Detecting Fake News in Vietnamese Media BGRA2025

by An 31231025020 - Thái Hoài

Submission date: 03-Aug-2025 12:39PM (UTC+0700)
Submission ID: 2724362664
File name: 29988_An_31231025020_-_Th_i_Ho_i_Detecting_Fake_News_in_Vietnamese_Media_BGRA2025_260_869222391.pdf (3.04M)
Word count: 19008
Character count: 118256

Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
API	Application Programming Interface
AUC	Area Under the Curve
BERT	Bidirectional Encoder Representations from Transformers
BiLSTM	Bidirectional Long Short-Term Memory
BPE	Byte-Pair Encoding
CBOW	Continuous Bag of Words
CNN	Convolutional Neural Network
COVID-19	Coronavirus Disease 2019
CPU	Central Processing Unit
CUDA	Compute Unified Device Architecture
DL	Deep Learning
DM	Data Mining
DS	Data Science
EDA	Exploratory Data Analysis
FN	False Negatives
FP	False Positives
FPR	False Positive Rate
GPU	Graphics Processing Unit
LLM	Large Language Model
LSTM	Long Short-Term Memory
mBERT	Multilingual Bidirectional Encoder Representations from Transformers
MoE	Mixture-of-Experts
NLP	Natural Language Processing

PLM	Pre-trained Language Model
RAM	Random Access Memory
ReINTEL	Reliable Information Identification
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
SNS	Social Network Sites
TN	True Negatives
TP	True Positives
TPR	True Positive Rate
UEH	University of Economics Ho Chi Minh City (Đại học Kinh tế TP. Hồ Chí Minh)
VRAM	Video Random Access Memory

Abstract

The proliferation of misinformation on Vietnamese social media platforms necessitates robust automated detection systems. This study presents a comprehensive comparative evaluation of machine learning approaches for Vietnamese fake news detection, systematically analyzing three major model families: traditional deep learning, transfer learning, and large language models.

Using the ReINTEL dataset containing 9,713 Vietnamese social media posts with severe class imbalance (83.2% real vs 16.8% fake news), we evaluated ten distinct model configurations. Traditional deep learning employed BiLSTM networks with three embedding strategies (random, Word2Vec, FastText). Transfer learning utilized PhoBERT in frozen and fine-tuned configurations. Large language models included Qwen2.5-7B, Llama-2-7B, and DeepSeek across zero-shot and few-shot paradigms.

Results demonstrate the overwhelming superiority of fine-tuned PhoBERT, achieving 96.30% accuracy with balanced performance across classes. This represents a 23-percentage-point improvement over the best LLM (Qwen2.5 at 73.25%). BiLSTM approaches consistently achieved 81.89% accuracy while failing on minority class detection (2.4% recall).

The evaluation reveals critical insights about Vietnamese language processing: the importance of language-specific pre-training, limitations of multilingual models for specialized tasks, and counterintuitive few-shot learning degradation. PhoBERT provides optimal efficiency balance (2GB memory, 20ms inference) compared to LLMs requiring 8GB memory and 3+ second inference times.

These findings establish performance benchmarks for Vietnamese fake news detection while validating continued importance of language-specific model development for critical applications in low-resource languages.

Keywords: Fake news detection, Vietnamese natural language processing, deep learning, transfer learning, large language models, PhoBERT, class imbalance, social media misinformation

Chapter 1

Introduction

1.1 Rationale of the research

The significant advancement of information technology has dramatically altered how news disseminates around the world, particularly in multi-platform journalism. As the Internet and social media platforms gained popularity, users gradually shifted their approach to news, moving away from traditional media such as newspapers and television to online sources. According to Shu et al. (2017a), there are two main reasons for this trend: Firstly, news published on social media platforms has a tendency to reach audiences faster and at a lower cost than traditional media outlets like newspapers or television. Secondly, the speed and efficiency of news dissemination are enhanced as users interact with the content by commenting or sharing it. This trend forces journalism and journalists to adapt their approaches towards readers. Multi-platform journalism enables readers to access the latest news across diverse platforms, from traditional newspapers to online news, live-streaming videos and social media. This diversity not only helps the journalism industry adapt to the developments of the times but also effectively meets the increasingly diverse and rapid demands of the public in the digital era (T. V. A. Nguyen, 2024).

