However, it is important to note that while the F1-scores decreased, they

are still relatively high, indicating that our model still performs well even after balancing the dataset. Figure 5.8 presents the F1-score results for the proposed BERTGuard after applying the random undersampling technique for handling class imbalance.

Regarding the baselines, downsampling produced mixed results. While there were advances in 'Pheme' and 'Politifact', other datasets such as 'COVID-Claims' and 'Snope' had difficulties, leading to decreased precision. Furthermore, 'GossipCop' and 'ISOT-small' revealed a trade-off between precision and recall, highlighting the difficulty of attaining balanced performance through downsampling alone. Figure 5.9 presents the F1-scores for the Baseline 1 (single-stage) detection model after applying the downsampling technique to handle the imbalance.

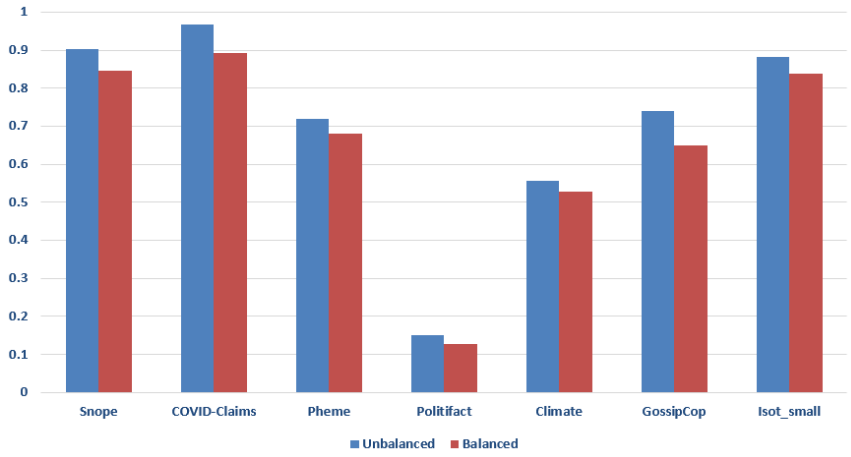

Figure 5.8: The effect of random undersampling on BERTGuard F1-scores.

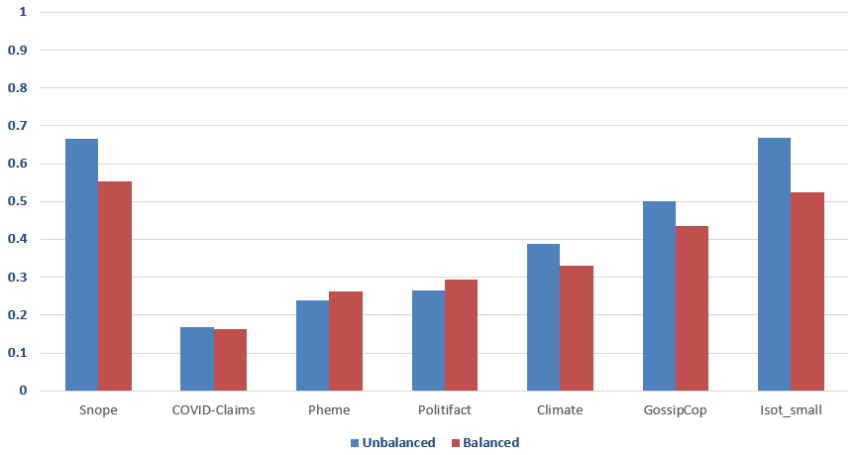

Figure 5.9: The effect of random undersampling on Baseline 1 F1-scores.

The performance of downsampling was context-dependent, with certain datasets favouring this approach over others. This underscores the importance of understanding the intricacies of each dataset to tailor the strategy accordingly. It performed well in some cases compared to the non-balanced case but had worse performance than the other imbalance handling techniques. Figure 5.10 compares the results of the F1 score when utilizing the three class imbalance handling techniques in the proposed domain-specific approach.

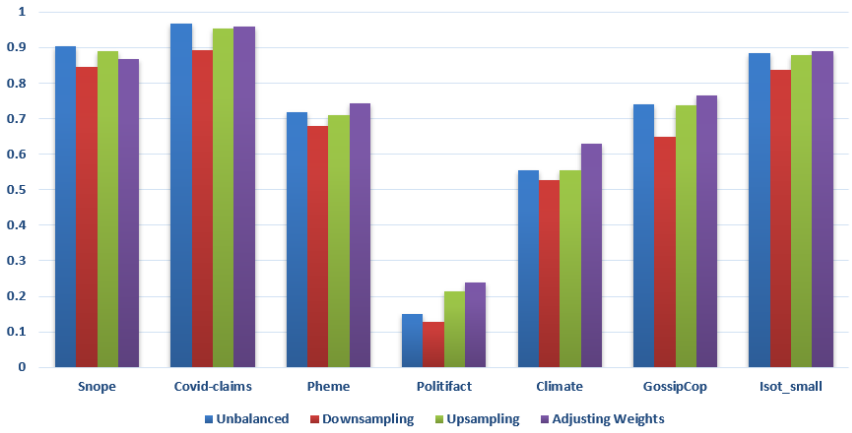

Figure 5.10: The effect of using upsampling, downsampling, and class weight adjustment vs. no handling technique on BERTGuard.

Another observation is that the F1-scores for some datasets, such as Climate, remained

relatively stable after balancing, while others, like Politifact and Climate, saw a more significant decrease. This variation could be due to the nature of the datasets and the distribution of the classes within them.

Downsampling was not effective in improving the precision-recall balance and was not helpful in most other datasets. Upsampling, while beneficial in some domains, has drawbacks, including susceptibility to information nuances, and it handled more datasets than the downsampling technique. Class weight modification emerges as a consistently adaptive method for preserving and improving detection accuracy across diverse datasets and domains. The domain-specific character of issues, as seen in 'Politifact', necessitates nuanced studies and future model changes.

For detailed statistics per each fake news dataset used, Appendix A shows testing results of our proposed two-tier model against both baselines.

5.6 Summary

In this chapter, we propose BERTGuard: a thorough, domain-specific strategy within the multi-domain FND framework. Our methodology builds upon the unique characteristics that exist in various news domains. BERTGuard, a two-tiered FND approach, includes domain categorization and domain-specific analyses using various BERT models. Our proposed approach demonstrates enhanced adaptability and precision across diverse information environments by leveraging multiple datasets from various domains simultaneously. To the best of our knowledge, this setting has not been explored previously. Our chapter introduced a methodical approach to the challenging obstacle of class imbalance in multi-domain FND. We determine the best-performing strategy by rigorously evaluating handling strategies such as random oversampling, random undersampling, and class weight adjustment. This strengthens the detection system against the difficulties of imbalanced datasets.

While our BERTGuard approach demonstrated strong performance in multi-domain FND, challenges related to noise and bias remain and require further attention. Although preprocessing steps were applied to reduce noise, dataset inconsistencies and ambiguities may still affect model performance. The dataset selection process aimed to incorporate diverse sources to improve representation, yet biases in news content and reporting styles remain an inherent difficulty. Additionally, while class imbalance was addressed through weighting adjustments, oversampling, and undersampling techniques to prevent the model from being skewed toward majority classes, such methods may introduce their own biases

or overfitting risks. Furthermore, cross-domain evaluation was performed to assess generalizability and reduce dataset-specific biases. Future work should explore more robust bias mitigation strategies to enhance model fairness and generalizability across domains.

In the next chapter, we will explore an advanced transfer learning approach that further builds upon the findings presented in this chapter, aiming to refine and optimize FND across various domains with tailoring prompts for the navigated news domains.

Chapter 6

ProFineLlama: A Prompt and Fine-Tuned Transfer Learning Approach for Multi-Domain Fake News Detection

In the previous chapter, we explored a two-tiered approach that improved multi-domain FND by addressing class imbalance. Building on that foundation, this chapter introduces a TL-based method that further enhances detection accuracy across diverse datasets. Specifically, we combine fine-tuning with prompt-tuning techniques to leverage pre-trained models in a way that adapts more effectively to the nuances of fake news in different domains. This chapter's work is based on the following article:

- Alnabhan, M.Q., Branco, P. (2024). ProFineLlama: A Prompt and Fine-Tuned TL Approach for Multi-Domain Fake News Detection. In: Foundations and Practice of Security. FPS 2024. Lecture Notes in Computer Science. Springer, Cham.

6.1 Introduction

In the digital age, the internet has become a vast reservoir of information. Still, not all of it can be trusted. From misleading reports on global conflicts to fabricated stories about celebrities, the pervasive nature of fake news poses a significant threat. The widespread dissemination of fake information has profound societal impacts, underscoring the need for effective detection systems to combat this issue [47].

With the expansion of accessible news content, the spread of fake news has become a growing concern, threatening the trustworthiness of the information people rely on. This challenge affects individuals and can disrupt social stability, potentially triggering widespread public anxiety. For example, in 2020, one in five USA adults relied primarily on social media to stay informed about the USA presidential election [291]. Therefore, detecting fake news is essential to ensure the accuracy and trustworthiness of the consumed information. The key source of this information is social media platforms like X (formerly Twitter), where users gather news [26].

To address the challenge of detecting fake news, researchers formulated different approaches. One of these approaches is a traditional detection method that involves manual checks. For example, journalists examine news by hand with fact-checking sites [299]. However, this method could not be applied to monitor events that have recently occurred. Meanwhile, manual detection requires significant work and time, especially in an era of data explosion [299]. Thus, it is necessary to build an automated model for FND.

To address this need, researchers proposed various computational alternatives for FND. One common strategy is to apply ML, which has been shown to perform well in detecting and curtailing fake news on social media platforms [124]. However, there are some limitations to the ML algorithms. Constructing such models requires handcraft-based features. In other words, these models are based on manual feature selection, which is data-specific and onerous [124]. In addition, the manual feature extraction process also depends on the work person's background and domain knowledge, thus leading to unstable data quality. Moreover, manually selected features are derived from shallow layers, mostly induced by statistical methods, and thus cannot capture and describe semantic information at deeper layers [124].

To overcome these limitations, researchers have turned to DL strategies, a specialized subset of ML, which offer a more accurate and automated approach to recognizing fake news. Instead of manually selecting features for ML, DL learns the features and recognizes the best feature set automatically through the deep neural network, minimizing in this way the time and effort costs [27]. In addition, multiple and non-linear layers could be built into the DL model to allow it to study semantic information automatically. Overall, DL has proven advantageous over ML, leading to its adoption in recent years for FND, where it has outperformed existing ML methods with improved results [25].

More recently, the usage of LLMs for FND has further advanced this field, offering even greater accuracy and adaptability to identify fake information [25, 26]. These LLMs have advanced the field of NLP and text classification, providing tools to combat the spread of deceptive information. In particular, OpenAI GPT [235] has established a benchmark for

model performance, with competitors such as Google's BERT [293] and Meta's Llama [13] entering the field. While BERT is smaller and less powerful than GPT, Llama's competitive performance makes it an appealing choice for FND.

However, fake news is not limited to a single domain. Many news domains include health, social, politics, and others. Hence, developing a learner based on one single news domain is not general and efficient. This is called single-domain FND in which an algorithm is specifically trained to identify fake news within a particular domain. However, models trained in this manner often produce inadequate results when applied to other domains [25]. Moreover, when training one model for FND in a particular domain, the size of the available dataset is smaller, and possibly important information from other domains is discarded. As such, multi-domain FND, where information from multiple domains is considered together, has been used to address the limitations of single-domain FND [146]. Additionally, dealing with multi-domain FND requires the integration of fine-tuned and prompt-tuned LLMs such as the Llama model [26]. Fine-tuning Llama enhances its performance by customizing the model to recognize deceptive patterns and domain-specific features. Building on this foundation, prompt tuning further refines the model's capabilities by optimizing its responses to targeted inputs [122].

This research builds on the current state-of-the-art in FND, BERTGuard [26], which utilizes a two-stage detection system to classify and verify the credibility of news across multiple domains. BERTGuard has proven to be highly effective by leveraging a first-layer domain classification and a second-layer domain-specific FND model. In this chapter, we propose a modification to BERTGuard by introducing a tailored version called ProFineLlama, which enhances the second layer by integrating a fine-tuned and prompt-tuned Llama 2 7B model. The first layer still relies on DistilBERT to classify the domain of news articles, while the second layer now uses the output from the first stage to apply domain-specific prompts, improving the model's ability to focus on the unique aspects of each news domain. This two-stage approach aims to improve the accuracy of the model in detecting fake news by first fine-tuning Llama 2 for the task and then applying prompt tuning to further guide the prediction of the model.

In the context of FND, prompt tuning refers to optimizing model responses to specific prompts through minor adjustments to the pre-trained model parameters, enhancing domain-specific performance. These prompts can direct the model's attention to specific aspects of the news content, such as the credibility of the source or the truthfulness of particular claims within a given news domain. By designing prompts that align with pre-existing knowledge of the model, prompt tuning can improve the accuracy of predictions without requiring extensive retraining. Two prompts are prepared and carefully evaluated for each news domain: (1) "Classify the following <DomainName> news article as true or

fake: <NewsContent>" and (2) "Evaluate the credibility of the following <DomainName> news: <NewsContent>." This refined approach, which incorporates both fine-tuning and prompt tuning, represents a significant improvement over previous methods, offering a more targeted and accurate multi-domain FND system.

To maintain the generalizability and comprehensiveness of the model, we trained the model on a comprehensive dataset that includes various news domains. The five distinct domain categories considered are politics, health, science, crime, and social. We also tested our model, ProFineLlama, against five other alternative baseline solutions, including the current state-of-the-art BERTGuard, in multi-domain FND.

Key Contributions

In this chapter, we introduce ProFineLlama, an innovative approach to multi-domain FND that leverages TL and domain-specific prompt tuning for improved cross-domain robustness. Unlike previous methods, which often train models in single-domain contexts and struggle with unseen data, ProFineLlama incorporates fine-tuning with Llama 2 and strategically designed prompts to achieve greater adaptability across domains. The main contributions of this chapter are as follows:

- **Novel Multi-Domain FND Framework:** We present a two-tiered approach, combining domain classification and domain-specific prompts, allowing the model to adapt its predictions based on the unique features of each domain using Llama 2 7B.
- **Enhanced Cross-Domain Generalization:** By integrating domain-adaptive prompts with a fine-tuned Llama 2 model, we address the common issue of domain dependency in FND.
- **Analysis of the effect of TL**: We provide a comprehensive study on the impact on the detection accuracy of fine-tuning a pre-trained model and prompt-tuning it.
- **Extensive performance evaluation**: ProFineLlama was tested on multiple benchmark datasets, achieving a notable improvement over existing state-of-the-art methods and solutions. These experiments demonstrate that ProFineLlama outperforms existing methods by achieving an 11% improvement in accuracy and a 21% increase in F1-score across multiple benchmark datasets, establishing a new state-of-the-art in FND.

Chapter Organization

This chapter is organized as follows. Section 6.2 presents the background and literature review. In Section 6.3, we present ProFineLlama, our proposed solution for multi-domain FND. In Section 6.4, we discuss the experimental methodology that we follow in our study in this chapter, detailing the datasets and the evaluation measures used. Section 6.5 presents the results of the experiments and analyzes the results obtained including a detailed error analysis section. Lastly, Section 6.6 summarizes the chapter.

6.2 Background and Literature Review

The literature review comprehensively examines the current state of FND, the application of TL approaches, and the various datasets utilized in this domain. This review aims to establish the foundation for the methodologies and models applied in the study.

6.2.1 Multi-Domain Fake News Detection

FND involves identifying and classifying intentionally false or misleading information, typically disseminated through digital platforms [27]. Fake news is often propagated by cyborgs and social robots, which engage in online discussions and distribute false information [218]. These bots contribute to the rapid spread of fake news, affecting key areas such as health, democracy, the media, finance, and social concerns [165].

Several DL models, such as CNN, RNN, and BERT, have been developed to tackle FND tasks [171,293]. Hybrid approaches, such as CNN-LSTM models, have shown improvements over traditional techniques by utilizing pre-trained word embeddings for feature extraction and long-term dependencies [230]. Transformer-based models, particularly BERT, have emerged as highly effective for FND, achieving superior performance in cross-domain tasks [25, 219]. Advanced models such as FakeBERT [146] and BERTGuard [26] combine transformers with additional techniques, like convolutional layers, to handle multi-domain fake news classification. Despite these advancements, challenges such as data set bias and limited generalizability persist, as demonstrated in studies showing significant performance drops when models trained on specific datasets were tested on unseen data [25]. To address this, cross-domain, cross-language, and prompt tuning techniques, as applied to models like LLAMA, have shown promise in improving FND's generalizability and detection accuracy across various domains [241, 265].

6.2.2 Transfer Learning approaches for FND

TL has been widely applied in FND, particularly with transformer models like BERT in recent research studies. TL is a technique used to advance a learner in one domain by transferring knowledge from a related domain [27]. It involves taking a model trained on a large general dataset and fine-tuning it on a smaller, task-specific dataset [27]. This approach is highly effective in FND because it allows models to leverage knowledge learned from vast text corpora and apply it to the nuanced task of identifying false information. Studies have shown that TL significantly reduces the time required for model training and improves the accuracy of FND models [7, 22].

TL is particularly crucial when there is a shortage of target training data [354]. This shortage can occur because data are rare, expensive to collect and label, or otherwise inaccessible. Using existing datasets that are related to, but not identical to, the target domain makes TL a valuable strategy, particularly as large data repositories become more available. This approach has proven effective in various machine and DL applications, such as text and image classification [27].

One of the most popular and effective methods in TL involves using pre-trained models. In this approach, a model trained on a large dataset serves as a starting point for training a new model on a smaller target dataset. The concept is that the knowledge gained from the source domain can be transferred to the target domain, even if they differ, thereby enhancing the performance of the target model [143]. After this initial training, the pre-trained models undergo a fine-tuning process. Models like BERT, Llama, and GPT (Generative Pre-trained Transformer), which are pre-trained on extensive text corpora, can be fine-tuned for tasks such as FND [26, 84] to achieve high detection accuracy while using relatively low computational resources.

Isa et al. [129] demonstrated that using a pre-trained BERT model as a TL technique achieved an accuracy of 94.66%. Similarly, pre-trained BERT models have also proven effective for FND with the ISOT dataset [84]. In another instance involving BERT variants, RoBERTa achieved accuracies of 92.77% and 91.7% in the Politifact and Gossipcop datasets, respectively [237]. These results outperformed state-of-the-art techniques that did not use TL, showing average improvements of 10.49% and 14.53% in accuracy on Politifact and Gossipcop, respectively.