As of early 2025, Vietnam had approximately 79.8 million Internet users, representing 78.8% of the total population, up from 78.4 million users (79%) in early 2024 (DataReportal, 2025). The country ranks among the top three in Southeast Asia for Internet users, trailing only Indonesia (about 212 million) and the Philippines (around 87 million) (Topics, 2025). With a high penetration rate and over 76.2 million active social media accounts (75.2% of the population), Vietnam demonstrates strong potential for adopting online journalism models through digital platforms such as social media and mobile applications (DataReportal, 2025). However, this digital expansion has also fueled the spread of fake news.

Despite the fact that social media offers various benefits, such as easy access to news at a low cost, the news quality on such platforms tends to be considerably worse than traditional journalism organizations. This has contributed to a serious issue, which is the rise of fake news – news that has purposely misinformation and is published online for financial or political benefits (Shu et al., 2017a). Additionally, with the influential power of mass media, certain

individuals or organizations have taken advantage of the ability to manipulate information to achieve their own goals. This has led to the emergence of articles that are not entirely truthful or even completely misleading. Notably, there even exist various websites created mainly for publishing fake news, purposely uploading fabricated information, propagandizing and misleading, while leveraging social media to drive traffic and amplify their effects (Granik & Mesyura, 2017).

The impact of fake news extends far beyond simply distorting information; it can have serious consequences across various aspects of social life. Economically, fake news or false rumors can erode trust and damage the reputation of businesses, potentially destabilizing entire economies during sensitive periods (P. Nguyen, 2023). For instance, in late March and early April 2022, false rumors circulating on social media about certain publicly listed companies being under investigation caused panic among investors. This led to a sharp decline in the stock prices of these companies, despite press releases that clarified the situation. In the realm of politics and national security, fake news is often exploited to execute political conspiracies, disrupt social order, and harm international relations (N. T. Le, 2022). During the 2016 U.S. presidential election, fake news was blamed for exacerbating political polarization and partisan conflict throughout the campaign (Riedel et al., 2017). Voters were easily swayed by misleading statements and policies, which had the potential to skew the election results (Zhang & Ghorbani, 2020). In Vietnam, fake news can create misunderstandings related to ethnic, religious, or regional issues, increasing tensions and potentially leading to violence or social and security instability (P. Nguyen, 2023).

As the prevalence of fake news continues to increase worldwide, automated fake news detection has emerged as an important area of research and development. Researchers around the globe are focusing on creating automated tools for fake news detection that utilize machine learning, deep learning, and natural language processing techniques. This area of study has garnered significant attention and is yielding promising results (Huang, 2020; Thota et al., 2018; Zhou et al., 2020). In Vietnam, the field of fake news detection is relatively new, but there have been significant advancements. The Vietnam Anti-Fake News Center (VAFC) has launched the website *tingia.gov.vn* to receive reports and publish alerts about fake news, helping to warn the public. Furthermore, academic research that utilizes machine learning and natural language processing (NLP) for the Vietnamese language is being actively promoted. This research enhances the ability to detect fake news in the Vietnamese context (N. T. Le, 2022; D. Q. Nguyen & Nguyen, 2020a; Pham et al., 2021). However, the development of fake news detection tools in Vietnam still faces several major challenges. One significant issue is the lack of necessary infrastructure for language processing, such as machine-readable dictionaries and large linguistic corpora (Dinh, 2013).

This underscores the imperative to continuously invest in research and the application of automated tools for Vietnamese news in light of the increasingly complex challenges posed to both readers and domestic researchers. Consequently, the primary objective of this study is to propose a fake news detection model for the Vietnamese language using deep learning

techniques. The research focuses on training and evaluating various deep learning models, comparing their effectiveness in detecting fake news. By leveraging advanced machine learning architectures, this study aims to identify the most optimal model for Vietnamese fake news detection, contributing to the development of automated tools for misinformation identification in the digital landscape.

1.2 Research objectives

The primary objective of this study is to develop and evaluate comprehensive machine learning approaches for Vietnamese fake news detection through a systematic comparison of different model families and learning paradigms. Specifically, the research aims to establish performance benchmarks by evaluating traditional deep learning architectures (BiLSTM), transfer learning approaches (PhoBERT), and large language models (Qwen2.5, Llama-2, DeepSeek) across zero-shot, few-shot, and fine-tuning methodologies. This comprehensive evaluation seeks to identify optimal approaches for accuracy, computational efficiency, and practical deployment while addressing the unique challenges posed by Vietnamese language characteristics and severe class imbalance in misinformation detection.