While using pre-trained models is an effective TL strategy leveraging knowledge from large datasets to minimize the need for extensive target domain data and potentially enhancing the performance of the target model, it is crucial to select the appropriate pre-trained model, architecture, and fine-tuning approach. This careful selection ensures that

the transferred knowledge is relevant and advantageous for the specific target task with the need to use prompts that guide these selected models.

6.2.3 Prompt-Tuning Technique

Another method of enhancing the model's domain-specific performance is by using a prompt tuning of the approach. Prompt tuning is a technique that involves crafting specific input prompts to guide the response of a pre-trained language model to a particular task [109]. Unlike traditional fine-tuning, which adjusts the model's parameters across the entire dataset, prompt tuning keeps the model's pre-trained parameters largely intact, instead optimizing a small set of parameters that influence how the model interprets the input prompts [92]. This method has been particularly effective in adapting LLMs to new tasks with minimal additional training, offering a more efficient alternative to complete fine-tuning.

In the context of FND, prompt tuning can be used to direct the model's attention to specific aspects of the news content, such as the credibility of the source or the truthfulness of specific claims. By designing prompts that frame the input data in a way that aligns with the model's pre-existing knowledge, prompt tuning can help improve the accuracy of predictions without requiring extensive retraining.

This approach generally involves designing prompts that can effectively extract relevant knowledge from pre-trained models, tailored to the specific context of the news content. For example, a study introduced the Prompt Tuning (FNDPT) FND model, which utilizes prompt-based templates to incorporate contextual information, thus improving model performance in the early stages FND [92]. Similarly, another framework, DPOD (domain-specific prompt tuning using out-of-domain data), improves the detection of out-of-context misinformation by leveraging out-of-domain data in a prompt tuning setting, thus improving generalizability and effectiveness across different domains [50].

The primary advantage of prompt tuning in FND is its efficiency [189]. By requiring fewer adjustments to the model's parameters, prompt tuning reduces the computational resources and time needed for training, making it an attractive option for rapid deployment in real-world scenarios. Additionally, prompt tuning allows for greater flexibility, as new prompts can be designed and tested without the need for full model retraining.

However, prompt tuning also presents challenges. Crafting effective prompts requires domain expertise and a deep understanding of how the model interprets language. Poorly designed prompts can lead to suboptimal performance or even introduce biases into the model predictions. In addition, prompt tuning is highly dependent on the quality of the

underlying pre-trained model; If the model's pre-training does not adequately cover the relevant domain, prompt tuning alone may not be sufficient to achieve high accuracy.

6.3 ProFineLlama: A Prompt and Fine-Tuned Transfer Learning Approach for Multi-Domain FND

In this section, we present our method, ProFineLlama, for FND, which utilizes a two-stage approach leveraging Llama 2 7B. In our solution, we combine both fine-tuning and prompt-tuning of the model to enhance the detection accuracy. Our proposed ProFineLlama works in two stages. The first stage involves obtaining the news domains using BERT to capture the distinct characteristics of various information sources. In the second stage, Llama 2 7B is employed to assess the authenticity of news within their respective domains using the knowledge obtained from the first stage. To achieve this, the domain obtained from the first stage is used as an input for the second stage. Figure 6.1 presents an overview of the proposed ProFineLlama detection approach.

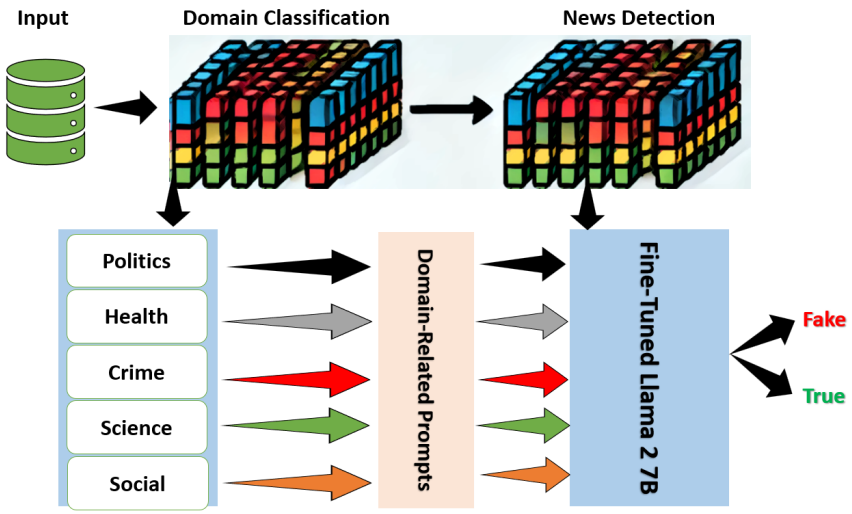

Figure 6.1: ProFineLlama FND Architecture

Figure 6.1 presents an overview of the ProFineLlama architecture. The first stage involves domain classification using DistilBERT, categorizing input news into specific domains. The second stage applies domain-specific prompts to the fine-tuned Llama 2 model, guiding it to make more accurate predictions based on the domain context.

We followed what was used in BERTGuard for the domain classification stage [26], where we found that DistilBERT, a distilled version of BERT, had better performance than other variants of BERT while being significantly faster to train and test. Thus, we use DistilBERT for the domain classification stage.

Following the domain classification stage, we enter the second stage of FND. The classified data are processed using the Llama 2 7B model with domain-specific prompts to guide the model to make more accurate predictions based on the domain context. The Llama 2 7B model used in this stage is fine-tuned for improved performance. The fine-tuning process involved systematically testing and adjusting hyperparameters to achieve optimal results.

We also carried out prompt-tuning of Llama 2 7B. A prompt engineering approach was applied to guide the model responses. Five different textual prompts were crafted and tested to determine the most effective formulation for inference. Each prompt varied in structure and wording to explore different ways of guiding the classification decisions of the model. The best-performing prompts were selected based on quantitative performance metrics, such as classification accuracy. This approach ensured that no additional fine-tuning or retraining was required after initial model adaptation. The predicted domain obtained from the DistilBERT model in the classification step was used in the selected prompts. The goal is that this information will guide the Llama model by providing a better context for the news. The Llama model outputs a Boolean value denoting whether the input is detected as fake or true.

Overall, our ProFineLlama goes beyond existing LLMs applications by combining prompt tuning with domain-specific fine-tuning, uniquely tailored to the specific challenges of multi-domain FND. Unlike prior models, ProFineLlama leverages domain-adaptive prompts to address variations across domains, achieving superior performance in handling nuanced and domain-specific language compared to other LLMs-based approaches for FND in various news domains, especially in health and politics domains.

6.3.1 Computational Environment

Given the complexity of fine-tuning large-scale DL, this research required high-performance computing resources beyond consumer-grade hardware. The Digital Research Alliance of Canada (Compute Canada) provided access to high-performance computing (HPC) clusters, enabling the efficient training and evaluation of Llama 2 and other transformer-based models.

6.3.1.1 Hardware and Computing Resources

Training LLMs such as Llama 2 demands significant VRAM to store model weights and process large datasets. Consumer-grade GPUs (e.g., Nvidia RTX 4090) lack the memory required for fine-tuning these models, while cloud-based solutions such as Google Colab and Kaggle provide insufficient computational guarantees for long-duration training. Due to these limitations, experiments were conducted using Compute Canada's Graham cluster, which provides access to Nvidia Tesla V100 32GB GPUs.

The primary hardware setup was as follows:

- Primary GPU: Nvidia Tesla V100 32GB
- Alternative GPU (when V100 was unavailable): Nvidia Tesla T4 16GB
- CPU: Intel Xeon E5-2683 v4
- Memory (RAM): 256GB DDR4
- OS: Linux Ubuntu 20.04
- Storage: Shared networked SSD storage

While attempts were made to use multiple V100 GPUs (up to six, totaling 192GB VRAM), the Transformers library did not fully support multi-GPU training in Compute Canada's cluster environment, restricting training to a single V100 GPU.

6.4 Research Methodology

6.4.1 Data Collection

We selected the most popular and publicly available datasets related to the following fake news domains as detailed in Table 5.2 in Section 5.4.1 of Chapter 5.

These datasets were used to train the models. The training was conducted on a combined training dataset comprising eight sub-datasets: FA-KES, COVID-FN, COVID-FNIR, FakeNews, ISOT, LIAR, Climate, and GossipCop. These datasets encompass five news domains: political, crime, science, social, and health, providing a diverse and representative corpus from which the model can learn. Each data set was subjected to a

preprocessing step that involved text cleaning, tokenization, and normalization to ensure consistency across different sources. The datasets were then merged into a single comprehensive training dataset.

The data is processed in a "text" column and a "label" column. The text column will contain all the content of the news article that will be used in our detection task. The label column will be '1' or '0', corresponding to 'fake' or 'not fake' news, and this will be the target of the detection model. We ensured a systematic approach by splitting the data into appropriate training and testing sets, incorporating the necessary preprocessing steps, and monitoring the duration of training.

6.4.2 Experimental Settings

In our experiments, we evaluate the proposed ProFineLlama model against five alternative baselines, which include the most recent state-of-the-art model for multi-domain FND, the BERTGuard model. These baselines are the following:

1. Baseline 1: blindly testing the original Llama 2 without fine-tuning or further training.
2. Baseline 2: fine-tune the Llama 2 7B model on the training dataset and evaluate its performance on the test datasets.
3. Baseline 3: a previously developed one-stage model [26] that does not pay attention to the news domains.
4. Baseline 4: FakeBERT, a well-known FND model [146]. It is a BERT-based model enhanced with 1D CNN layers of varying kernel sizes for improved FND. We retrained this model on our dataset to adapt it to different news domains and improve detection accuracy.
5. Baseline 5: BERTGuard, the current state-of-the-art model for multi-domain FND [26]. It employs a two-tiered approach, where the first tier involves domain classification, and the second tier focuses on FND within those specific domains and uses a specialized model for each news domain.

We designed our experiments to understand two aspects: (i) the TL effect and (ii) the impact of domain categorization and prompt tuning. To achieve this, our first set of experiments compares the impact of fine-tuning against blind testing. In this set of

experiments, we used baseline 1 and baseline 2. Our second set of experiments focuses on assessing the impact of domain categorization and prompt tuning using DistilBERT for domain classification and domain-specific prompts. These experiments included the following:

- Domain Classification with DistilBERT to classify the news articles into one of the five news domains. It will be trained on a huge labelled dataset to categorize each news article into the respective domain.
- Prompt Tuning for Llama 2 7B: Use the domain classification results in prompts for directing the Llama 2 7B model. This includes creating training and testing datasets with these prompts.
- Domain-specific prompts were created to provide context for the Llama 2 7B model and further guide the fine-tuned model for better performance. Each prompt included a clear instruction related to the domain based on the domain classification model's output, as follows:

  Prompt 1: *"Classify the following [NewsDomainName] news article as true or fake: [NewsContent]"*;

  Prompt 2: *"Evaluate the credibility of the following [NewsDomainName] news: [NewsContent]"*.

We conducted six independent runs to ensure the reliability and stability of our proposed ProFineLlama model. These multiple runs helped minimize the impact of potential biases introduced by single-run fluctuations and improved overall confidence in the reported performance metrics. By averaging the results across these runs, we ensured that the model's reported performance was not influenced by random initialization or specific data splits.

The fine-tuning process began with the pre-trained Llama 2 7B model, which was downloaded from Hugging Face and integrated with the Transformers library. The model was initially trained using a standard learning rate and batch size, but multiple configurations were tested to optimize performance. The fine-tuning involved adjusting the model weights to minimize the loss function on the training dataset. Key hyperparameters such as learning rate, batch size, the number of epochs, and gradient accumulation steps were iteratively adjusted to avoid overfitting and improve model convergence. They were carefully selected based on preliminary experiments to ensure optimal performance. This selection includes: Learning Rate: 1e-5, Batch Size: 8, Number of Epochs: 5, Optimizer: AdamW.

To ensure robust training and evaluation, we carefully curated a multi-domain fake news dataset comprising news articles from five domains: politics, health, crime, science, and social media. The 70% training division of our merged multi-domain dataset was used to update the detection model weights during fine-tuning, while the 10% validation division of the dataset was used to tune the hyperparameters and monitor generalization during training.

Regarding the prompt-tuning process, initial experiments were conducted to evaluate several prompts. The validation split of 10% of our merged multi-domain datasets was used to evaluate the effectiveness of each prompt. This validation split was used to compare different prompt variations and select the most effective ones based on accuracy, precision, recall, and F1-score. Then, a test split of 10% of the dataset was used to evaluate the performance of the selected prompt structure on completely unseen data. This iterative process allowed us to determine which prompts provided the most accurate predictions while ensuring that the model remained generalizable across domains. The prompts "Classify the following [NewsDomainName] news article as true or fake: [NewsContent]" and "Evaluate the credibility of the following [NewsDomainName] news: [NewsContent]" were ultimately chosen due to their superior performance in initial trials, as it effectively leveraged domain-specific context to enhance model accuracy across all tested domains, which yielded to a butter detection accuracy. Other prompt structures were explored to maximize the accuracy of cross-domain detection, given the diverse nature of fake news content. These variations were tested to examine whether alternative phrasings could improve accuracy in specific domains. These additional prompts included:

- Prompt 1: "What do you think about this news: [NewsContent]"
- Prompt 2: "Judge the factual accuracy of this news: [NewsContent]"
- Prompt 3: "Is this [NewsDomainName] story credible: [NewsContent]?"

These two selected prompts, the direct classification prompt ("Classify the following [NewsDomainName] news article as true or fake: [NewsContent])" and the credibility evaluation prompt ("Evaluate the credibility of the following [NewsDomainName] news: [NewsContent]"), helped guide the model and gave better detection accuracy. Our findings indicate that the direct classification prompt generally yielded higher precision and recall across most domains, particularly in the health and political domains. In contrast, the Credibility Evaluation Prompt showed improved performance in the social and crime domains, suggesting that the model benefits from prompts that align closely with the nature

of the content being evaluated. These prompts effectively leveraged domain-specific context, which proved essential in guiding the model toward more accurate predictions across domains such as health and politics.

It is worth mentioning that we recognize that LLMs such as Llama 2 may inherit biases present in their pre-training data, which could influence FND outcomes. In our approach, we minimized bias by employing domain-specific prompts to reduce reliance on potentially biased pre-trained associations.

6.4.3 Evaluation Metrics

The performance of the system was evaluated using various metrics, as detailed in Chapter 4, Section 4.3.6. These metrics were selected to provide a well-rounded evaluation of the model's capability in detecting fake news, especially in the context of imbalanced datasets. By including not only accuracy but also several metrics that are particularly suitable for imbalanced scenarios, such as F1-score, precision, recall, and G-Mean, we aim to gain a deeper understanding of how well the models perform under such challenging conditions [94].

6.5 Results and Discussion

This section presents the results of the various models tested, including the five baseline models and our proposed solution, ProFineLlama.

6.5.1 Baselines Performance

Five different baselines were used to validate our proposed model. Table 6.1 shows the overall performance of these baselines in all testing datasets. The results presented are the average obtained for all six test datasets.

We observe that ProFineLlama has a better average performance compared to all the baselines considered. We also observe that the fine-tuned Llama model (Baseline 2) provides better results than the BERTGuard model. These results show that considering a two-layered approach and entering the predicted domain in the second layer for FND detection provide important performance gains across all performance metrics.

Table 6.1: Performance comparison between ProFineLlama and various baseline models.

Model	Precision	Recall	Accuracy	F1-Score	G-Mean
Baseline 1 (Blind Llama)	44%	51%	46%	51%	47%
Baseline 2 (Fine-Tuned Llama)	86%	88%	87%	87%	87%
Baseline 3 (BERT Domainless)	42%	46%	55%	41%	43%
Baseline 4 (FakeBERT Domainless)	32%	34%	38%	31%	32%
Baseline 5 (BERTGuard Domain-Aware)	81%	68%	81%	70%	75%
ProFineLlama (Our Solution)	**90%**	**91%**	**91%**	**91%**	**91%**

These results reflect the average values for testing using various unseen datasets

Table 6.1 provides an ablation study of ProFineLlama, demonstrating the effect of each component of the model (fine tuning, prompt tuning, domain classification) on model performance across domains. By comparing all these baselines, we observe a clear improvement in accuracy and generalization with the components we added to the original Llama 2 model.

Our proposed model has shown positive effects across all datasets and performance metrics as presented in Table 6.2.