In addition, this study aims to conduct a comprehensive comparison of the most prominent model families currently used in text classification. Three major categories are examined: traditional deep learning architectures such as BiLSTM, transfer learning approaches based on BERT-like models, and state-of-the-art LLMs such as Qwen2.5, Llama2 and DeepSeek. By evaluating the performance, strengths, and limitations of each model family under different conditions, the study seeks to identify the optimal approach for maximizing accuracy and generalizability in fake news detection systems.

1.3 Contributions

First, it develops and analyses three representative model families for Vietnamese text classification. The research provides a detailed description of the evaluation pipeline, including model architectures, technical specifications, prompt engineering strategies, and comprehensive performance assessment across multiple learning paradigms. This systematic approach ensures that the models are assessed consistently and fairly across all experiments.

Second, the study conducts a thorough benchmarking of these models against each other to evaluate their performance under different conditions. The results highlight the superior performance of language-specific pre-trained models over multilingual approaches, demonstrating the critical importance of Vietnamese-specific training for achieving high-quality fake news detection. Furthermore, the comparative analysis underscores the relative strengths and weaknesses of each model family, offering clear insights into which techniques are most effective for fake news detection.

By combining rigorous model development with transparent evaluation, these contributions

provide both researchers and practitioners with a robust foundation for selecting and deploying AI systems that are accurate, interpretable, and scalable for combating misinformation.

Chapter 2

Literature Review

This chapter provides a comprehensive review of the theoretical foundations and empirical research underlying fake news detection methodologies. We begin by examining the conceptual framework of fake news and its definitional challenges, followed by an analysis of the major technological approaches that have emerged to address this problem. The review traces the evolution from traditional machine learning techniques through deep learning architectures to the current state-of-the-art large language models, with particular attention to their application in Vietnamese language contexts.

The literature review is structured to provide both theoretical grounding and practical insights relevant to our comparative evaluation, examining not only the technical capabilities of different approaches but also their limitations and suitability for deployment in real-world Vietnamese fake news detection scenarios.

2.1 Fake News

Fake news has long existed, yet there is still no universally accepted definition in journalism or academic research (Shu et al., 2017b; Zhou & Zafarani, 2020). This lack of consensus creates difficulties in analyzing and evaluating existing studies on fake news and adds complexity to the development of reliable detection methods.

The Cambridge Dictionary (Cambridge Dictionary, n.d.) defines fake news as "false stories that appear to be news, spread on the internet or using other media, usually created to influence political views or as a joke" (retrieved July 27, 2025). This definition highlights three core components: (1) form, (2) environment, and (3) purpose. Tandoc Jr. et al. (2017) emphasized that the concept of "fake news" is broad, encompassing political satire, parody, state propaganda, and even false advertising. Subsequently, Tandoc Jr. (2019) refined the concept, defining fake news as "a specific type of disinformation: It is false, it is intended to deceive people, and it does so by trying to look like real news" (p.2). Similarly, Shu et al. (2017b) and Allcott and Gentzkow (2017) considered fake news to be "a news article that is intentionally and verifiably false," thus classifying it as a deceptive form of news (Zhou & Zafarani, 2020).

Although the definitions vary in scope, they converge on two fundamental characteristics

of fake news: inauthenticity and an explicit intent to mislead. Recognizing these attributes is crucial because they directly shape the strategies and technologies deployed in detection efforts.

2.2 Deep Learning-based Models

The growing prevalence of fake news has motivated researchers to adopt advanced machine learning techniques, particularly deep learning (DL), to address the detection challenge. DL, a subfield of machine learning, utilizes multi-layered computational models to learn increasingly abstract data representations (LeCun et al., 2015). These deep neural networks can automatically extract hierarchical features from raw input data, thereby reducing reliance on manual feature engineering that traditional machine learning techniques required (Mishra et al., 2021).

Popular DL architectures include convolutional neural networks (CNNs), recurrent neural networks (RNNs), and their variants such as long short-term memory (LSTM) networks. More recently, Transformer-based models, which rely on attention mechanisms to capture global dependencies, have been widely adopted for complex natural language processing (NLP) tasks (Li et al., 2024). This ability to model high-dimensional data makes deep learning an effective backbone for fake news detection systems, which often rely on subtle textual cues and contextual relationships.

2.2.1 Application to Fake News Detection

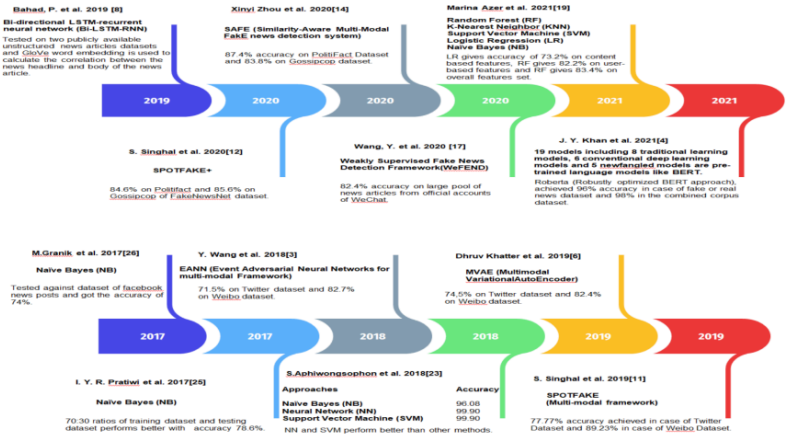

Figure 2.1: Timeline of Fake News Detection. Source: Rani and Virmani (2022)

The evolution of fake news detection approaches, illustrated in Figure 2.1, demonstrates the progression from traditional methods to modern AI-based solutions.

Deep learning approaches have demonstrated particular effectiveness in fake news detection due to their ability to capture subtle linguistic patterns and contextual relationships that traditional feature-based methods often miss (Zhou & Zafarani, 2020). Kaliyar et al. (2021) showed that CNN-based models could effectively identify deceptive language patterns by learning hierarchical representations of textual features. Similarly, Nasir et al. (2021) demonstrated that LSTM networks excel at capturing sequential dependencies in news articles that distinguish authentic from fabricated content.

However, traditional deep learning approaches face significant challenges when applied to low-resource languages like Vietnamese, where limited training data and unique linguistic characteristics can severely impact model performance (Nguyễn & Nguyễn, 2020). These limitations have motivated the development of transfer learning approaches and cross-lingual models specifically adapted for Vietnamese text processing.

2.3 Transformer Architecture

Among DL architectures, the Transformer introduced by Vaswani et al. (2017) represents a major breakthrough in NLP. Unlike RNNs or CNNs, Transformers rely entirely on self-attention mechanisms to capture relationships between tokens. This design enables parallel computation and models long-range dependencies effectively, making Transformers the foundation for many state-of-the-art NLP models. Core components include multi-head self-attention, feed-forward layers, residual connections, and positional encoding (Vaswani et al., 2017).

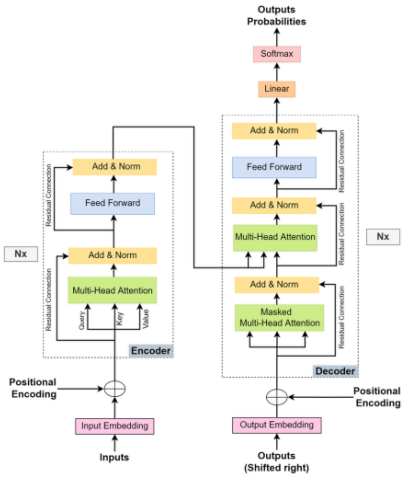

Figure 2.2: Transformer architecture. Source: Vaswani et al. (2017)

The Transformer architecture shown in Figure 2.2 revolutionized natural language process-

ing through its attention-based design.

Building upon this architecture, a series of influential models have emerged. BERT (Bidirectional Encoder Representations from Transformers) introduced deep bidirectional language representations using masked language modeling and next sentence prediction, enabling efficient fine-tuning for downstream tasks (Devlin et al., 2019). Variants such as DistilBERT (Sanh et al., 2019) reduce size and inference costs via knowledge distillation, while ALBERT (Lan et al., 2020) achieves memory efficiency through parameter sharing and embedding factorization.

Despite these advances, Transformer-based models pose challenges such as high computational cost and memory usage, particularly for long sequences, as well as the risk of propagating biases present in training data (Fields et al., 2023). Nevertheless, their flexibility and superior performance have established the Transformer family — especially BERT variants — as a preferred backbone for fake news detection systems.

2.4 Large Language Models (LLMs)

The success of Transformers has paved the way for Large LLMs, which leverage massive Transformer-based architectures trained on trillions of tokens to perform diverse language understanding and generation tasks (Raiaan et al., 2024). LLMs extend the concept of pre-trained language models (PLMs) by integrating advanced tokenization strategies, self-attention mechanisms, and large-scale pre-training objectives (e.g., masked language modeling, autoregressive generation) to capture complex linguistic dependencies and semantic nuances (Naveed et al., 2024).