Table 6.2: Results per Dataset for Baseline 1, Baseline 2, and ProFineLlama

Dataset	Model	Precision	Recall	Accuracy	F1	G-Mean
Snope	Baseline 1: Llama 2 7B	0.383 (±0.016)	0.619 (±0.020)	0.619 (±0.019)	0.473 (±0.016)	0.487 (±0.017)
	Baseline 2: Fine-tuned Llama 2 7B	0.712 (±0.018)	0.791 (±0.008)	0.731 (±0.010)	0.750 (±0.008)	0.750 (±0.012)
	ProFineLlama	**0.826** (±0.003)	**0.880** (±0.005)	**0.852** (±0.004)	**0.852** (±0.004)	**0.852** (±0.004)
COVID-Claims	Baseline 1: Llama 2 7B	0.189 (±0.018)	0.435 (±0.022)	0.435 (±0.020)	0.263 (±0.017)	0.287 (±0.019)
	Baseline 2: Fine-tuned Llama 2 7B	0.964 (±0.004)	**0.984** (±0.003)	0.974 (±0.003)	0.974 (±0.004)	0.974 (±0.003)
	ProFineLlama	**0.984** (±0.003)	0.980 (±0.003)	**0.987** (±0.002)	**0.982** (±0.002)	**0.982** (±0.002)
Politifact	Baseline 1: Llama 2 7B	0.513 (±0.013)	0.480 (±0.016)	0.480 (±0.013)	0.496 (±0.012)	0.496 (±0.014)
	Baseline 2: Fine-tuned Llama 2 7B	0.916 (±0.011)	0.915 (±0.002)	0.915 (±0.002)	0.916 (±0.002)	0.916 (±0.007)
	ProFineLlama	**0.940** (±0.004)	**0.950** (±0.004)	**0.940** (±0.004)	**0.945** (±0.003)	**0.945** (±0.004)
Pheme	Baseline 1: Llama 2 7B	0.463 (±0.019)	0.473 (±0.022)	0.473 (±0.021)	0.468 (±0.019)	0.468 (±0.021)
	Baseline 2: Fine-tuned Llama 2 7B	0.902 (±0.018)	0.900 (±0.011)	0.900 (±0.015)	0.901 (±0.015)	0.901 (±0.014)
	ProFineLlama	**0.923** (±0.011)	**0.922** (±0.011)	**0.922** (±0.011)	**0.923** (±0.010)	**0.923** (±0.011)
ISOT-science	Baseline 1: Llama 2 7B	0.636 (±0.020)	0.636 (±0.020)	0.636 (±0.021)	0.636 (±0.020)	0.636 (±0.020)
	Baseline 2: Fine-tuned Llama 2 7B	0.796 (±0.016)	0.802 (±0.011)	0.802 (±0.013)	0.799 (±0.017)	0.799 (±0.014)
	ProFineLlama	**0.823** (±0.006)	**0.820** (±0.008)	**0.829** (±0.010)	**0.822** (±0.001)	**0.822** (±0.005)
ISOT-social	Baseline 1: Llama 2 7B	0.458 (±0.017)	0.442 (±0.019)	0.442 (±0.019)	0.450 (±0.018)	0.450 (±0.018)
	Baseline 2: Fine-tuned Llama 2 7B	0.884 (±0.009)	0.8871 (±0.019)	0.887 (±0.007)	0.885 (±0.007)	0.885 (±0.013)
	ProFineLlama	**0.919** (±0.004)	**0.913** (±0.002)	**0.920** (±0.004)	**0.916** (±0.001)	**0.916** (±0.003)

Upon fine-tuning the Llama 2 7B model as a TL approach, the results improved dramatically across all datasets compared to Baseline 1 (testing Llama 2 without fine-tuning). Notably, the model achieved nearly perfect scores in the COVID-Claims dataset, with a Precision of 96.38%, Recall of 98.4%, Accuracy of 97.4%, and F1 score of 97.37%. This improvement indicates that TL significantly improves the model's ability to generalize and

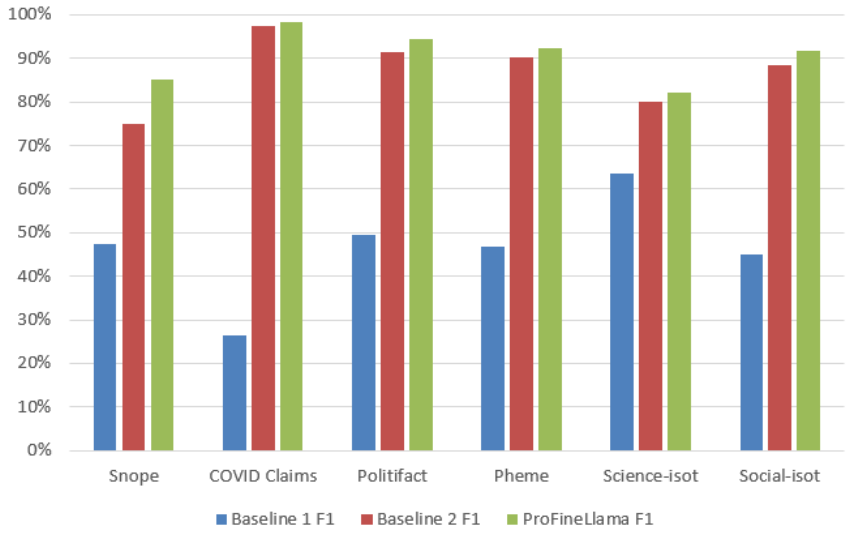

Figure 6.2: F1-scores of the alternative baselines vs. the proposed ProFineLlama

detect fake news, especially in health-related content. The Politifact and Pheme datasets also exhibited substantial improvements, with F1 scores of 91.56% and 90.13%, respectively, indicating a better handling of political FND. Across all datasets, the F1 scores and overall accuracy were consistently high, reflecting the efficiency of the fine-tuned model. Additionally, the G-Mean results further confirm the balanced classification performance of the model, particularly in data sets with a higher class imbalance ratio. The higher G-Mean values indicate that the model is capable of correctly identifying both fake and real news without disproportionately favoring one class over the other. For example, in the COVID-Claims dataset, the G-Mean score matches the F1-score, reinforcing that both positive and negative instances were handled effectively.

6.5.2 ProFineLlama Performance: Prompt-tuning the Fine-tuned Llama 2 7B

Further improvements were observed when applying prompt tuning in combination with fine-tuning which helped in directing the model in the detection process. Figure 6.2 depicts the improvement that occurred with the use of prompt tuning compared to using Baseline 1 (the blind Llama) and Baseline 2 (the fine-tuned Llama).

For the Snope dataset, the model's Precision improved to 82.5%, Recall to 88%, Accu-

racy to 85.2%, F1 score to 85.2%, and G-Mean to 85.2%. Across most datasets, Precision, Recall, F1 scores, and G-Mean all increased, particularly in the Politifact and Pheme datasets, where the F1 score increased to 94.5% and 92.3%, respectively. The COVID-Claims dataset reached its highest performance of 98.2% for both G-Mean and F1. The additional layer of prompt tuning seems to have enhanced the model's understanding of the contextual patterns in FND across domains.

Additionally, the proposed ProFineLlama with Prompt-tuned Fine-tuned Llama 2 7B recorded a noticeable performance when it was compared with the best-performing models in FND in the literature. For instance, BERTGuard [26], Baseline 5 in this study, has achieved the best detection performance in the multi-stage detection systems. This baseline underperformed significantly in the Politics domain with the Politifact dataset, with an F1 score of only 15%, compared to our model's score of 94.5%. In the Health domain, COVID Claims dataset presents a more competitive comparison, with BERTGuard achieving an F1 score of 96.6%, which is slightly lower than our model's score of 98.2%. The same pattern is visible on the Social-isot dataset where our proposed model outperformed BERTGuard, with F1 scores of 88% for BERTGuard and around 92% for our model.

These results demonstrate the superiority of utilizing ProFineLlama, our proposed two-layer Prompt fine-tuned Llama 2 model. This is particularly clear in the Politics domain with the Politifact dataset and the Health domain with the COVID Claims dataset, where BERTGuard struggled the most. Figure 6.3 shows the performance of all baselines compared to our proposed two-stages prompt-tuned the fine-tuned Llama 2 7B, ProFineLlama.

The F1-score results show that the proposed model, ProFineLlama, significantly outperforms all baselines across most news domains. The prompt tuning step in the proposed ProFineLlama model significantly enhanced the representation of several challenging news domains that had previously struggled with low accuracy or F1 scores. For example, the crime, health, and social news domains saw dramatic improvements, with F1 scores reaching 85.2%, 98.2%, and 91.6%, respectively. Beyond F1-score, ProFineLlama also achieved exceptionally high G-Mean values, demonstrating its ability to maintain a strong balance between precision and recall across all domains. Notably, the political dataset Politifact' improved to a G-Mean of 94.5%, and Pheme' reached 92.3%, showcasing the model's ability to capture both fake and real news instances effectively. This substantial improvement stems from the fine-tuning and prompt-tuning strategies, which allow the model to adapt more precisely to domain-specific language patterns, reducing misclassification and enhancing generalization. Moreover, the political dataset Pheme, which had previously performed poorly in all previous work and experiments, experienced a substantial boost, achieving 92.3% with ProFineLlama. In contrast, earlier methods, such as BERT Domainless and FakeBERT, failed to address the low performance in these domains, particularly in datasets

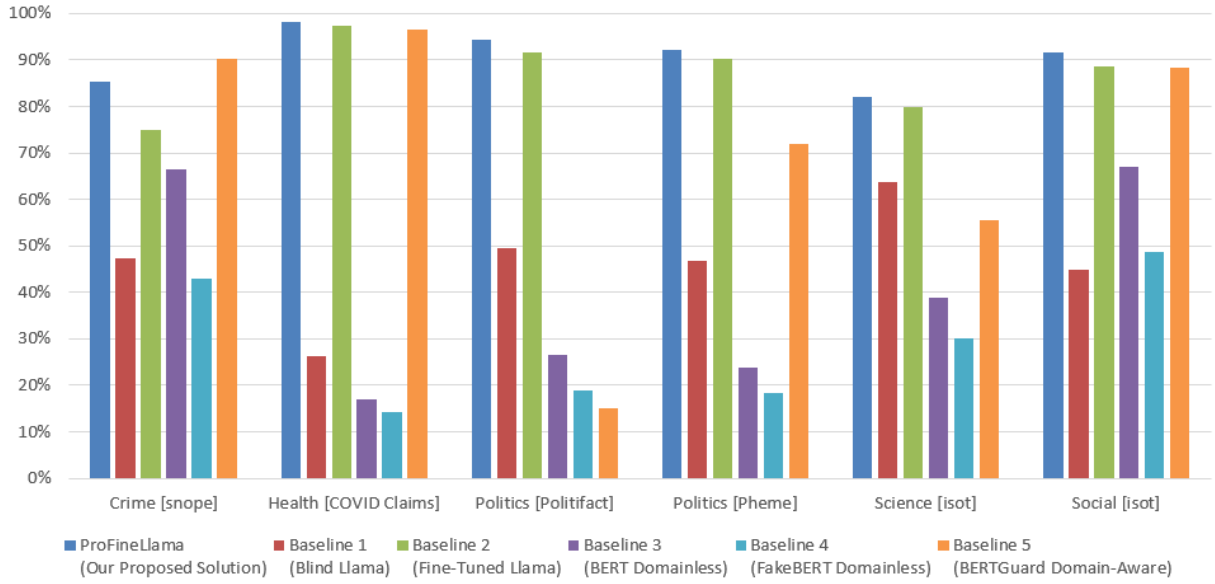

Figure 6.3: F1-scores of the Baselines vs. the Proposed ProFineLlama

belonging to crime, social, and political (specifically PHEME) domains, where accuracy consistently lagged. None of the previous models were able to sufficiently resolve these issues, highlighting the critical role of domain-specific prompt tuning in improving FND across diverse datasets. The combination of prompt tuning and fine-tuning in ProFineLlama ensures that the model not only achieves high accuracy but also maintains strong robustness across domains, as reflected in the consistently high G-Mean values. Overall, the results demonstrate the superiority of ProFineLlama in handling multi-domain FND, particularly when compared to various models that either lack domain adaptation or use domain-specialized approaches.

6.5.3 Error Analysis

In this section, we conduct a comprehensive error analysis of our proposed model, ProFineLlama, which employs a two-tiered approach for FND. The analysis focuses on the two levels used in the model: domain classification and fake news detection, taking into account the prompts used. This analysis aims to identify the types of errors made by the model, understand their causes, and suggest potential improvements.

1. Domain Classification Errors

The first stage of ProFineLlama involves classifying news articles into specific domains using DistilBERT. Errors at this stage can significantly impact the subsequent FND process. We analyzed the misclassified instances and identified the following types of errors:

(a) Misclassification of Ambiguous Domains: Some news articles contain content that spans multiple domains, making it challenging for the model to classify them accurately. For example, an article discussing the political implications of a health crisis was misclassified as political rather than health news, depending on the emphasis of the text.

(b) Misjudgment of Contextual Keywords: The model sometimes misclassified articles based on the presence of high-frequency keywords that are strongly associated with a particular domain. For instance, articles containing terms like "election" or "vaccine" were misclassified since the context did not align with the typical usage of these keywords in the training data.

(c) Annotation Errors: Errors in the labeling of the training dataset led to misclassification. If the ground truth labels were incorrect or inconsistent, the model's predictions reflected these inaccuracies. For example, an article labeled as "social" actually belonged to the "political" domain, leading to confusion during training and testing.

2. News Detection Errors

The second stage involves detecting fake news within the classified domains using a fine-tuned and prompt-tuned Llama 2 model. We analyzed the misclassified instances in this stage and identified the following types of errors:

(a) False Positives Errors: Instances where real news articles were incorrectly classified as fake news. This type of error occurred due to:

- Misleading Headlines: Real news articles with sensational or misleading headlines were flagged as fake news.
- Complex Language: Articles with complex language or nuanced arguments were misinterpreted by the model as fake news.

(b) False Negatives Errors: Instances where fake news articles were incorrectly classified as real news. This type of error occurred due to:

- Sophisticated Fake News: Well-crafted fake news articles that closely mimicked the style and content of real news evaded detection.

- Lack of Context: The model missed subtle cues indicating fake news if the context was not fully captured by the prompts.

(c) Prompt-Related Errors: Despite using the best-performing prompts, some errors persisted. The effectiveness of prompt tuning can vary based on the design of the prompts. We identified the following issues:

- Inadequate Prompts: Prompts that did not fully capture the context or specific characteristics of the domain led to incorrect predictions. For example, a prompt like "Classify the following political news article as true or fake" did not provide enough context for nuanced political articles.
- Overfitting to Prompts: The model overfitted to specific prompts used during training, leading to reduced generalization when encountering slightly different prompts or real-world scenarios.

Table 6.3 provides examples of misclassified instances by showing the types of errors identified, error examples, and the percentage of errors, which presents the proportion of each specific type of error relative to the total number of misclassified instances. Instances with incorrect domain classification are excluded from the next stage of FND analysis to prevent the propagation of errors between stages.

Table 6.3: Examples of Misclassified Instances and Their Distribution

Error Type	Example	Ground Truth	Predicted	Percentage of Errors
Domain Classification Errors				
Misclassification of Ambiguous Domains	"The political impact of the COVID-19 vaccine rollout."	Health	Political	15%
Misjudgment of Contextual Keywords	"Election results show a significant impact on vaccine distribution."	Health	Political	10%
Annotation Errors	An article labeled as 'social' actually belonged to the 'political' domain.	Political	Social	15%
News Detection Errors				
False Positives (Type I Errors)	"Breaking: New vaccine shows 95% efficacy in trials."	Real	Fake	20%
False Negatives (Type II Errors)	"Miracle cure for COVID-19 discovered in remote village."	Fake	Real	25%
Prompt-Related Errors	"Classify the following political news article as true or fake: [NewsContent]"	Real	Fake	15%

The error analysis highlights the complexity of the FND task and the challenges associated with domain classification and news detection. Misclassification of ambiguous domains, misjudgment of contextual keywords, and annotation errors were significant sources of errors in the domain classification stage. In the news detection stage, false positives, false negatives, and prompt-related errors were prominent. Addressing these issues requires

improving the quality and consistency of the training data, refining the design of prompts, and enhancing the model's ability to capture contextual nuances.

To improve the model's accuracy and reliability in both domain classification and news detection tasks, it is essential to enhance the quality of training data by ensuring it is comprehensive, accurately labeled, and representative of various domains and types of news. Regular updates and expansions of the dataset will help maintain its relevance and effectiveness. Additionally, refining prompt design by developing clear, precise, and contextually relevant prompts can significantly reduce prompt-related errors. This process involves iterative testing and feedback loops to continually improve prompt quality. Furthermore, leveraging advanced NLP techniques to better capture the nuances of context and domain-specific language can enhance the model's performance. This includes using contextual embeddings and attention mechanisms to focus on relevant parts of the text.

6.5.4 Scalability and Deployment Considerations

While ProFineLlama demonstrates robust performance across multiple domains, the scalability and deployment of such a large model present significant challenges. Llama 2 7B, being a substantial model with 7 billion parameters, requires considerable computational resources for both training and inference. This poses potential limitations for deployment in resource-constrained environments, such as edge devices or organizations with limited access to high-performance computing infrastructure.

6.5.4.1 Computational Requirements

ProFineLlama's deployment necessitates powerful GPUs to handle real-time inference efficiently. The model's memory footprint and processing demands can lead to increased latency and higher operational costs, especially when scaling to handle large volumes of data or integrating into real-time systems. To mitigate these challenges, several strategies could be considered:

- Model Optimization: Techniques such as model pruning, quantization, and knowledge distillation can reduce the model size and computational requirements without significantly compromising performance.
- Distributed Computing: Leveraging cloud-based platforms and distributed computing frameworks can distribute the computational load, improving scalability, and reducing latency.

- Batch Processing: Implementing batch processing for inference can optimize resource utilization and improve throughput in environments where real-time processing is not critical.

6.5.4.2 Deployment Strategies

For large-scale deployment, ProFineLlama could be integrated into cloud-based services where computational resources can be dynamically allocated based on demand. Additionally, containerization technologies such as Docker and orchestration tools like Kubernetes can facilitate scalable and efficient deployment, ensuring that the model can handle varying workloads seamlessly.

6.5.5 Discussion of Different News Content

One important aspect of the detection process is the ability to handle other types of news, such as sarcasm and clickbait, as misinformation can take various forms beyond straightforward false claims.

Sarcasm presents a unique challenge in FND, as it often conveys meaning that contradicts the literal text. Since Llama 2 relies on textual patterns, sarcasm detection may require additional contextual understanding, such as sentiment shifts, historical author behavior, or multimodal cues (e.g., images or tone in speech). Future work could involve evaluating ProFineLlama on sarcasm-heavy datasets or integrating external tools, such as sentiment analysis models, to improve interpretation.

Similarly, clickbait headlines use exaggeration and sensationalism to mislead readers. While ProFineLlama was not explicitly tested on clickbait detection, its ability to recognize manipulative language patterns could make it suitable for this task. A targeted evaluation of clickbait datasets would help determine how effectively the model distinguishes between engagement-driven exaggeration and actual fake news.

Preliminary experiments with an unseen news category suggest that ProFineLlama can generalize beyond its trained domains. We conducted an additional experiment using a news sample outside the categories covered in the study from the 'entertainment domain'. The model successfully aligned the sample with the most relevant domain and correctly classified it to 'social' domain, demonstrating its adaptability to unseen news topics. However, further testing on different news kinds, such as sarcasm, clickbait, and other news domains, would be beneficial in understanding the model's full adaptability and limitations.

6.6 Summary

In this chapter, we proposed a novel approach, ProFineLlama, which leverages fine-tuning and prompt tuning along with a domain classification step to enhance multi-domain FND. The proposed model significantly outperformed all alternatives in five news domains: Crime, Health, Politics, Science, and Social. The results demonstrated that our model achieved the highest overall performance, indicating its robustness across diverse domains. In comparison, baseline models performed poorly, particularly baseline 1 (Blind Llama), which lacked domain knowledge and fine-tuning and resulted in lower accuracy scores. This underscores the importance of applying TL and task-specific training along with domain adaptation. Baseline 2, which was fine-tuned but lacked prompt tuning, showed marked improvement across domains but still underperformed compared to our proposed model. The outcomes of the remaining non-domain aware baselines show even worse performance, which suggests that domain-specific adaptation is critical for effective FND. The following Chapter 7 offers the research conclusions and a discussion of future work.