The development pipeline of LLMs typically involves pre-training on vast corpora, fine-tuning on domain-specific or instruction-focused data, and alignment with human feedback to improve safety and reliability (Naveed et al., 2024). Architecturally, LLMs encompass decoder-only models (e.g., GPT series), encoder-decoder models (e.g., T5), and autoencoding models (e.g., BERT), while emerging approaches like mixture-of-experts (MoE) enhance scalability (Raiaan et al., 2024).

Beyond technical improvements, LLMs demonstrate emergent capabilities such as few-shot learning, reasoning, and in-context adaptation, which make them particularly suitable for tasks like text classification, summarization, question answering, and misinformation detection (Naveed et al., 2024). However, their reliance on uncurated, large-scale data introduces risks, including the propagation of social biases and harmful outputs. This has spurred extensive research into fairness and bias mitigation strategies at all stages of the training pipeline (Gallegos et al., 2024).

Yet, applying LLMs to fake news detection poses significant challenges. High computational requirements limit deployability in real-time systems, and fine-tuning LLMs to achieve high accuracy remains a difficult task (Zhang et al., 2024). Moreover, as fake news is increasingly produced by LLMs themselves (e.g., GPT-4, LLaMA), detection becomes more complex because synthetic text often mimics authentic content. This similarity can lead detectors to misclassify

human-written fake news as real or introduce bias toward LLM-generated outputs (Su et al., 2023).

2.5 Related Studies

Given these conceptual foundations, researchers worldwide have investigated various strategies to build sophisticated fake news detection systems using deep learning, transfer learning, and more recently LLMs.

On a global scale, Roumeliotis et al. (2025) conducted a comprehensive comparative study evaluating traditional CNN models, BERT, and GPT-4 Omni (including its mini version). The fine-tuned GPT-4 Omni model achieved an outstanding accuracy of 98.6%, significantly surpassing conventional methods (CNN achieved ∼58.6%). Interestingly, the smaller GPT-4o mini achieved results comparable to its larger counterpart, suggesting its potential for tasks with limited computational resources.

Focusing on the Vietnamese context, Nguyễn and Nguyễn (2020) explored transfer learning models such as PhoBERT and bert4news, which were pre-trained specifically for Vietnamese. Their combined model achieved an impressive AUC of 94.52% on the ReINTEL dataset, clearly demonstrating the effectiveness of transfer learning in low-resource languages. They also experimented with deep learning models like TextCNN and BiLSTM integrated with embeddings (FastText, PhoW2V), but these approaches performed significantly worse than transfer learning.

More recently, Võ and Đỗ (2023) built a dataset of authentic and fabricated Vietnamese articles collected from online newspapers and social media. They applied three deep learning models (LSTM, Bi-LSTM, CNN-BiLSTM) and found that the hybrid CNN-BiLSTM model performed best by effectively capturing spatial and sequential features. The study also pointed out unique challenges in Vietnamese language processing, such as syntactic ambiguity and difficulty in verifying source credibility. Building on this dataset, our research intends to address these challenges by incorporating natural language inference (NLI) and knowledge graph approaches with the support of LLMs to improve detection accuracy.

2.6 Research Gaps and Motivation

While the reviewed literature demonstrates significant progress in fake news detection methodologies, several critical gaps remain, particularly in the Vietnamese language context:

Limited Vietnamese Language Coverage: Despite the growing sophistication of fake news detection systems, comprehensive evaluations of state-of-the-art approaches on Vietnamese text remain scarce. Most existing studies focus on high-resource languages like English, leaving Vietnamese-specific challenges largely unexplored (Nguyễn & Nguyễn, 2020).

Incomplete Model Family Comparisons: Previous Vietnamese studies typically evaluate only one or two model families, failing to provide comprehensive comparisons across traditional

deep learning, transfer learning, and large language model approaches. This limitation prevents researchers and practitioners from making informed decisions about optimal model selection.

Insufficient Analysis of Learning Paradigms: The potential of zero-shot and few-shot learning for Vietnamese fake news detection remains largely untapped. Given the limited availability of labeled Vietnamese misinformation data, these paradigms could offer significant practical advantages that warrant systematic investigation.

Class Imbalance Considerations: Real-world fake news datasets exhibit severe class imbalance, yet most studies fail to adequately address this challenge or evaluate model robustness under such conditions. This gap is particularly critical for practical deployment scenarios.