Chapter 7

Conclusion and Future Work

This chapter presents the conclusion of the thesis, summarizing the key contributions, findings, and limitations. The work addresses three core challenges in FND: cross-domain generalization, class imbalance, and the effect of TL techniques in FND. Through the development of DL-based and LLMs-based solutions, the research demonstrated significant improvements in the detection of fake news across multiple datasets.

7.1 Conclusion

The primary objective of this thesis was to enhance the performance of FND models, particularly in cross-domain generalization and class imbalance mitigation. This research contributed novel solutions, including a two-tiered multi-domain detection system, BERT-Guard, and a TL-based approach leveraging prompt-tuning and fine-tuning of pre-trained models, ProFineLlama.

At first, our comprehensive SLR provided a solid foundation for understanding the limitations of current FND methods. Second, the evaluation of DL models in cross-domain settings revealed significant performance degradation, prompting the development of a new multi-domain detection framework. Third, the introduction of class imbalance handling techniques along with domain-specific model for accurate FND further improved the robustness of the models. Finally, the proposed solutions have been embedded with tailored prompts to guide the models throughout the detection process. The research findings, validated through extensive experimentation, set a new standard for FND across diverse domains.

In conclusion, this thesis made four key contributions:

1. A comprehensive review of existing literature on FND was conducted, highlighting the strengths and limitations of various DL models, TL techniques, and strategies to address class imbalance. This review set the foundation for the novel contributions presented in subsequent chapters.

2. Various DL architectures, including BERT and BiLSTM, were evaluated for their performance in cross-domain FND. The assessment revealed how well these models performed across multiple datasets and identified the limitations of current approaches, particularly their inability to generalize to new domains.

3. A novel two-tiered FND system, BERTGuard, was introduced to enhance cross-domain generalization and address class imbalance. This approach combined feature extraction with imbalance handling techniques to achieve superior results across multiple domains.

4. A refined TL strategy, ProFineLlama, incorporating prompt-tuning and fine-tuning of Llama 2 models was presented. This approach showed significant improvements in FND, especially with tackling the domain classification before the detection.

These contributions collectively advanced the field of FND by addressing critical challenges and proposing innovative solutions.

7.2 Limitations

While this thesis achieved significant improvements in FND across multiple domains, several limitations remain, which are critical to address for further advancements.

Firstly, generalization across entirely new domains continues to be a significant challenge. Although the models have shown improved cross-domain performance, their ability to generalize effectively is limited when applied to datasets that contain vastly different structures or topics. This results in performance drops when the models encounter unfamiliar data patterns, reducing their adaptability to new or evolving domains.

Another limitation lies in the availability of large and well-annotated datasets. High-quality domain-specific data are essential for training robust models in FND, yet accessible data sets are often limited, particularly on diverse topics. This bottleneck restricts both the

training and the evaluation stages, which in turn hinders the comprehensive applicability and accuracy of the models across domains.

In addition, coping with Ambiguous News Content is also a noticeable limitation. Although our proposed solution demonstrates robust performance across various domains, the challenge of classifying ambiguous news content remains an area for future improvement. Ambiguous news items, often containing partially true or misleading information, pose unique challenges to classification systems as these articles do not fit strictly within "true" or "false" labels. This solution could benefit from incorporating additional subclassifications, such as satire or opinion, which would refine its detection abilities.

To handle real-world complexity, future enhancements could include training ProFineLlama on a dataset with more nuanced categories, allowing it to discern degrees of truthfulness. In addition, a confidence scoring mechanism can be implemented for borderline cases, providing users with a clearer understanding of uncertainty in classifications. Such adaptations will enable the ProFineLlama solution to better manage the intricacies of authentic news content and improve its applicability in dynamic news environments.

Finally, the computational requirements of transformer-based architectures present a major challenge. The models developed in this research are computationally intensive, requiring significant resources for training and fine-tuning, which complicates their deployment in real-time or at large scales. This limitation impacts the feasibility of applying these models in environments where computational resources are limited or the response time is critical.

Addressing these limitations requires several strategic approaches. To improve generalization, incorporating domain adaptation techniques and exploring meta-learning approaches could enhance the flexibility of the model when handling diverse domains. Expanding and diversifying available datasets through collaborative data sharing and data augmentation techniques could help mitigate the data scarcity issue, making training more robust. To overcome computational constraints, exploring model optimization techniques such as quantization, pruning, and distillation, as well as utilizing cloud-based resources, can make transformer-based models more accessible for real-time applications. By tackling these areas, future research can push towards more adaptable, efficient, and inclusive solutions in FND.

7.3 Future Work

Future research will focus on addressing the remaining challenges highlighted in this thesis:

- Enhancing Generalization: Exploring domain adaptation techniques and more advanced TL strategies can help improve the model's ability to generalize across entirely new domains.

- Creating New Datasets: Collaborating with media organizations and fact-checking bodies to develop larger and more diverse datasets will provide the necessary resources to train more robust models. The collection and curating of high-quality datasets from the sports, entertainment, and finance domains will improve the generality of models' training.

- Coping with Ambiguous News Content: Future work will explore ProFineLlama's ability to classify ambiguous or partially true news content. Integrating additional sub-classifications, such as satire, clickbait, sarcasm, or opinion, would improve the model's detection abilities in real-world scenarios where content often contains elements of truth mixed with fake news. Implementing confidence scores for borderline cases could further enhance the usability of ProFineLlama, allowing users to gauge the certainty of classifications.

- Resource-Efficient Models: Future work should focus on optimizing DL models for efficiency, including the use of model distillation and pruning techniques to reduce the computational cost of large transformer models without sacrificing performance.

- Real-World Application in Dynamic News Streams: While ProFineLlama has shown robust performance across multiple domains in a static testing environment, its adaptability to dynamic real-world news streams is essential for practical application. In a live environment, news topics and fake news trends evolve rapidly, requiring the model to continuously learn and adapt. This is a challenging deployment setting for a FND solution. ProFineLlama's modular architecture, which integrates domain classification with domain-specific prompt tuning, makes it particularly suitable for deployment in dynamic settings.

  An approach for a real-world application could involve deploying ProFineLlama within a continuous learning framework. Such a framework would allow periodic retraining on updated datasets, capturing emerging trends and evolving language patterns in fake news across domains like politics, health, and social networks. Furthermore, incorporating real-time news feeds would enable ongoing assessments, ensuring that the model remains effective in detecting fake news as it develops. This adaptability demonstrates ProFineLlama's potential to serve as a reliable tool for real-time FND in dynamic media landscapes in the future.

- Expanding Domain Diversity: To further establish ProFineLlama's robustness, future efforts should consider expanding testing to additional domains such as sports, entertainment, and finance. These domains introduce unique linguistic styles, terminologies, and types of misinformation that differ from those in politics, health, and social media. Evaluating ProFineLlama in these new contexts will provide a more comprehensive assessment of its generalization capabilities and identify any domain-specific limitations.

References

[1] Qamber Abbas, Muhammad Umar Zeshan, and Muhammad Asif. A cnn-rnn based fake news detection model using deep learning. In *2022 International Seminar on Computer Science and Engineering Technology (SCSET)*, pages 40–45. IEEE, 2022.

[2] A Abdullah, M Awan, M Shehzad, and M Ashraf. Fake news classification bimodal using convolutional neural network and long short-term memory. *Int. J. Emerg. Technol. Learn*, 11:209–212, 2020.

[3] Ayat Abedalla, Aisha Al-Sadi, and Malak Abdullah. A closer look at fake news detection: A deep learning perspective. In *Proceedings of the 2019 3rd international conference on advances in artificial intelligence*, pages 24–28, 2019.

[4] Fatima K. Abu Salem, Roaa Al Feel, Shady Elbassuoni, Mohamad Jaber, and May Farah. Fa-kes: A fake news dataset around the syrian war. *Proceedings of the International AAAI Conference on Web and Social Media*, 13(01):573–582, July 2019.

[5] Giuseppe Aceto, Domenico Ciuonzo, Antonio Montieri, and Antonio Pescapé. Mobile encrypted traffic classification using deep learning: Experimental evaluation, lessons learned, and challenges. *IEEE Transactions on Network and Service Management*, 16(2):445–458, 2019.

[6] Isha Y Agarwal and Dipti P Rana. Fake news and imbalanced data perspective. In *Data Preprocessing, Active Learning, and Cost Perceptive Approaches for Resolving Data Imbalance*, pages 195–210. IGI Global, 2021.

[7] Akshay Aggarwal, Aniruddha Chauhan, D. Kumar, Mamta Mittal, and Sharad Verma. Classification of fake news by fine-tuning deep bidirectional transformers based language model. *EAI Endorsed Trans. Scalable Inf. Syst.*, 7:e10, 2018.

[8] David W Aha, Dennis Kibler, and Marc K Albert. Instance-based learning algorithms. *Machine learning*, 6:37–66, 1991.

[9] Iftikhar Ahmad, Muhammad Yousaf, Suhail Yousaf, and Muhammad Ovais Ahmad. Fake news detection using machine learning ensemble methods. *Complexity*, 2020:1–11, 2020.

[10] Tahir Ahmad, Muhammad Shahzad Faisal, Atif Rizwan, Reem Alkanhel, Prince Waqas Khan, and Ammar Muthanna. Efficient fake news detection mechanism using enhanced deep learning model. *Applied Sciences*, 12(3):1743, 2022.

[11] Hadeer Ahmed, Issa Traore, and Sherif Saad. Detection of online fake news using n-gram analysis and machine learning techniques. In *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments: First International Conference, ISDDC 2017, Vancouver, BC, Canada, October 26-28, 2017, Proceedings 1*, pages 127–138. Springer, 2017.

[12] Nishtha Ahuja and Shailender Kumar. S-han: Hierarchical attention networks with stacked gated recurrent unit for fake news detection. In *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*, pages 873–877. IEEE, 2020.

[13] Meta AI. Llama 2: Open Foundation and Fine-Tuned Chat Models. https://ai.meta.com/research/publications/llama-2-open-foundation-and-fine-tuned-chat-models/, 2023. Accessed: 2024-11-29.

[14] Oluwaseun Ajao, Deepayan Bhowmik, and Shahrzad Zargari. Fake news identification on twitter with hybrid cnn and rnn models. In *Proceedings of the 9th international conference on social media and society*, pages 226–230, 2018.

[15] Abdulhameed Al Obaid, Hassan Khotanlou, Muharram Mansoorizadeh, and Davood Zabihzadeh. Multimodal fake-news recognition using ensemble of deep learners. *Entropy*, 24(9):1242, 2022.

[16] Mohammed Al-Sarem, Abdullah Alsaeedi, Faisal Saeed, Wadii Boulila, and Omair AmeerBakhsh. A novel hybrid deep learning model for detecting covid-19-related rumors on social media based on lstm and concatenated parallel cnns. *Applied Sciences*, 11(17):7940, 2021.

[17] Firoj Alam, Fahim Dalvi, Shaden Shaar, Nadir Durrani, Hamdy Mubarak, Alex Nikolov, Giovanni Da San Martino, Ahmed Abdelali, Hassan Sajjad, Kareem Darwish, et al. Fighting the covid-19 infodemic in social media: a holistic perspective

and a call to arms. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 913–922, 2021.

[18] Mohammed N Alenezi and Zainab M Alqenaei. Machine learning in detecting covid-19 misinformation on twitter. *Future Internet*, 13(10):244, 2021.

[19] Jawaher Alghamdi, Yuqing Lin, and Suhuai Luo. Does context matter? effective deep learning approaches to curb fake news dissemination on social media. *Applied Sciences*, 13(5):3345, 2023.

[20] Jawaher Alghamdi, Yuqing Lin, and Suhuai Luo. Towards covid-19 fake news detection using transformer-based models. *Knowledge-Based Systems*, 274:110642, 2023.

[21] H Ali, MS Khan, A AlGhadhban, M Alazmi, A Alzamil, K Al-utaibi, and J Qadir. Analyzing the robustness of fake-news detectors under black-box adversarial attacks. *IEEE Access*, 9:81678–81692, 2021.

[22] Shadi A. Aljawarneh and Safa Swedat. Fake news detection using enhanced bert. *IEEE Transactions on Computational Social Systems*, 2022.

[23] Sarah A Alkhodair, Steven HH Ding, Benjamin CM Fung, and Junqiang Liu. Detecting breaking news rumors of emerging topics in social media. *Information Processing & Management*, 57(2):102018, 2020.

[24] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–236, 2017.

[25] Mohammad Q Alnabhan and Paula Branco. Evaluating deep learning for cross-domains fake news detection. In *International Symposium on Foundations and Practice of Security*, pages 40–51. Springer, 2023.

[26] Mohammad Q. Alnabhan and Paula Branco. Bertguard: A two-tiered multi-domain fake news detection with class imbalance mitigation. *Big Data and Cognitive Computing*, 8(8), 2024.

[27] Mohammad Q. Alnabhan and Paula Branco. Fake news detection using deep learning: A systematic literature review. *IEEE Access*, 12:114435–114459, 2024.

[28] Njood Mohammed AlShariah, A Khader, and J Saudagar. Detecting fake images on social media using machine learning. *International Journal of Advanced Computer Science and Applications*, 10(12):170–176, 2019.

[29] Naomi S Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992.

[30] Zaid Alyafeai, Maged Saeed AlShaibani, and Irfan Ahmad. A survey on transfer learning in natural language processing. *arXiv preprint arXiv:2007.04239*, 2020.

[31] Shatha Alyoubi, Manal Kalkatawi, and Felwa Abukhodair. The detection of fake news in arabic tweets using deep learning. *Applied Sciences*, 13(14):8209, 2023.

[32] Eslam Amer, Kyung-Sup Kwak, and Shaker El-Sappagh. Context-based fake news detection model relying on deep learning models. *Electronics*, 11(8):1255, 2022.

[33] Ayush Anand, Raghavendra Kulkarni, and Pragati Agrawal. Fake news identification: An effective combined approach using ml and dl techniques. In *2023 2nd International Conference on Paradigm Shifts in Communications Embedded Systems, Machine Learning and Signal Processing (PCEMS)*, pages 1–6. IEEE, 2023.

[34] G Anusha, G Praveen, D Mounika, U Sai Krishna, and R Cristin. Detection of fake news using recurrent neural network. In *2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)*, pages 1–5. IEEE, 2022.

[35] Fatma Arslan, Naeemul Hassan, Chengkai Li, and Mark Tremayne. A benchmark dataset of check-worthy factual claims. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 821–829, 2020.

[36] Muhammad Zubair Asghar, Ammara Habib, Anam Habib, Adil Khan, Rehman Ali, and Asad Khattak. Exploring deep neural networks for rumor detection. *Journal of Ambient Intelligence and Humanized Computing*, 12:4315–4333, 2021.

[37] Matin N Ashtiani and Bijan Raahemi. An efficient resampling technique for financial statements fraud detection: A comparative study. In *2023 3rd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, pages 1–7. IEEE, 2023.

[38] Nida Aslam, Irfan Ullah Khan, Farah Salem Alotaibi, Lama Abdulaziz Aldaej, and Asma Khaled Aldubaikil. Fake detect: A deep learning ensemble model for fake news detection. *complexity*, 2021:1–8, 2021.

[39] Avast. The year of fake news covid related scams and ransomware. https://www.prnewswire.com/news-releases/

2020-the-year-of-fake-news-covid-related-scams-and-ransomware-301180568.html, 2020. Accessed: 2024-12-12.

[40] Raad Sadi Aziz, Ahmed T Sadiq, Monji Kherallah, and Ali Douik. Arabic fake news detection for covid-19 using deep learning and machine learning. *Periodicals of Engineering and Natural Sciences*, 11(6):56–72, 2023.

[41] Jessica Babineau. Product review: Covidence (systematic review software). *Journal of the Canadian Health Libraries Association/Journal de l'Association des bibliothèques de la santé du Canada*, 35(2):68–71, 2014.

[42] Na Bai, Fanrong Meng, Xiaobin Rui, and Zhixiao Wang. Rumour detection based on graph convolutional neural net. *IEEE Access*, 9:21686–21693, 2021.

[43] Shashikant Mahadu Bankar and Sanjeev Kumar Gupta. Fake news detection using lstm-based deep learning approach and word embedding feature extraction. In *International Conference on Communication, Electronics and Digital Technology*, pages 129–141. Springer, 2023.

[44] Santosh Kumar Bharti, Ramkrushna Pradhan, Korra Sathya Babu, and Sanjay Kumar Jena. Sarcasm analysis on twitter data using machine learning approaches. *Trends in Social Network Analysis: Information Propagation, User Behavior Modeling, Forecasting, and Vulnerability Assessment*, pages 51–76, 2017.

[45] Thanaphan Bhatia, Bundit Manaskasemsak, and Arnon Rungsawang. Detecting fake news sources on twitter using deep neural network. In *2023 11th international conference on information and education technology (ICIET)*, pages 508–512. IEEE, 2023.

[46] Carlo Bianchini, Ivana Truccolo, Ettore Bidoli, CRO Information Quality Assessment Group, and Mauro Mazzocut. Avoiding misleading information: a study of complementary medicine online information for cancer patients. *Library & Information Science Research*, 41(1):67–77, 2019.

[47] D. Boissonneault and E. Hensen. Fake news detection with large language models on the liar dataset. *Research Square*, 2024.

[48] Rabia Bounaama and Mohammed El Amine Abderrahim. Classifying covid-19 related tweets for fake news detection and sentiment analysis with bert-based models. *arXiv preprint arXiv:2304.00636*, 2023.

[49] Alexandre Bovet and Hernán A Makse. Influence of fake news in twitter during the 2016 us presidential election. *Nature communications*, 10(1):7, 2019.

[50] Debarshi Brahma, Amartya Bhattacharya, Suraj Nagaje Mahadev, Anmol Asati, Vikas Verma, and Soma Biswas. Leveraging out-of-domain data for domain-specific prompt tuning in multi-modal fake news detection. *ArXiv*, abs/2311.16496, 2023.

[51] Paula Branco, Luís Torgo, and Rita P Ribeiro. A survey of predictive modeling on imbalanced domains. *ACM computing surveys (CSUR)*, 49(2):1–50, 2016.

[52] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[53] John Brummette, Marcia DiStaso, Michail Vafeiadis, and Marcus Messner. Read all about it: The politicization of "fake news" on twitter. *Journalism & Mass Communication Quarterly*, 95(2):497–517, 2018.

[54] Jannis Bulian, Jordan Boyd-Graber, Markus Leippold, Massimiliano Ciaramita, and Thomas Diggelmann. Climate-fever: A dataset for verification of real-world climate claims. In *NeurIPS 2020 Workshop on Tackling Climate Change with Machine Learning*, 2020.

[55] Cody Buntain and Jennifer Golbeck. Automatically identifying fake news in popular twitter threads. In *2017 IEEE international conference on smart cloud (smartCloud)*, pages 208–215. IEEE, 2017.

[56] Joanna M Burkhardt. *Combating fake news in the digital age*, volume 53. American Library Association Chicago, IL, USA, 2017.

[57] Leonardo Bursztyn, Aakaash Rao, Christopher P Roth, and David H Yanagizawa-Drott. Misinformation during a pandemic. Technical report, National Bureau of Economic Research, 2020.

[58] Marius Cristian Buzea, Stefan Trausan-Matu, and Traian Rebedea. Automatic fake news detection for romanian online news. *Information*, 13(3):151, 2022.

[59] Biwei Cao, Lulu Hua, Jiuxin Cao, Jie Gui, Bo Liu, and James Tin-Yau Kwok. No place to hide: Dual deep interaction channel network for fake news detection based on data augmentation. *arXiv preprint arXiv:2303.18049*, 14(8):1–10, 2023.

[60] Emerson F Cardoso, Renato M Silva, and Tiago A Almeida. Towards automatic filtering of fake reviews. *Neurocomputing*, 309:106–116, 2018.

[61] Petar Veličković Guillem Cucurull Arantxa Casanova, Adriana Romero Pietro Lio, and Yoshua Bengio. Graph attention networks. *ICLR. Petar Velickovic Guillem Cucurull Arantxa Casanova Adriana Romero Pietro Liò and Yoshua Bengio*, 2018.

[62] Sonia Castelo, Thais Almeida, Anas Elghafari, Aécio Santos, Kien Pham, Eduardo Nakamura, and Juliana Freire. A topic-agnostic approach for identifying fake news pages. In *Companion proceedings of the 2019 World Wide Web conference*, pages 975–980, 2019.

[63] Arati Chabukswar, P Deepa Shenoy, and KR Venugopal. Fake news detection using optimized deep learning model through effective feature extraction. In *2023 International Conference on Recent Advances in Information Technology for Sustainable Development (ICRAIS)*, pages 118–123. IEEE, 2023.

[64] Tavishee Chauhan and Hemant Palivela. Optimization and improvement of fake news detection using deep learning approaches for societal benefit. *International Journal of Information Management Data Insights*, 1(2):100051, 2021.

[65] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[66] Nitesh V Chawla, Nathalie Japkowicz, and Aleksander Kotcz. Special issue on learning from imbalanced data sets. *ACM SIGKDD explorations newsletter*, 6(1):1–6, 2004.

[67] Mu-Yen Chen, Yi-Wei Lai, and Jiunn-Woei Lian. Using deep learning models to detect fake news about covid-19. *ACM Transactions on Internet Technology*, 2022.

[68] Qingqing Chen. Coronavirus rumors trigger irrational behaviors among chinese netizens. *Retrieved October*, 19:2020, 2020.

[69] Yimin Chen, Niall J Conroy, and Victoria L Rubin. Misleading online content: recognizing clickbait as" false news". In *Proceedings of the 2015 ACM on workshop on multimodal deception detection*, pages 15–19, 2015.

[70] 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.

[71] Dan C. Cireşan, Ueli Meier, Jonathan Masci, Luca M. Gambardella, and Jürgen Schmidhuber. Fake news challenge. http://arxiv.org/abs/1102.0183, 2011. Accessed: 2024-11-11.

[72] Demo Rangarirai Collen, Leslie Kudzai Nyandoro, and Kudakwashe Zvarevashe. Fake news detection using 5l-cnn. In *2022 1st Zimbabwe Conference of Information and Communication Technologies (ZCICT)*, pages 1–7. IEEE, 2022.

[73] America News Community. The onion. https://www.theonion.com/), 2021. Online, Accessed: 2024-11-11.

[74] FourChan Community. 4chan web. https://boards.4chan.org/, 2018. Accessed: 2024-11-13.

[75] Gab Community. Gab. https://gab.com/, 2018. Accessed: 2024-11-13.

[76] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20:273–297, 1995.

[77] S Deepak and Bhadrachalam Chitturi. Deep neural approach to fake-news identification. *Procedia Computer Science*, 167:2236–2243, 2020.

[78] Javier Del Ser, Miren Nekane Bilbao, Ibai Laña, Khan Muhammad, and David Camacho. Efficient fake news detection using bagging ensembles of bidirectional echo state networks. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2022.

[79] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[80] Amaram Divija, Smitha Kiran Kurkuri, Priyanka Telukuntla, and Shailaja Mantha. Fake news classifier. *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, 10:1716–1722, 2022.

[81] Yash Doke, Prajwalita Dongare, Mansi Gaikwad, Mayuri Gaikwad, and Vaibhav Marathe. Deep fake detection through deep learning. *International Journal for Research in Applied Science & Engineering Technology (IJRA)*, 11(5):861–866, 2023.

[82] Manqing Dong, Lina Yao, Xianzhi Wang, Boualem Benatallah, Quan Z Sheng, and Hao Huang. Dual: A deep unified attention model with latent relation representations for fake news detection. In *Web Information Systems Engineering–WISE 2018:*

19th International Conference, Dubai, United Arab Emirates, November 12-15, 2018, Proceedings, Part I 19, pages 199–209. Springer, 2018.

[83] Mohamed K Elhadad, Kin Fun Li, and Fayez Gebali. An ensemble deep learning technique to detect covid-19 misleading information. In *Advances in Networked-Based Information Systems: The 23rd International Conference on Network-Based Information Systems (NBiS-2020) 23*, pages 163–175. Springer, 2021.

[84] Imane Ennejjai, Anass Ariss, Nassim Kharmoum, Wajih Rhalem, Soumia Ziti, and Mostafa Ezziyyani. Artificial intelligence for fake news. In *International Conference on Advanced Intelligent Systems for Sustainable Development*, pages 77–91. Springer, 2022.

[85] Ehab Essa, Karima Omar, and Ali Alqahtani. Fake news detection based on a hybrid bert and lightgbm models. *Complex & Intelligent Systems*, pages 1–12, 2023.

[86] Anindika Riska Intan Fauzy, Erwin Budi Setiawan, et al. Detecting fake news on social media combined with the cnn methods. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 7(2):271–277, 2023.

[87] William Ferreira and Andreas Vlachos. Emergent: a novel data-set for stance classification. In *Proceedings of NAACL-HLT*, pages 1163–1168, 2016.

[88] Thomas Finley and Thorsten Joachims. Supervised clustering with support vector machines. In *Proceedings of the 22nd international conference on Machine learning*, pages 217–224, 2005.

[89] Khaled M Fouad, Sahar F Sabbeh, and Walaa Medhat. Arabic fake news detection using deep learning. *CMC-Comput. Mater. Contin*, 71:3647–3665, 2022.

[90] G. McIntire. Fake or real news Dataset. https://github.com/joolsa/fake_real_news_dataset. Accessed: 2024-11-22.

[91] Huiji Gao and Huan Liu. Data analysis on location-based social networks. *Mobile social networking: an innovative approach*, pages 165–194, 2014.

[92] Wang Gao, Mingyuan Ni, Hongtao Deng, Xun Zhu, Peng Zeng, and Xi Hu. Few-shot fake news detection via prompt-based tuning. *J. Intell. Fuzzy Syst.*, 44:9933–9942, 2023.

[93] Rachit Garg and S Jeevaraj. Effective fake news classifier and its applications to covid-19. In *2021 IEEE Bombay Section Signature Conference (IBSSC)*, pages 1–6. IEEE, 2021.

[94] Jean-Gabriel Gaudreault and Paula Branco. Empirical analysis of performance assessment for imbalanced classification. *Machine Learning*, pages 1–43, 2024.

[95] Jean-Gabriel Gaudreault, Paula Branco, and João Gama. An analysis of performance metrics for imbalanced classification. In *Discovery Science: 24th International Conference, DS 2021, Halifax, NS, Canada, October 11–13, 2021, Proceedings 24*, pages 67–77. Springer, 2021.

[96] Alexander Genkin, David D Lewis, and David Madigan. Large-scale bayesian logistic regression for text categorization. *technometrics*, 49(3):291–304, 2007.

[97] Md Zahin Hossain George, Naimul Hossain, Md Rafiuzzaman Bhuiyan, Abu Kaisar Mohammad Masum, and Sheikh Abujar. Bangla fake news detection based on multichannel combined cnn-lstm. In *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–5. IEEE, 2021.

[98] Kushankur Ghosh, Colin Bellinger, Roberto Corizzo, Paula Branco, Bartosz Krawczyk, and Nathalie Japkowicz. The class imbalance problem in deep learning. *Machine Learning*, pages 1–57, 2022.

[99] Sherry Girgis, Eslam Amer, and Mahmoud Gadallah. Deep learning algorithms for detecting fake news in online text. In *2018 13th international conference on computer engineering and systems (ICCES)*, pages 93–97. IEEE, 2018.

[100] Mohammad Hadi Goldani, Reza Safabakhsh, and Saeedeh Momtazi. Convolutional neural network with margin loss for fake news detection. *Information Processing & Management*, 58(1):102418, 2021.

[101] S Gonwirat, A Choompol, and N Wichapa. A combined deep learning model based on the ideal distance weighting method for fake news detection. *International Journal of Data and Network Science*, 6(2):347–354, 2022.

[102] Nuno Guimarães, Álvaro Figueira, and Luís Torgo. Can fake news detection models maintain the performance through time? a longitudinal evaluation of twitter publications. *Mathematics*, 9(22):2988, 2021.

[103] Gülselin Güler and Sedef GÜNDÜZ. Deep learning based fake news detection on social media. *International Journal of Information Security Science*, 12(2):1–21, 2023.

[104] Richard Gunther, Paul A Beck, and Erik C Nisbet. Fake news did have a significant impact on the vote in the 2016 election: Original full-length version with methodological appendix. *Unpublished manuscript, Ohio State University, Columbus, OH*, 2018.

[105] Bin Guo, Yasan Ding, Lina Yao, Yunji Liang, and Zhiwen Yu. The future of false information detection on social media: New perspectives and trends. *ACM Computing Surveys (CSUR)*, 53(4):1–36, 2020.

[106] Suhaib Kh Hamed, Mohd Juzaiddin Ab Aziz, and Mohd Ridzwan Yaakub. A review of fake news detection models: Highlighting the factors affecting model performance and the prominent techniques used. *International Journal of Advanced Computer Science and Applications*, 14(7), 2023.

[107] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

[108] Xintong Han, Vlad Morariu, Peng IS Larry Davis, et al. Two-stream neural networks for tampered face detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 19–27, 2017.

[109] Xu Han, Weilin Zhao, Ning Ding, Zhiyuan Liu, and Maosong Sun. Ptr: Prompt tuning with rules for text classification. *AI Open*, 3:182–192, 2022.

[110] Maayan Harel and Shie Mannor. Learning from multiple outlooks. *arXiv preprint arXiv:1005.0027*, 2010.

[111] Ehtesham Hashmi, Sule Yildirim Yayilgan, Muhammad Mudassar Yamin, Subhan Ali, and Mohamed Abomhara. Advancing fake news detection: hybrid deep learning with fasttext and explainable ai. *IEEE Access*, 2024.

[112] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pages 1322–1328. IEEE, 2008.

[113] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 770–778, 2016.

[114] Stefan Helmstetter and Heiko Paulheim. Weakly supervised learning for fake news detection on twitter. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 274–277. IEEE, 2018.

[115] Stefan Helmstetter and Heiko Paulheim. Collecting a large scale dataset for classifying fake news tweets using weak supervision. *Future Internet*, 13(5):114, 2021.

[116] Peter Hernon. Disinformation and misinformation through the internet: Findings of an exploratory study. *Government information quarterly*, 12(2):133–139, 1995.

[117] Gabriel Emile Hine, Jeremiah Onaolapo, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Riginos Samaras, Gianluca Stringhini, and Jeremy Blackburn. Kek, cucks, and god emperor trump: A measurement study of 4chan's politically incorrect forum and its effects on the web. In *Eleventh international AAAI conference on web and social media*, 2017.

[118] Chaitra K Hiramath and GC Deshpande. Fake news detection using deep learning techniques. In *2019 1st International Conference on Advances in Information Technology (ICAIT)*, pages 411–415. IEEE, 2019.

[119] Prajwal Hiremath, Srujan Sateesh Kalagi, et al. Analysis of fake news detection using graph neural network (gnn) and deep learning. In *2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS)*, pages 1805–1811. IEEE, 2023.

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

[121] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018.

[122] W. Hu. A multi-modal prompt learning framework for early detection of fake news. *Proceedings of the International Aaai Conference on Web and Social Media*, 18:651–662, 2024.

[123] Binxuan Huang and Kathleen M Carley. Disinformation and misinformation on twitter during the novel coronavirus outbreak. *arXiv preprint arXiv:2006.04278*, 2020.

[124] Lu Huang. Deep learning for fake news detection: Theories and models. In *Proceedings of the 2022 6th International Conference on Electronic Information Technology and Computer Engineering*, pages 1322–1326, 2022.

[125] Yin-Fu Huang and Po-Hong Chen. Fake news detection using an ensemble learning model based on self-adaptive harmony search algorithms. *Expert Systems with Applications*, 159:113584, 2020.

[126] Yinqiu Huang, Min Gao, Jia Wang, and Kai Shu. Dafd: Domain adaptation framework for fake news detection. In *Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part I 28*, pages 305–316. Springer, 2021.

[127] Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana Inkpen. Deep learning for depression detection of Twitter users. In Kate Loveys, Kate Niederhoffer, Emily Prud'hommeaux, Rebecca Resnik, and Philip Resnik, editors, *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 88–97, New Orleans, LA, June 2018. Association for Computational Linguistics.

[128] Vlad-Iulian Ilie, Ciprian-Octavian Truică, Elena-Simona Apostol, and Adrian Paschke. Context-aware misinformation detection: A benchmark of deep learning architectures using word embeddings. *IEEE Access*, 9:162122–162146, 2021.

[129] Sani Muhamad Isa, Gary Nico, and Mikhael Permana. Indobert for indonesian fake news detection. *ICIC Express Letters*, 16(3):289–297, 2022.

[130] Md Rafiqul Islam, Muhammad Ashad Kabir, Ashir Ahmed, Abu Raihan M Kamal, Hua Wang, and Anwaar Ulhaq. Depression detection from social network data using machine learning techniques. *Health information science and systems*, 6:1–12, 2018.

[131] Klaudia Ivancová, Martin Sarnovskỳ, and Viera Maslej-Krcšňáková. Fake news detection in slovak language using deep learning techniques. In *2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, pages 000255–000260. IEEE, 2021.

[132] Vidit Jain, Rohit Kumar Kaliyar, Anurag Goswami, Pratik Narang, and Yashvardhan Sharma. Aenet: an attention-enabled neural architecture for fake news detection using contextual features. *Neural Computing and Applications*, 34(1):771–782, 2022.

[133] Arunima Jaiswal, Himika Verma, and Nitin Sachdeva. Swarm optimized fake news detection on social-media textual content using deep learning. In *2023 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)*, pages 1–8. IEEE, 2023.

[134] Samireh Jalali and Claes Wohlin. Systematic literature studies: database searches vs. backward snowballing. In *Proceedings of the ACM-IEEE international symposium on Empirical software engineering and measurement*, pages 29–38, 2012.

[135] Zainab A Jawad and Ahmed J Obaid. Combination of convolution neural networks and deep neural networks for fake news detection. *arXiv preprint arXiv:2210.08331*, 2022.

[136] Nirosh Jayakody, Azeem Mohammad, and Malka N Halgamuge. Fake news detection using a decentralized deep learning model and federated learning. In *IECON 2022–48th Annual Conference of the IEEE Industrial Electronics Society*, pages 1–6. IEEE, 2022.

[137] 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.

[138] Yu Ji. Fake news detection based on a bi-directional lstm with cnn. In *Computing and Data Science: Third International Conference, CONF-CDS 2021, Virtual Event, August 12-17, 2021, Proceedings*, pages 36–44. Springer, 2022.

[139] Tao Jiang, Jian Ping Li, Amin Ul Haq, and Abdus Saboor. Fake news detection using deep recurrent neural networks. In *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pages 205–208. IEEE, 2020.

[140] Nitin Jindal and Bing Liu. Opinion spam and analysis. In *Proceedings of the 2008 international conference on web search and data mining*, pages 219–230, 2008.

[141] George H John and Pat Langley. Estimating continuous distributions in bayesian classifiers. *arXiv preprint arXiv:1302.4964*, 2013.

[142] Jruvika. Fake news detection. https://www.kaggle.com/jruvika/fake-news-detection/version/1, Accessed: 2024-12-12.

[143] Christoph Käding, Erik Rodner, Alexander Freytag, and Joachim Denzler. Fine-tuning deep neural networks in continuous learning scenarios. In *Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part III 13*, pages 588–605. Springer, 2017.

[144] Kaggle Community Prediction Team. Fake News Dataset. https://www.kaggle.com/competitions/fake-news/data, Accessed: 2024-11-18.

[145] Rohit Kumar Kaliyar. Fake news detection using a deep neural network. In *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, pages 1–7. IEEE, 2018.

[146] Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. Fakebert: Fake news detection in social media with a bert-based deep learning approach. *Multimedia tools and applications*, 80(8):11765–11788, 2021.

[147] Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. A hybrid model for effective fake news detection with a novel covid-19 dataset. In *ICAART (2)*, pages 1066–1072, 2021.

[148] Rohit Kumar Kaliyar, Anurag Goswami, Pratik Narang, and Soumendu Sinha. Fndnet–a deep convolutional neural network for fake news detection. *Cognitive Systems Research*, 61:32–44, 2020.