Computational Efficiency Analysis: While accuracy comparisons are common, systematic evaluation of computational efficiency and deployment feasibility across different model families remains limited, hindering practical adoption decisions.

These gaps motivate our comprehensive evaluation framework that systematically compares diverse approaches while addressing the unique challenges of Vietnamese fake news detection in realistic deployment scenarios.

Chapter 3

Methodology

This chapter presents the comprehensive methodology employed for evaluating machine learning approaches to Vietnamese fake news detection. The methodology encompasses four primary components: exploratory data analysis of the ReINTEL dataset, implementation of traditional deep learning approaches using BiLSTM architectures, deployment of transfer learning techniques with PhoBERT, and systematic evaluation of large language models across multiple learning paradigms. Each component is designed to provide fair and comprehensive comparison while addressing the unique challenges of Vietnamese language processing and severe class imbalance in misinformation detection.

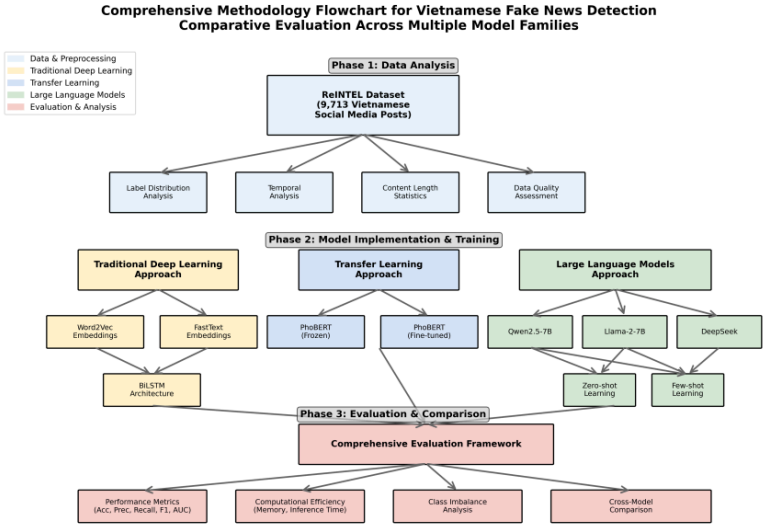

Figure 3.1: Methodology flowchart

Figure 3.1 provides a comprehensive overview of our systematic evaluation framework, illustrating the three-phase approach from data analysis through model implementation to comparative evaluation.

3.1 Exploratory Data Analysis (EDA)

In this study, we utilize the ReINTEL dataset (D.-T. Le et al., 2020) - a Vietnamese-language news dataset curated for the task of fake news detection. The dataset comprises news articles collected from various online media sources in Vietnam, annotated with binary labels: real or fake. Each record includes multiple fields such as the article content (post_message), timestamp (timestamp_post), source (page_name), and label (label).

3.1.1 Purpose and Context

The primary goal of the ReINTEL challenge is to identify whether a piece of information shared on Vietnamese social network sites (SNSs), such as Facebook, Zalo, or Lotus, is reliable or unreliable. This task specifically focuses on classifying Vietnamese news/posts as "unreliable" or "reliable".

The development of this task was driven by the rapid growth of SNSs in Vietnam and the increasing proliferation of unreliable information, especially in the context of events like the COVID-19 pandemic and significant political/economic developments. Detecting unreliable news has gained considerable attention.

The task addresses the challenge of categorizing unreliable news collected in Vietnamese, which is considered a low-resource language for natural language preprocessing.

3.1.2 Dataset Overview

The ReINTEL dataset employed in this study consists of 9,713 social media news posts, distributed across three subsets: 8,741 samples in the training set, and 486 samples each in the validation and test sets. Each instance contains seven features: `user_name`, `post_message`, `timestamp_post`, `num_like_post`, `num_comment_post`, `num_share_post`, and the binary `label` indicating whether the news is real (`0`) or fake (`1`).

Table 3.1: Structure and data types of the training dataset

#	Column Name	Data Type	Description
0	`user_name`	`object`	An anonymized identifier for the user who posted the content.
1	`post_message`	`object`	The full text content of the social media post.
2	`timestamp_post`	`object`	The timestamp indicating when the post was published.
3	`num_like_post`	`object`	The number of likes received by the post.
4	`num_comment_post`	`object`	The number of comments received by the post.
5	`num_share_post`	`object`	The number of times the post was shared.
6	`label`	`int64`	The ground-truth label of the post: 0 for real news, 1 for fake news.