[149] Rohit Kumar Kaliyar, Pawan Kumar, Manish Kumar, Meenal Narkhede, Sreyas Namboodiri, and Sneha Mishra. Deepnet: an efficient neural network for fake news detection using news-user engagements. In *2020 5th International Conference on Computing, Communication and Security (ICCCS)*, pages 1–6. IEEE, 2020.

[150] Rohit Kumar Kaliyar, Arjun Mohnot, R Raghhul, VK Prathyushaa, Anurag Goswami, Navya Singh, and Palavi Dash. Multideepfake: Improving fake news detection with a deep convolutional neural network using a multimodal dataset. In *Advanced Computing: 10th International Conference, IACC 2020, Panaji, Goa, India, December 5–6, 2020, Revised Selected Papers, Part I 10*, pages 267–279. Springer, 2021.

[151] Rohit Kumar Kaliyar, Rajbeer Singh, Sai N Laya, M Sujay Sudharshan, Anurag Goswami, and Deepak Garg. Rumeval2020-an effective approach for rumour detection with a deep hybrid c-lstm model. In *Advanced Computing: 10th International*

Conference, IACC 2020, Panaji, Goa, India, December 5–6, 2020, Revised Selected Papers, Part I 10, pages 300–312. Springer, 2021.

[152] Cannannore Nidhi Kamath, Syed Saqib Bukhari, and Andreas Dengel. Comparative study between traditional machine learning and deep learning approaches for text classification. In *Proceedings of the ACM Symposium on Document Engineering 2018*, pages 1–11, 2018.

[153] Uday Kamath, John Liu, and James Whitaker. *Deep learning for NLP and speech recognition*, volume 84. Springer, 2019.

[154] N Kanagavalli, S Baghavathi Priya, and D Jeyakumar. Design of hyperparameter tuned deep learning based automated fake news detection in social networking data. In *2022 6th International Conference on Computing Methodologies and Communication (ICCMC)*, pages 958–963. IEEE, 2022.

[155] M Kanchana, Vel Murugesh Kumar, TP Anish, and Pv Gopirajan. Deep fake bert: Efficient online fake news detection system. In *2023 International Conference on Networking and Communications (ICNWC)*, pages 1–6. IEEE, 2023.

[156] Venkatachalam Kandasamy, Štěpán Hubálovskỳ, and Pavel Trojovskỳ. Deep fake detection using a sparse auto encoder with a graph capsule dual graph cnn. *PeerJ Computer Science*, 8:e953, 2022.

[157] Piyush Katariya, Vedika Gupta, Rohan Arora, Adarsh Kumar, Shreya Dhingra, Qin Xin, and Jude Hemanth. A deep neural network-based approach for fake news detection in regional language. *International Journal of Web Information Systems*, 18(5/6):286–309, 2022.

[158] Shingo Kato, Linshuo Yang, and Daisuke Ikeda. Domain bias in fake news datasets consisting of fake and real news pairs. In *2022 12th International Congress on Advanced Applied Informatics (IIAI-AAI)*, pages 101–106, 2022.

[159] Shingo Kato, Linshuo Yang, and Daisuke Ikeda. Domain bias in fake news datasets consisting of fake and real news pairs. In *2022 12th International Congress on Advanced Applied Informatics (IIAI-AAI)*, pages 101–106. IEEE, 2022.

[160] Staffs Keele et al. Guidelines for performing systematic literature reviews in software engineering. Technical report, Technical report, ver. 2.3 ebse technical report. ebse, 2007.

[161] Ashfia Jannat Keya, Shahid Afridi, Afroza Siddique Maria, Snaha Sadhu Pinki, Joy Ghosh, and MF Mridha. Fake news detection based on deep learning. In *2021 International Conference on Science & Contemporary Technologies (ICSCT)*, pages 1–6. IEEE, 2021.

[162] Ashfia Jannat Keya, Md Anwar Hussen Wadud, MF Mridha, Mohammed Alatiyyah, and Md Abdul Hamid. Augfake-bert: handling imbalance through augmentation of fake news using bert to enhance the performance of fake news classification. *Applied Sciences*, 12(17):8398, 2022.

[163] Razib Hayat Khan, Asm Shihavuddin, MM Mahbubul Syeed, Rakib Ul Haque, and Mohammad Faisal Uddin. Improved fake news detection method based on deep learning and comparative analysis with other machine learning approaches. In *2022 International Conference on Engineering and Emerging Technologies (ICEET)*, pages 1–6. IEEE, 2022.

[164] Asmaa Khoudi, Nessrine Yahiaoui, and Feriel Rebahi. Detect misinformation of covid-19 using deep learning: A comparative study based on word embedding. In *2023 1st International Conference on Advanced Innovations in Smart Cities (ICAISC)*, pages 1–5. IEEE, 2023.

[165] Raghad Khweiled, Mahmoud Jazzar, and Derar Eleyan. Cybercrimes during covid-19 pandemic. *International Journal of Information Engineering & Electronic Business*, 13(2), 2021.

[166] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[167] Azka Kishwar and Adeel Zafar. Fake news detection on pakistani news using machine learning and deep learning. *Expert Systems with Applications*, 211:118558, 2023.

[168] A Kitanovski, M Toshevska, and G Mirceva. Distilbert and roberta models for identification of fake news. In *2023 46th MIPRO ICT and Electronics Convention (MIPRO)*, pages 1102–1106. IEEE, 2023.

[169] Piotr Kłosowski. Deep learning for natural language processing and language modelling. In *2018 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, pages 223–228. IEEE, 2018.

[170] Pakindessama M Konkobo, Rui Zhang, Siyuan Huang, Toussida T Minoungou, Jose A Ouedraogo, and Lin Li. A deep learning model for early detection of fake

news on social media. In *2020 7th International Conference on Behavioural and Social Computing (BESC)*, pages 1–6. IEEE, 2020.

[171] Rafał Kozik, Sebastian Kula, Michał Choraś, and Michał Woźniak. Technical solution to counter potential crime: Text analysis to detect fake news and disinformation. *Journal of Computational Science*, 60:101576, 2022.

[172] Viera Maslej Krešňáková, Martin Sarnovskỳ, and Peter Butka. Deep learning methods for fake news detection. In *2019 IEEE 19th International Symposium on Computational Intelligence and Informatics and 7th IEEE International Conference on Recent Achievements in Mechatronics, Automation, Computer Sciences and Robotics (CINTI-MACRo)*, pages 000143–000148. IEEE, 2019.

[173] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2012.

[174] Brian Kulis, Kate Saenko, and Trevor Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *CVPR 2011*, pages 1785–1792. IEEE, 2011.

[175] Chaitanya Kulkarni, P Monika, S Shruthi, MS Deepak Bharadwaj, and D Uday. Covid-19 fake news detection using glove and bi-lstm. In *Proceedings of Second International Conference on Sustainable Expert Systems: ICSES 2021*, pages 43–56. Springer, 2022.

[176] Abhinav Kumar, Jyoti Prakash Singh, and Amit Kumar Singh. Covid-19 fake news detection using ensemble-based deep learning model. *IT Professional*, 24(2):32–37, 2022.

[177] Bharani Kumar. Bert variants and their differences. Technical report, 360DigiTMG, 2023.

[178] K Kranthi Kumar, S Hanumantha Rao, G Srikar, and M Bharat Chandra. A novel approach for detection of fake news using long short term memory (lstm). *International Journal*, 10(5), 2021.

[179] Sanjay Kumar, Akshi Kumar, Abhishek Mallik, and Rishi Ranjan Singh. Optnetfake: Fake news detection in socio-cyber platforms using grasshopper optimization and deep neural network. *IEEE Transactions on Computational Social Systems*, 2023.

[180] Srijan Kumar and Neil Shah. False information on web and social media: A survey. *arXiv preprint arXiv:1804.08559*, 2018.

[181] Srijan Kumar, Robert West, and Jure Leskovec. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proceedings of the 25th international conference on World Wide Web*, pages 591–602, 2016.

[182] Juhi Kumari, Raman Choudhary, Swadha Kumari, and Gopal Krishna. A deep learning based approach for classification of news as real or fake. In *Data Science and Security: Proceedings of IDSCS 2021*, pages 239–246. Springer, 2021.

[183] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

[184] Dong-Ho Lee, Yu-Ri Kim, Hyeong-Jun Kim, Seung-Myun Park, and Yu-Jun Yang. Fake news detection using deep learning. *Journal of Information Processing Systems*, 15(5):1119–1130, 2019.

[185] Kyumin Lee, James Caverlee, and Steve Webb. Uncovering social spammers: social honeypots+ machine learning. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 435–442, 2010.

[186] Qian Li, Qingyuan Hu, Youshui Lu, Yue Yang, and Jingxian Cheng. Multi-level word features based on cnn for fake news detection in cultural communication. *Personal and Ubiquitous Computing*, 24:259–272, 2020.

[187] Xinlong Li, Xingyu Fu, Guangluan Xu, Yang Yang, Jiuniu Wang, Li Jin, Qing Liu, and Tianyuan Xiang. Enhancing bert representation with context-aware embedding for aspect-based sentiment analysis. *IEEE Access*, 8:46868–46876, 2020.

[188] Jinshuo Liu, Chenyang Wang, Chenxi Li, Ningxi Li, Juan Deng, and Jeff Z Pan. Dtn: Deep triple network for topic specific fake news detection. *Journal of Web Semantics*, 70:100646, 2021.

[189] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.

[190] 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. *arXiv preprint arXiv:1907.11692*, 2019.

[191] Rushi Longadge and Snehalata Dongre. Class imbalance problem in data mining review. *arXiv preprint arXiv:1305.1707*, 2013.

[192] Yandan Lu and Hongmei Ye. Detection method of fake news spread in social network based on deep learning. In *International Conference on Advanced Hybrid Information Processing*, pages 473–488. Springer, 2022.

[193] Cristian Lumezanu, Nick Feamster, and Hans Klein. # bias: Measuring the tweeting behavior of propagandists. In *Proceedings of the International AAAI Conference on Web and social media*, volume 6, pages 210–217, 2012.

[194] Yangyang Luo, Yuliang Shi, and ShaoShuai Li. Social media fake news detection algorithm based on multiple feature groups. In *2023 IEEE 3rd International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA)*, volume 3, pages 91–95. IEEE, 2023.

[195] Ben Lutkevich. Bert language model. Technical report, Technical Features Write, 2020.

[196] Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI 2016)*, pages 3818–3824. AAAI Press, 2016.

[197] Mirmorsal Madani, Homayun Motameni, and Hosein Mohamadi. Fake news detection using deep learning integrating feature extraction, natural language processing, and statistical descriptors. *Security and Privacy*, 5(6):e264, 2022.

[198] Mirmorsal Madani, Homayun Motameni, and Reza Roshani. Fake news detection using feature extraction, natural language processing, curriculum learning, and deep learning. *International Journal of Information Technology & Decision Making*, pages 1–36, 2023.

[199] Govind Singh Mahara and Sharad Gangele. Fake news detection: A rnn-lstm, bi-lstm based deep learning approach. In *2022 IEEE 1st International Conference on Data, Decision and Systems (ICDDS)*, pages 01–06. IEEE, 2022.

[200] R Mahesh, Bonam Poornika, Nimma Sharaschandrika, Survi Deeksha Goud, and Padirla Uday Kumar. Identification of fake news using deep learning architecture. In *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)*, pages 1246–1253. IEEE, 2021.

[201] Fahim Belal Mahmud, Mahi Md Sadek Rayhan, Mahdi Hasan Shuvo, Islam Sadia, and Md Kishor Morol. A comparative analysis of graph neural networks and commonly used machine learning algorithms on fake news detection. In *2022 7th International Conference on Data Science and Machine Learning Applications (CDMA)*, pages 97–102. IEEE, 2022.

[202] Bhaskar Majumdar, Md RafiuzzamanBhuiyan, Md Arid Hasan, Md Sanzidul Islam, and Sheak Rashed Haider Noori. Multi class fake news detection using lstm approach. In *2021 10th International Conference on System Modeling & Advancement in Research Trends (SMART)*, pages 75–79. IEEE, 2021.

[203] Ruchika Malhotra, Anushree Mahur, et al. Covid-19 fake news detection system. In *2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pages 428–433. IEEE, 2022.

[204] Suhail Malik, Ashish Kumar Chakraverti, and Ali Imam Abidi. Enhancing fake news detection using classification algorithms and deep learning. In *2023 10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, volume 10, pages 780–787. IEEE, 2023.

[205] Chandrakant Mallick, Sarojananda Mishra, and Manas Ranjan Senapati. A cooperative deep learning model for fake news detection in online social networks. *Journal of Ambient Intelligence and Humanized Computing*, 14(4):4451–4460, 2023.

[206] G Mareeswari and Erana Veerappa Dinesh. Deep neural networks based detection and analysis of fake tweets. In *2023 4th International Conference on Signal Processing and Communication (ICSPC)*, pages 56–61. IEEE, 2023.

[207] Andy Marlatt. Records suggest 2020 election conspiracy involved 80m people. https://www.satirewire.com/claim-anti-trump-conspiracy-involved-80-million-people/, 2020. Online, Accessed: 2024-12-10.

[208] Elio Masciari, Vincenzo Moscato, Antonio Picariello, and Giancarlo Sperlí. Detecting fake news by image analysis. In *Proceedings of the 24th symposium on international database engineering & Applications*, pages 1–5, 2020.

[209] Mohammad Masum and Hossain Shahriar. A transfer learning with deep neural network approach for network intrusion detection. *International journal of intellligent computing research*, 12(1), 2021.

[210] Anand Matheven and Burra Venkata Durga Kumar. Fake news detection using deep learning and natural language processing. In *2022 9th International Conference on Soft Computing & Machine Intelligence (ISCMI)*, pages 11–14. IEEE, 2022.

[211] Samarth Mengji, Saransh Ambarte, Sai Viswa Teja Arumilli, Shankar Mhamane, and Rashmi Rane. Fake news detection using rnn-lstm. *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, 9:1731–1737, 2021.

[212] Microsoft. Microsoft machine learning. https://docs.microsoft.com/en-us/machine-learning/, 2021. Accessed: 2024-11-11.

[213] Microsoft Corporation. Microsoftml: A package for machine learning with r. Version 1.5.0, 2018.

[214] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[215] Fahad Mira. Deep learning technique for recognition of deep fake videos. In *2023 IEEE IAS Global Conference on Emerging Technologies (GlobConET)*, pages 1–4. IEEE, 2023.

[216] Rishabh Misra. Politifact Fact Check Dataset. https://www.kaggle.com/datasets/rmisra/politifact-fact-check-dataset, 2022. Accessed: 2024-12-11.

[217] Asutosh Mohapatra, Nithin Thota, and P Prakasam. Fake news detection and classification using hybrid bilstm and self-attention model. *Multimedia Tools and Applications*, 81(13):18503–18519, 2022.

[218] Patricia Moravec, Antino Kim, and Alan Dennis. Flagging fake news: System 1 vs. system 2. In *Thirty ninth International Conference on Information Systems, San Francisco*, 2018.

[219] Ahmadreza Mosallanezhad, Mansooreh Karami, Kai Shu, Michelle V Mancenido, and Huan Liu. Domain adaptive fake news detection via reinforcement learning. In *Proceedings of the ACM Web Conference 2022*, pages 3632–3640, 2022.

[220] Despoina Mouratidis, Maria Nefeli Nikiforos, and Katia Lida Kermanidis. Deep learning for fake news detection in a pairwise textual input schema. *Computation*, 9(2):20, 2021.

[221] Jane Wambui Waweru Muigai. Understanding fake news. *International Journal of Scientific and Research Publications*, 9(1):29–38, 2019.

[222] Alhassan Mumuni and Fuseini Mumuni. Data augmentation: A comprehensive survey of modern approaches. *Array*, 16:100258, 2022.

[223] Rekha Muppidi and V. Biksham. Deep convolutional neural network for fake news detection over online social networks. *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, 6(4):1–10, 2022.

[224] Jaechang Nam and Sunghun Kim. Heterogeneous defect prediction. In *Proceedings of the 2015 10th joint meeting on foundations of software engineering*, pages 508–519, 2015.

[225] Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. Mdfend: Multi-domain fake news detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3343–3347, 2021.

[226] Qiong Nan, Danding Wang, Yongchun Zhu, Qiang Sheng, Yuhui Shi, Juan Cao, and Jintao Li. Improving fake news detection of influential domain via domain- and instance-level transfer. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2834–2848, October 2022.

[227] C Nandhakumar, C Kowsika, R Reshema, and L Sandhiya. Fake news detection using machine learning and deep learning classifiers. In *International Conference on Information and Communication Technology for Intelligent Systems*, pages 165–175. Springer, 2023.

[228] Udit Narayan, Ajad Kumar, and Kuldeep Kumar. Fake news detection using hybrid of deep neural network and stacked lstm. In *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, pages 385–390. IEEE, 2021.

[229] M Badri Narayanan, Arun Kumar Ramesh, KS Gayathri, and A Shahina. Fake news detection using a deep learning transformer based encoder-decoder architecture. *Journal of Intelligent & Fuzzy Systems*, pages 1–13, 2023.

[230] Jamal Abdul Nasir, Osama Subhani Khan, and Iraklis Varlamis. Fake news detection: A hybrid cnn-rnn based deep learning approach. *International Journal of Information Management Data Insights*, 1(1):100007, 2021.

[231] Ali Bou Nassif, Ashraf Elnagar, Omar Elgendy, and Yaman Afadar. Arabic fake news detection based on deep contextualized embedding models. *Neural Computing and Applications*, 34(18):16019–16032, 2022.

[232] Okuhle Ngada and Bertram Haskins. Investigating fake news detection by means of deep learning on a limited data set. In *2022 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, pages 1–6. IEEE, 2022.

[233] Minal Nirav Shah and Amit Ganatra. A systematic literature review and existing challenges toward fake news detection models. *Social Network Analysis and Mining*, 12(1):168, 2022.

[234] Emerson Yoshiaki Okano, Zebin Liu, Donghong Ji, and Evandro Eduardo Seron Ruiz. Fake news detection on fake. br using hierarchical attention networks. In *Computational Processing of the Portuguese Language: 14th International Conference, PROPOR 2020, Evora, Portugal, March 2–4, 2020, Proceedings 14*, pages 143–152. Springer, 2020.

[235] OpenAI. Gpt-3: Language models are few-shot learners. https://openai.com/research/gpt-3, 2024. Accessed: 2024-11-29.

[236] Timothy O'shea and Jakob Hoydis. An introduction to deep learning for the physical layer. *IEEE Transactions on Cognitive Communications and Networking*, 3(4):563–575, 2017.

[237] Balasubramanian Palani and Sivasankar Elango. Ctrl-fnd: content-based transfer learning approach for fake news detection on social media. *International Journal of System Assurance Engineering and Management*, 14(3):903–918, 2023.

[238] Balasubramanian Palani, Sivasankar Elango, and Vignesh Viswanathan K. Cb-fake: A multimodal deep learning framework for automatic fake news detection using capsule neural network and bert. *Multimedia Tools and Applications*, 81(4):5587–5620, 2022.

[239] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

[240] Kanaga Devan K Parimala and Anandha GS Mala. An optimal detection of fake news from twitter data using dual-stage deep capsule autoencoder. *Journal of Experimental & Theoretical Artificial Intelligence*, 36(2):287–313, 2024.

[241] Bohdan M Pavlyshenko. Analysis of disinformation and fake news detection using fine-tuned large language model. *arXiv preprint arXiv:2309.04704*, 2023.

[242] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[243] Warren A Peterson and Noel P Gist. Rumor and public opinion. *American Journal of Sociology*, 57(2):159–167, 1951.

[244] Ihor Pilkevych, Dmytro Fedorchuk, Olena Naumchak, and Mykola Romanchuk. Fake news detection in the framework of decision-making system through graph neural network. In *2021 IEEE 4th International Conference on Advanced Information and Communication Technologies (AICT)*, pages 153–157. IEEE, 2021.

[245] Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. Declare: Debunking fake news and false claims using evidence-aware deep learning. *arXiv preprint arXiv:1809.06416*, 2018.

[246] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*, 2017.

[247] Om Prakash and Rajeev Kumar. Fake news detection in social networks using attention mechanism. In *Proceedings of the International Conference on Cognitive and Intelligent Computing: ICCIC 2021, Volume 2*, pages 453–462. Springer, 2023.

[248] Peter Prettenhofer and Benno Stein. Cross-language text classification using structural correspondence learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 1118–1127, 2010.

[249] Anu Priya and Abhinav Kumar. Deep ensemble approach for covid-19 fake news detection from social media. In *2021 8th International Conference on Signal Processing and Integrated Networks (SPIN)*, pages 396–401. IEEE, 2021.

[250] Ethar Qawasmeh, Mais Tawalbeh, and Malak Abdullah. Automatic identification of fake news using deep learning. In *2019 Sixth international conference on social networks analysis, Management and Security (SNAMS)*, pages 383–388. IEEE, 2019.

[251] Momina Qazi, Muhammad US Khan, and Mazhar Ali. Detection of fake news using transformer model. In *2020 3rd international conference on computing, mathematics and engineering technologies (iCoMET)*, pages 1–6. IEEE, 2020.

[252] Awf Qdroo and Muhammet Baykara. A new approach to detect fake news related to covid-19 pandemic using deep neural network. *Journal of Applied Science and Technology Trends*, 3(02):27–34, 2022.

[253] J Ross Quinlan. *C4. 5: programs for machine learning*. Elsevier, 2014.

[254] Nishant Rai, Deepika Kumar, Naman Kaushik, Chandan Raj, and Ahad Ali. Fake news classification using transformer based enhanced lstm and bert. *International Journal of Cognitive Computing in Engineering*, 3:98–105, 2022.

[255] Chahat Raj and Priyanka Meel. Convnet frameworks for multi-modal fake news detection. *Applied Intelligence*, pages 1–17, 2021.

[256] RR Rajalaxmi, LV Narasimha Prasad, B Janakiramaiah, CS Pavankumar, N Neelima, and VE Sathishkumar. Optimizing hyperparameters and performance analysis of lstm model in detecting fake news on social media. *Transactions on Asian and Low-Resource Language Information Processing*, 2022.

[257] SP Ramya and R Eswari. Attention-based deep learning models for detection of fake news in social networks. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 15(4):1–25, 2021.

[258] Vivek Ranjan and Pragati Agrawal. Fake news detection: Ga-transformer and ig-transformer based approach. In *2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pages 487–493. IEEE, 2022.

[259] Kenneth Rapoza. Can fake news impact the stock market? https://www.forbes.com/sites/kenrapoza/2017/02/26/can-fake-news-impact-the-stock-market, 2017. Accessed: 2024-12-11.

[260] Shubhangi Rastogi and Divya Bansal. A review on fake news detection 3t's: Typology, time of detection, taxonomies. *International Journal of Information Security*, 22(1):177–212, 2023.

[261] Jay Rautela, VV Ramalingam, and Hamaad Makhdoomi. Fake news detection through deep learning techniques. *International Journal of Health Sciences*, 6(5):2107—2111, 2022.

[262] Shaina Raza and Chen Ding. Fake news detection based on news content and social contexts: a transformer-based approach. *International Journal of Data Science and Analytics*, 13(4):335–362, 2022.

[263] Yuxiang Ren, Bo Wang, Jiawei Zhang, and Yi Chang. Adversarial active learning based heterogeneous graph neural network for fake news detection. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 452–461. IEEE, 2020.

[264] Junaid Ali Reshi and Rashid Ali. Online fake news detection using pre-trained embeddings. In *2022 5th International Conference on Multimedia, Signal Processing and Communication Technologies (IMPACT)*, pages 1–5. IEEE, 2022.

[265] Lalu M. Riza Rizky and S. Suyanto. Improving stance-based fake news detection using bert model with synonym replacement and random swap data augmentation technique. *2021 IEEE 7th Information Technology International Seminar (ITIS)*, pages 1–6, 2021.

[266] Steven A Rosenberg, Batya Elbaum, Cordelia Robinson Rosenberg, Yvonne Kellar-Guenther, and Beth M McManus. From flawed design to misleading information: The us department of education's early intervention child outcomes evaluation. *American Journal of Evaluation*, 39(3):350–363, 2018.

[267] Victoria L Rubin, Yimin Chen, and Nadia K Conroy. Deception detection for news: three types of fakes. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4, 2015.

[268] Fariba Sadeghi, Amir Jalaly Bidgoly, and Hossein Amirkhani. Fake news detection on social media using a natural language inference approach. *Multimedia Tools and Applications*, 81(23):33801–33821, 2022.

[269] Ammar Saeed and Eesa Al Solami. Fake news detection using machine learning and deep learning methods. *CMC-COMPUTERS MATERIALS & CONTINUA*, 77(2):2079–2096, 2023.

[270] Julio A. Saenz, Sindhu Reddy Kalathur Gopal, and Diksha Shukla. Covid-19 Fake News Infodemic Research Dataset (CoVID19-FNIR Dataset). https://dx.doi.org/10.21227/b5bt-5244, 2021. Accessed: 2024-11-28.

[271] Somya Ranjan Sahoo and Brij B Gupta. Multiple features based approach for automatic fake news detection on social networks using deep learning. *Applied Soft Computing*, 100:106983, 2021.

[272] Kajal Saini and Ruchi Jain. A hybrid lstm-bert and glove-based deep learning approach for the detection of fake news. In *2023 3rd International Conference on Smart Data Intelligence (ICSMDI)*, pages 400–406. IEEE, 2023.

[273] Ilhem Salah, Khaled Jouini, and Ouajdi Korbaa. On the use of text augmentation for stance and fake news detection. *Journal of Information and Telecommunication*, pages 1–17, 2023.

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

[275] Fatima K Abu Salem, Roaa Al Feel, Shady Elbassuoni, Mohamad Jaber, and May Farah. Fa-kes: A fake news dataset around the syrian war. In *Proceedings of the international AAAI conference on web and social media*, volume 13, pages 573–582, 2019.

[276] Mohammadreza Samadi and Saeedeh Momtazi. Fake news detection: deep semantic representation with enhanced feature engineering. *International Journal of Data Science and Analytics*, pages 1–12, 2023.

[277] Mohammadreza Samadi, Maryam Mousavian, and Saeedeh Momtazi. Persian fake news detection: Neural representation and classification at word and text levels. *Transactions on Asian and Low-Resource Language Information Processing*, 21(1):1–11, 2021.

[278] K Sangeeta, Thota Gowri Priya, Susmitha Patro, Allu Goutham, Kinjarapu Jeevana Jyothi, and Muvvala Anil Kumar. Fake news detection using feature selection and deep learning. *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, 10:3878–3887, 2022.

[279] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2020.

[280] Gaurav Sarin and Pradeep Kumar. Convgrutext: a deep learning method for fake text detection on online social media. *Pasific Asia Conference on Information Systems, PACIS 2020 Proceedings*, 2020.

[281] Martin Sarnovskỳ, Viera Maslej-Krešňáková, and Klaudia Ivancová. Fake news detection related to the covid-19 in slovak language using deep learning methods. *Acta Polytechnica Hungarica*, 19(2):43–57, 2022.

[282] I Kadek Sastrawan, IPA Bayupati, and Dewa Made Sri Arsa. Detection of fake news using deep learning cnn–rnn based methods. *ICT Express*, 8(3):396–408, 2022.

[283] MSVPJ SATHVIK, Mukesh Kumar Mishra, and Sibasankar Padhy. Fake news detection by fine tuning of bidirectional encoder representations from transformers. *IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS, NO. 0*, 00:1–8, 2023.

[284] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.

[285] Tal Schuster, Roei Schuster, Darsh J Shah, and Regina Barzilay. The limitations of stylometry for detecting machine-generated fake news. *Computational Linguistics*, 46(2):499–510, 2020.

[286] Mina Schütz, Alexander Schindler, Melanie Siegel, and Kawa Nazemi. Automatic fake news detection with pre-trained transformer models. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10-15, 2021, Proceedings, Part VII*, pages 627–641. Springer, 2021.

[287] Ahmed Sedik, Amr A Abohany, Karam M Sallam, Kumudu Munasinghe, and T Medhat. Deep fake news detection system based on concatenated and recurrent modalities. *Expert Systems with Applications*, 208:117953, 2022.

[288] Chris Seiffert, Taghi M Khoshgoftaar, Jason Van Hulse, and Amri Napolitano. A comparative study of data sampling and cost sensitive learning. In *2008 IEEE international conference on data mining workshops*, pages 46–52. IEEE, 2008.

[289] Youngkyung Seo and Chang-Sung Jeong. Fagon: Fake news detection model using grammatic transformation on neural network. In *2018 Thirteenth International Conference on Knowledge, Information and Creativity Support Systems (KICSS)*, pages 1–5. IEEE, 2018.

[290] Shaban Shabani and Maria Sokhn. Hybrid machine-crowd approach for fake news detection. In *2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC)*, pages 299–306. IEEE, 2018.

[291] Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu. Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(3):1–42, 2019.

[292] Pushkar Sharma and Rakesh Sahu. Fake news detection using deep learning based approach. In *2023 International Conference on Circuit Power and Computing Technologies (ICCPCT)*, pages 651–656. IEEE, 2023.

[293] Srishti Sharma, Mala Saraswat, and Anil Kumar Dubey. Fake news detection using deep learning. In *Knowledge Graphs and Semantic Web: Third Iberoamerican Conference and Second Indo-American Conference, KGSWC 2021, Kingsville, Texas, USA, November 22–24, 2021, Proceedings 3*, pages 249–259. Springer, 2021.

[294] Sumit Sharma, Rajan Kumar Dudeja, Gagangeet Singh Aujla, Rasmeet Singh Bali, and Neeraj Kumar. Detras: deep learning-based healthcare framework for iot-based assistance of alzheimer patients. *Neural Computing and Applications*, pages 1–13, 2020.

[295] Md Rakibul Hasan Shezan, Mohammed Nasif Zawad, Yeasir Arafat Shahed, and Shamim Ripon. Bangla fake news detection using hybrid deep learning models. In *Applied Informatics for Industry 4.0*, pages 46–60. Chapman and Hall/CRC, 2023.

[296] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

[297] Priya Shrivastava and Dilip Kumar Sharma. Fake content identification using pretrained glove-embedding. In *2021 5th International Conference on Information Systems and Computer Networks (ISCON)*, pages 1–6. IEEE, 2021.

[298] Priya Shrivastava and Dilip Kumar Sharma. Covid-19 fake news detection using pretuned bert-based transfer learning models. In *2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART)*, pages 64–68. IEEE, 2022.

[299] Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 395–405, 2019.

[300] 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(3):171–188, 2020.

[301] Elena Shushkevich, John Cardiff, and Anna Boldyreva. Detection of truthful, semi-truthful, false and other news with arbitrary topics using bert-based models. In *2023 33rd Conference of Open Innovations Association (FRUCT)*, pages 250–256. IEEE, 2023.

[302] Amila Silva, Ling Luo, Shanika Karunasekera, and Christopher Leckie. Embracing domain differences in fake news: Cross-domain fake news detection using multi-modal data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 557–565, 2021.

[303] Simon Fraser University. Snopes Dataset. http://fakenews.research.sfu.ca/. Accessed: 2024-11-28.

[304] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[305] Yuvraj Singh and Pawan Singh. Fake news detection using lstm in tensorflow and deep learning. *Journal of Applied Science and Education (JASE)*, 3(2):1–14, 2023.

[306] Jayesh Soni. An efficient lstm model for fake news detection. *Computer Science & Engineering: An International Journal (CSEIJ)*, 12(2), 2022.

[307] Pokkuluri Kiran Sree, Gurujukota Ramesh Babu, PBV Raja Rao, Phaneendra Varma Chintalapati, M Prasad, et al. Fake news detection using cellular automata based deep learning. In *2023 3rd International Conference on Computing and Information Technology (ICCIT)*, pages 167–171. IEEE, 2023.

[308] Statistics Canada. Concerns with misinformation online. https://www150.statcan.gc.ca/n1/daily-quotidien/231220/dq231220b-eng.htm, 2023. Accessed: 2024-12-04.

[309] Xuanyu Su, Yansong Li, Paula Branco, and Diana Inkpen. Ssl-gan-roberta: A robust semi-supervised model for detecting anti-asian covid-19 hate speech on social media. *Natural Language Engineering*, pages 1–20, 2023.

[310] Paliwal Mohan Subhash, Deepa Gupta, Suja Palaniswamy, and Manju Venugopalan. Fake news detection using deep learning and transformer-based model. In *2023 14th*

International Conference on Computing Communication and Networking Technologies (ICCCNT), pages 1–6. IEEE, 2023.

[311] Raquiba Sultana and Tetsuro Nishino. Fake news detection system: An implementation of bert and boosting algorithm. *Proceedings of 38th International Confer*, 91:124–137, 2023.

[312] Banik Sumit. COVID Fake News Dataset. https://doi.org/10.5281/zenodo.4282522, November 2020. Accessed: 2024-12-11.

[313] Mengtao Sun, Ibrahim A Hameed, Hao Wang, and Mark Pasquine. Perceiving the narrative style for fake news detection using deep learning. In *IEEE 23rd Int Conf on High Performance Computing & Communications; 7th Int Conf on Data Science & Systems; 19th Int Conf on Smart City; 7th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys)*, pages 1195–1202. IEEE, 2021.

[314] Shraddha Suratkar and Faruk Kazi. Deep fake video detection using transfer learning approach. *Arabian Journal for Science and Engineering*, 48(8):9727–9737, 2023.

[315] S Suryavardan, Shreyash Mishra, Megha Chakraborty, Parth Patwa, Anku Rani, Aman Chadha, Aishwarya Reganti, Amitava Das, Amit Sheth, Manoj Chinnakotla, et al. Findings of factify 2: multimodal fake news detection. *arXiv preprint arXiv:2307.10475*, 2023.

[316] Mateusz Szczepański, Marek Pawlicki, Rafał Kozik, and Michał Choraś. New explainability method for bert-based model in fake news detection. *Scientific reports*, 11(1):23705, 2021.

[317] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. https://arxiv.org/abs/1409.4842, 2015. Accessed: 2024-12-12.

[318] Samaneh Taheri, Seyed Hadi Hashemi, Albert Y Zomaya, and Jianming Yong. Sequence graph transform: A general approach to substructure-aware sequence encoding. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 117–126. IEEE, 2018.

[319] Kian Long Tan, Chin Poo Lee, and Kian Ming Lim. Fn-net: A deep convolutional neural network for fake news detection. In *2021 9th International Conference on Information and Communication Technology (ICoICT)*, pages 331–336. IEEE, 2021.

[320] Hongyan Tang, Junning Liu, Ming Zhao, and Xudong Gong. Progressive layered extraction (ple): A novel multi-task learning (mtl) model for personalized recommendations. In *Proceedings of the 14th ACM Conference on Recommender Systems*, pages 269–278, 2020.

[321] Wei Tang, Zuyao Ma, Haifeng Sun, and Jingyu Wang. Learning sparse alignments via optimal transport for cross-domain fake news detection. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[322] Yahya Tashtoush, Balqis Alrababah, Omar Darwish, Majdi Maabreh, and Nasser Alsaedi. A deep learning framework for detection of covid-19 fake news on social media platforms. *Data*, 7(5):65, 2022.

[323] Jitendra Vikram Tembhurne, Md Moin Almin, and Tausif Diwan. Mc-dnn: Fake news detection using multi-channel deep neural networks. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 18(1):1–20, 2022.

[324] Madusha Prasanjith Thilakarathna, Vihanga Ashinsana Wijayasekara, Yasiru Gamage, Kavindi Hanshani Peiris, Chanuka Abeysinghe, Intizar Rafaideen, and Prathieshna Vekneswaran. Hybrid approach and architecture to detect fake news on twitter in real-time using neural networks. In *2020 5th International Conference on Information Technology Research (ICITR)*, pages 1–6. IEEE, 2020.

[325] Yi Tian, Jialing Gu, Yiiun Jia, and Richard O Sinnott. An exploration of machine and deep learning models for fake news detection in social media. In *2021 8th International Conference on Behavioral and Social Computing (BESC)*, pages 1–6. IEEE, 2021.

[326] Vijay Srinivas Tida, Dr Sonya Hsu, and Dr Xiali Hei. Unified fake news detection using transfer learning of bert model. *IEEE*, 2020.

[327] Fatoumata Wongbé Rosalie Tokpa, Beman Hamidja Kamagaté, Vincent Monsan, and Souleymane Oumtanaga. Fake news detection in social media: Hybrid deep learning approaches. *J. Adv. Inf. Technol*, 14(3):606–615, 2023.