Id: 0.
User id: 2167074723833130000.
Post message: Cần các bậc phụ huynh xã Ngũ Thái lên tiếng, không ngờ xã mình cũng nhận thịt nhiễm sán... Cho các cháu Mầm non ăn uống thế này thật vô nhân tính! VTV đăng tin rồi nhé các anh chị.
English translation: *Needing the parents of Ngu Thai commune to speak up, astonishing my commune accept contaminated meat ... Feeding preschool children like this is so inhumane! VTV posted the news, guys.*
Timestamp post: 1584426000.
Number of post's like: 45.
Number of post's comment: 15.
Number of post's share: 8.
Label: 1 (unreliable).
Image: NAN.

Figure 3.2: An example extracted from the dataset.

A preliminary inspection of the training set confirms that all records are complete, with no missing values in any column. The `label` field is of type `int`, while the remaining columns are stored as string (`object`) types, including engagement metrics such as likes, comments, and shares. These metrics may require further type conversion and normalization before being used in modeling or statistical analysis.

Overall, the dataset exhibits a clean and well-structured format, suitable for preprocessing and downstream feature engineering. The presence of engagement metrics alongside textual

content presents opportunities to incorporate both linguistic and social signals into the fake news classification task.

3.1.3 Label Distribution

The dataset exhibits a significant class imbalance across all subsets. In the training set, real news instances account for 7,269 samples (83.2%), while fake news instances make up only 1,472 samples (16.8%), resulting in a class ratio of approximately 0.20. A similar distribution is observed in both the validation and test sets, each containing 404 real news samples (83.1%) and 82 fake news samples (16.9%), maintaining the same ratio of 0.20.

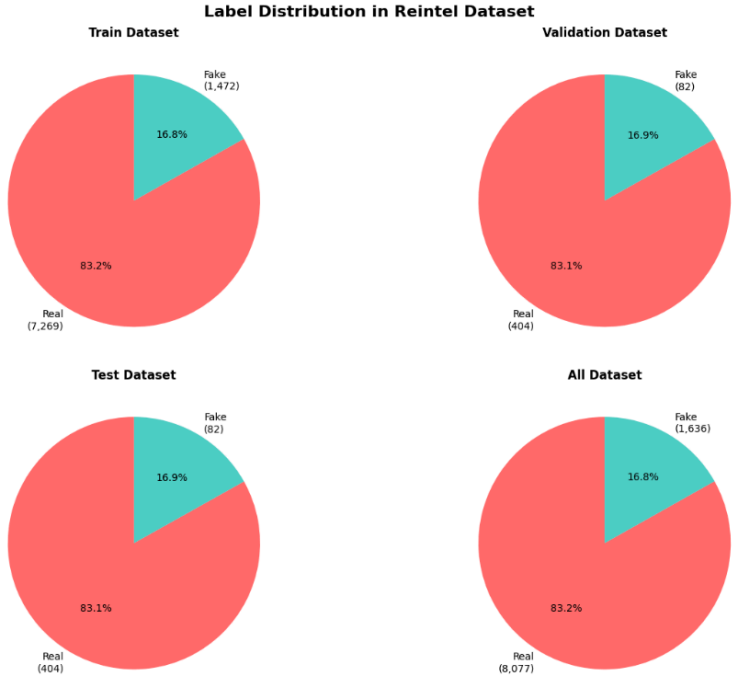

Figure 3.3: Label distribution across all ReINTEL dataset splits, visualized via pie charts.

When aggregated across all data splits, the dataset contains a total of 8,077 real news articles and 1,636 fake ones, corresponding to the same class distribution (83.2% vs. 16.8%). This level of imbalance is considered substantial and may introduce bias into the learning process, particularly for models that are sensitive to skewed class proportions. Consequently,

special attention must be given to model evaluation, including the use of performance metrics such as precision, recall, and F1-score, rather than relying solely on accuracy. Visualizations in the form of pie charts were generated to further illustrate the class distribution and ensure consistency across the dataset splits.

3.1.4 Time Analysis

Among the 9,705 valid timestamps in the dataset, the overwhelming majority of posts were published in the year 2020, accounting for approximately 96.0% of the total. The temporal distribution is highly concentrated, with only a small number of samples spanning earlier years: 50 in 2019, 26 in 2017, and negligible counts in 2014–2016. This pattern holds consistently across the training, validation, and test sets, where each subset exhibits a dominant spike in content during 2020.