[328] Fatemeh Torgheh, Mohammad Reza Keyvanpour, Behrooz Masoumi, and Seyed Vahab Shojaedini. A novel method for detecting fake news: Deep learning based on propagation path concept. In *2021 26th International Computer Conference, Computer Society of Iran (CSICC)*, pages 1–5. IEEE, IEEE, 2021.

[329] Tina Esther Trueman, Ashok Kumar, P Narayanasamy, and J Vidya. Attention-based c-bilstm for fake news detection. *Applied Soft Computing*, 110:107600, 2021.

[330] Ciprian-Octavian Truică and Elena-Simona Apostol. Misrobærta: Transformers versus misinformation. *Mathematics*, 10(4):569, 2022.

[331] Amina Yusuf Umar, Ibrahim Said Ahmad, and Khadijah Muhammad. Fake news detection using cnn/gru deep learning model. *Sule Lamido University Journal of Science & Technology*, 7(1):57–67, 2023.

[332] 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.

[333] Bibek Upadhayay and Vahid Behzadan. Hybrid deep learning model for fake news detection in social networks (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 13067–13068, 2022.

[334] P Ushashree, Archana Naik, Siddarth Gurav, Ankit Kumar, SR Chethan, and BS Madhumala. Fake news detection using neural network. In *2023 IEEE International Conference on Integrated Circuits and Communication Systems (ICICACS)*, pages 01–05. IEEE, 2023.

[335] Bruno Vaz, Vítor Bernardes, and Álvaro Figueira. On creation of synthetic samples from gans for fake news identification algorithms. In *World Conference on Information Systems and Technologies*, pages 316–326. Springer, 2022.

[336] Matthew Veres and Medhat Moussa. Deep learning for intelligent transportation systems: A survey of emerging trends. *IEEE Transactions on Intelligent transportation systems*, 21(8):3152–3168, 2019.

[337] Abhishek Verma, Vanshika Mittal, and Suma Dawn. Find: Fake information and news detections using deep learning. In *2019 twelfth international conference on contemporary computing (IC3)*, pages 1–7. IEEE, 2019.

[338] Pawan Kumar Verma, Prateek Agrawal, Vishu Madaan, and Radu Prodan. Mcred: multi-modal message credibility for fake news detection using bert and cnn. *Journal of Ambient Intelligence and Humanized Computing*, 14(8):10617–10629, 2023.

[339] Humberto Fernandes Villela, Fábio Corrêa, Jurema Suely de Araújo Nery Ribeiro, Air Rabelo, and Dárlinton Barbosa Feres Carvalho. Fake news detection: a systematic literature review of machine learning algorithms and datasets. *Journal on Interactive Systems*, 14(1):47–58, 2023.

[340] Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 647–653, 2017.

[341] Jan Vom Brocke, Alan Hevner, and Alexander Maedche. Introduction to design science research. *Design science research. Cases*, pages 1–13, 2020.

[342] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *science*, 359(6380):1146–1151, 2018.

[343] Chang Wang and Sridhar Mahadevan. Heterogeneous domain adaptation using manifold alignment. In *IJCAI proceedings-international joint conference on artificial intelligence*, volume 22, page 1541, 2011.

[344] Jingqiong Wang and Xinzhu Li. Radiation fears prompt panic buying of salt. *China Daily*, 18:2011–03, 2011.

[345] William Yang Wang. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*, 2017.

[346] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, pages 849–857, 2018.

[347] Zhichao Wang. Deep learning methods for fake news detection. In *2022 IEEE 2nd International Conference on Data Science and Computer Application (ICDSCA)*, pages 472–475. IEEE, 2022.

[348] Zuhui Wang, Zhaozheng Yin, and Young Anna Argyris. Detecting medical misinformation on social media using multimodal deep learning. *IEEE journal of biomedical and health informatics*, 25(6):2193–2203, 2020.

[349] Apurva Wani, Isha Joshi, Snehal Khandve, Vedangi Wagh, and Raviraj Joshi. Evaluating deep learning approaches for covid19 fake news detection. In *Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event, February 8, 2021, Revised Selected Papers 1*, pages 153–163. Springer, 2021.

[350] Claire Wardle. Fake news challenge. http://www.fakenewschallenge.org/, 2017. Accessed: 2024-12-11.

[351] Claire Wardle. Fake news. it's complicated. https://firstdraftnews.org/articles/fake-news-complicated, 2017. Accessed: 2024-11-11.

[352] Sunita Warjri, Partha Pakray, Saralin A Lyngdoh, and Arnab K Maji. Fake news detection using social media data for khasi language. In *2023 International Conference on Intelligent Systems, Advanced Computing and Communication (ISACC)*, pages 1–6. IEEE, 2023.

[353] Przemyslaw M Waszak, Wioleta Kasprzycka-Waszak, and Alicja Kubanek. The spread of medical fake news in social media–the pilot quantitative study. *Health policy and technology*, 7(2):115–118, 2018.

[354] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3(1):1–40, 2016.

[355] Michael Wessel, Ferdinand Thies, Alexander Benlian, et al. A lie never lives to be old: The effects of fake social information on consumer decision-making in crowdfunding. In *ECIS*, 2015.

[356] Ian H Witten, Eibe Frank, Mark A Hall, Christopher J Pal, and MINING DATA. Practical machine learning tools and techniques. In *Data Mining*, volume 2, 2005.

[357] Liang Wu, Jundong Li, Xia Hu, and Huan Liu. Gleaning wisdom from the past: Early detection of emerging rumors in social media. In *Proceedings of the 2017 SIAM international conference on data mining*, pages 99–107. SIAM, 2017.

[358] 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.

[359] Lianwei Wu, Yuan Rao, Hualei Yu, Yiming Wang, and Ambreen Nazir. False information detection on social media via a hybrid deep model. In *Social Informatics:*

10th International Conference, SocInfo 2018, St. Petersburg, Russia, September 25-28, 2018, Proceedings, Part II 10, pages 323–333. Springer, 2018.

[360] Ning Xiang et al. Deep learning-based fake information detection and influence evaluation. *Computational Intelligence and Neuroscience*, 2022, 2022.

[361] Feng Xing and Caili Guo. Mining semantic information in rumor detection via a deep visual perception based recurrent neural networks. In *2019 IEEE International Congress on Big Data (BigDataCongress)*, pages 17–23. IEEE, 2019.

[362] Hu Xu, Bing Liu, Lei Shu, and Philip S Yu. Bert post-training for review reading comprehension and aspect-based sentiment analysis. *arXiv preprint arXiv:1904.02232*, 2019.

[363] Arun Kumar Yadav, Suraj Kumar, Dipesh Kumar, Lalit Kumar, Kapil Kumar, Sandeep Kumar Maurya, Mohit Kumar, and Divakar Yadav. Fake news detection using hybrid deep learning method. *SN Computer Science*, 4(6):845, 2023.

[364] Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. Deepfakebench: A comprehensive benchmark of deepfake detection. *Advances in Neural Information Pro-cessing Systems*, 2(6):1–32, 2023.

[365] Yang Yang, Lei Zheng, Jiawei Zhang, Qingcai Cui, Zhoujun Li, and Philip S Yu. Ti-cnn: Convolutional neural networks for fake news detection. *arXiv preprint arXiv:1806.00749*, page 1806.00749, 2018.

[366] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.

[367] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.

[368] Savvas Zannettou, Barry Bradlyn, Emiliano De Cristofaro, Haewoon Kwak, Michael Sirivianos, Gianluca Stringini, and Jeremy Blackburn. What is gab: A bastion of free speech or an alt-right echo chamber. In *Companion Proceedings of the The Web Conference 2018*, pages 1007–1014, 2018.

[369] Alexandros Zervopoulos, Aikaterini Georgia Alvanou, Konstantinos Bezas, Asterios Papamichail, Manolis Maragoudakis, and Katia Kermanidis. Deep learning for fake

news detection on twitter regarding the 2019 hong kong protests. *Neural Computing and Applications*, 34(2):969–982, 2022.

[370] Qin Zhang, Zhiwei Guo, Yanyan Zhu, Pandi Vijayakumar, Aniello Castiglione, and Brij B Gupta. A deep learning-based fast fake news detection model for cyber-physical social services. *Pattern Recognition Letters*, 168:31–38, 2023.

[371] Yudong Zhang, Juan Manuel Gorriz, and Zhengchao Dong. Deep learning in medical image analysis. *Journal of Imaging*, 7:74, 2021.

[372] Feng Zhou, Yong Hu, and Xukun Shen. Msanet: multimodal self-augmentation and adversarial network for rgb-d object recognition. *The Visual Computer*, 35(11):1583–1594, 2019.

[373] Xinyi Zhou and Reza Zafarani. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40, 2020.

[374] Haichao Zhu and Richard O Sinnott. A performance comparison of fake news detection approaches. In *2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, pages 1–7. IEEE, 2021.

[375] Yin Zhu, Yuqiang Chen, Zhongqi Lu, Sinno Pan, Gui-Rong Xue, Yong Yu, and Qiang Yang. Heterogeneous transfer learning for image classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 25, pages 1304–1309, 2011.

[376] Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR)*, 51(2):1–36, 2018.

[377] Arkaitz Zubiaga, Maria Liakata, and Rob Procter. Learning reporting dynamics during breaking news for rumour detection in social media. *arXiv preprint arXiv:1610.07363*, 2016.

APPENDICES

Appendix A

Extra Experimental Tests

A.1 BERTGuard Detailed Results

Table A.1 represents the testing results of all models using the Snope dataset, Table A.2 shows the testing results of all models using the COVID-Claims dataset, Table A.3 shows the testing results of all models using the Pheme datasets, Table A.4 shows the testing results of all models using the Politifact dataset, Table A.5 shows the testing results of all models using the Climate dataset, and Table A.6 shows the testing results of all models using the GossipCop. dataset.

Table A.1: Testing Results on Snope

Model	Precision	Recall	F1-score	Accuracy
Baseline 1 (Single-Stage)	0.583011583	0.774358974	0.665198238	0.517460317
Baseline 2 (FakeBERT)	0.384977855	0.483582913	0.428683218	0.447892589
BERTGuard	**0.837719298**	**0.979487179**	**0.903073286**	**0.86984127**
Baseline 1 (Single-Stage) + Adjusting Class Weights	0.472440945	0.625	0.538116592	0.463541667
Baseline 2 (FakeBERT) + Adjusting Class Weights	0.412659847	0.489921405	0.447983806	0.432982697
BERTGuard + Adjusting Class Weight	**0.772357724**	**0.989583333**	**0.867579909**	**0.848958333**
Baseline 1 (Single-Stage) + Random Oversampling	0.579630124	0.775894752	0.663553992	0.509871545
Baseline 2 (FakeBERT) + Random Oversampling	0.409235843	0.509821473	0.454024394	0.418259451
BERTGuard + Random Oversampling	**0.829632541**	**0.963254891**	**0.891464337**	**0.889193257**
Baseline 1 (Single-Stage) + Random Downsampling	0.503514973	0.614358974	0.553441545	0.556482013
Baseline 2 (FakeBERT) + Random Downsampling	0.394272455	0.479155478	0.43258934	0.469781209
BERTGuard + Random Downsampling	**0.829914686**	**0.860479554**	**0.844920791**	**0.821669209**

Table A.2: Testing Results on Covid-Claims

Model	Precision	Recall	F1-score	Accuracy
Baseline 1 (Single-Stage)	0.224425887	0.135135135	0.168693605	0.246979389
Baseline 2 (FakeBERT)	0.189236934	0.115833972	0.143704727	0.219358114
BERTGuard	**0.956869994**	**0.976115651**	**0.966397013**	**0.961620469**
Baseline 1 (Single-Stage) + Adjusting Class Weights	0.492921493	0.391615542	0.436467236	0.494376278
Baseline 2 (FakeBERT) + Adjusting Class Weights	0.301254982	0.498246303	0.375482026	0.489244721
BERTGuard + Adjusting Class Weight	**0.943564356**	**0.974437628**	**0.958752515**	**0.95807771**
Baseline 1 (Single-Stage) + Random Oversampling	0.298140269	0.260536923	0.278073096	0.325971273
Baseline 2 (FakeBERT) + Random Oversampling	0.225971472	0.293258705	0.2552552	0.223804178
BERTGuard + Random Oversampling	**0.938367825**	**0.971478833**	**0.954636306**	**0.958731071**
Baseline 1 (Single-Stage) + Random Downsampling	0.215957211	0.130287603	0.162524007	0.159320159
Baseline 2 (FakeBERT) + Random Downsampling	0.118349661	0.229854687	0.156248619	0.39254215
BERTGuard + Random Downsampling	**0.887931557**	**0.898320015**	**0.893095577**	**0.899914330**

Table A.3: Testing Results on Pheme

Model	Precision	Recall	F1-score	Accuracy
Baseline 1 (Single-Stage)	0.313032887	0.192509363	0.238404453	0.744396015
Baseline 2 (FakeBERT)	0.211972237	0.159728818	0.182179063	0.208359716
BERTGuard	**0.847560976**	**0.624719101**	**0.71927555**	**0.89866127**
Baseline 1 (Single-Stage) + Adjusting Class Weights	0.678362573	0.434456929	0.529680365	0.61423221
Baseline 2 (FakeBERT) + Adjusting Class Weights	0.587471475	0.452015893	0.510918077	0.498259711
BERTGuard + Adjusting Class Weight	**0.942528736**	**0.61423221**	**0.743764172**	**0.788389513**
Baseline 1 (Single-Stage) + Random Oversampling	0.382557854	0.198977128	0.261790831	0.483258776
Baseline 2 (FakeBERT) + Random Oversampling	0.338326978	0.226862148	0.271603191	0.236484122
BERTGuard + Random Oversampling	**0.853290018**	**0.608954712**	**0.710708634**	**0.899101146**
Baseline 1 (Single-Stage) + Random Downsampling	0.310138297	0.226820179	0.262015136	0.386214479
Baseline 2 (FakeBERT) + Random Downsampling	0.183925587	0.159728818	0.223484548	0.208359716
BERTGuard + Random Downsampling	**0.820173808**	**0.579873577**	**0.679401462**	**0.601333679**

Table A.4: Testing Results on Politifact

Model	Precision	Recall	F1-score	Accuracy
Baseline 1 (Single-Stage)	0.486368313	0.18228263	0.265180199	0.49520736
Baseline 2 (FakeBERT)	0.386986823	0.125522351	0.189559518	0.389158614
BERTGuard	**0.800490597**	**0.083248299**	**0.150812601**	**0.478772693**
Baseline 1 (Single-Stage) + Adjusting Class Weights	0.57832193	0.387034582	0.463726269	0.552415954
Baseline 2 (FakeBERT) + Adjusting Class Weights	0.499128995	0.342180216	0.40601497	0.448917520
BERTGuard + Adjusting Class Weight	**0.834051724**	**0.139896373**	**0.239603756**	**0.556030847**
Baseline 1 (Single-Stage) + Random Oversampling	0.451789226	0.201493781	0.278693057	0.475873331
Baseline 2 (FakeBERT) + Random Oversampling	0.448209722	0.289173815	0.351541657	0.398731518
BERTGuard + Random Oversampling	**0.791581883**	**0.123983301**	**0.214387652**	**0.479918739**
Baseline 1 (Single-Stage) + Random Downsampling	0.416754822	0.228279145	0.294981161	0.441993275
Baseline 2 (FakeBERT) + Random Downsampling	0.236581751	0.298245412	0.263858782	0.317295515
BERTGuard + Random Downsampling	**0.780304702**	**0.069358933**	**0.127394182**	**0.328347991**

Table A.5: Testing Results on Climate

Model	Precision	Recall	F1-score	Accuracy
Baseline 1 (Single-Stage)	0.349206349	0.434782609	0.387323944	0.61631753
Baseline 2 (FakeBERT)	0.269025792	0.340992147	0.300763885	0.331289459
BERTGuard	**0.534798535**	**0.577075099**	**0.55513308**	**0.742006615**
Baseline 1 (Single-Stage) + Adjusting Class Weights	0.655462185	0.386138614	0.485981308	0.591584158
Baseline 2 (FakeBERT) + Adjusting Class Weights	0.283515820	0.332652552	0.306124966	0.329518024
BERTGuard + Adjusting Class Weight	**0.7125**	**0.564356436**	**0.629834254**	**0.668316832**
Baseline 1 (Single-Stage) + Random Oversampling	0.479190576	0.44372519	0.460776459	0.628347923
Baseline 2 (FakeBERT) + Random Oversampling	0.301710366	0.368231514	0.331668367	0.331289459
BERTGuard + Random Oversampling	**0.519136522**	**0.593255101**	**0.553726554**	**0.750148775**
Baseline 1 (Single-Stage) + Random Downsampling	0.281452588	0.401997021	0.331094204	0.412865201
Baseline 2 (FakeBERT) + Random Downsampling	0.238154502	0.411355454	0.301661745	0.398451714
BERTGuard + Random Downsampling	**0.529663888**	**0.523665556**	**0.526647643**	**0.528911174**

Table A.6: Testing Results on GossipCop

Model	Precision	Recall	F1-score	Accuracy
Baseline 1 (Single-Stage)	0.378611058	0.743565658	0.501743044	0.644941283
Baseline 2 (FakeBERT)	0.321026966	0.619652353	0.422939275	0.601397116
BERTGuard	**0.835781991**	**0.66259628**	**0.739180551**	**0.887579042**
Baseline 1 (Single-Stage) + Adjusting Class Weights	0.707474227	0.644668859	0.674612927	0.689055895
Baseline 2 (FakeBERT) + Adjusting Class Weights	0.337935218	0.582315714	0.427676802	0.375923658
BERTGuard + Adjusting Class Weight	**0.933965459**	**0.647721935**	**0.764942449**	**0.800962893**
Baseline 1 (Single-Stage) + Random Oversampling	0.418263754	0.74316255	0.535269361	0.628739785
Baseline 2 (FakeBERT) + Random Oversampling	0.329187265	0.573599781	0.41830849	0.389217214
BERTGuard + Random Oversampling	**0.819732297**	**0.670117381**	**0.737412462**	**0.890125896**
Baseline 1 (Single-Stage) + Random Downsampling	0.320173659	0.681455469	0.435658438	0.665847845
Baseline 2 (FakeBERT) + Random Downsampling	0.298966547	0.661492152	0.411811616	0.562321354
BERTGuard + Random Downsampling	**0.786928771**	**0.550944282**	**0.648124134**	**0.728914474**