The observed concentration reflects the dataset's topical focus, which likely aligns with major social and political events during that period—most notably the COVID-19 pandemic. This narrow time window, while helpful for analyzing contemporary misinformation, introduces a potential risk of temporal bias. Therefore, any machine learning model trained on this dataset must be interpreted with caution when applied to future or temporally distant data.

To summarize, while the timestamps are mostly valid and complete, the strong temporal skew toward 2020 highlights the importance of considering time-awareness in both model training and evaluation strategies.

3.1.5 Content Length Statistics

The post_message field, which holds the main textual content of each social media post, shows substantial variation in length across the dataset. In the training set, the average message length is 760 characters, while the median is 252, indicating a skewed distribution. The standard deviation is high (1,445 characters), and extreme values are present (up to 28,689 characters). Similar statistics are observed in the validation and test sets, with overall consistency in distributional patterns. When grouped into fixed length bins, a majority of the messages fall into the short (50–200) and medium (200–500) categories, though a notable proportion exceed 1,000 characters. Additionally, label-wise analysis reveals that fake news posts are typically longer and more variable than real ones across all splits.

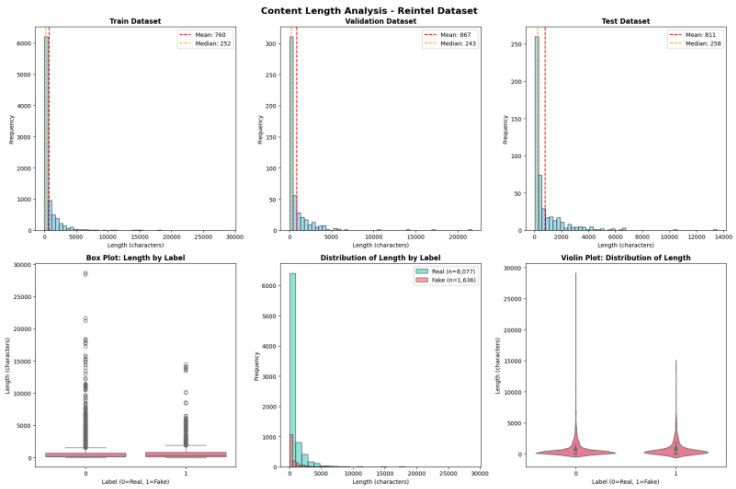

Figure 3.4: Content length distribution in the ReINTEL dataset.

The histograms (top row) confirm the right-skewed nature of content length, with most posts being under 1,000 characters and a long tail of rare, lengthy messages. Vertical dashed lines represent the mean (red) and median (orange), both indicating strong asymmetry. The bottom-left box plot reveals the presence of numerous outliers in both real and fake news posts. Interestingly, while real news exhibits a greater number of extreme outliers, this is likely attributed to the larger sample size of the real class. In contrast, the spread and interquartile range are visibly broader for fake news, indicating higher variability in content length. The bar chart in the center compares the overall length distribution by label and confirms that fake news posts, despite being fewer in number, are generally longer on average. This finding is reinforced by the violin plot, which illustrates a wider and flatter distribution for fake news, with both the median and density peak shifted toward higher character counts. These visualizations collectively suggest that fake news tends to be more verbose and exhibits greater variability in textual structure.

Overall, the analysis reveals that content length is a potentially valuable feature in fake news detection. The consistent difference in length between real and fake posts, particularly the longer and more variable nature of fake content, could help models better discriminate between classes. This suggests that incorporating length-based features—either directly or via engineered input representations—may enhance classification performance.

3.1.6 Data Quality and Redundancy Analysis

A comprehensive assessment of data quality and redundancy was conducted to ensure the integrity of subsequent model training and evaluation. Encouragingly, the dataset shows high reliability in terms of message uniqueness. All three subsets contain entirely unique `post_message` entries, with no internal duplication observed. This indicates strong consistency in content-level sampling.

In contrast, user-level redundancy is present across all subsets. The training set includes 8,741 posts but only 3,567 unique users, implying that many users contribute multiple posts. Similar trends are found in the validation (391 unique users) and test (411 unique users) sets. While not inherently problematic, such repetition may influence learning dynamics, especially in tasks where author bias or posting behavior plays a role.

Regarding missing data, no null values were found across key fields, further confirming the dataset's structural integrity and readiness for modeling.

Critically, cross-split analysis revealed no overlapping messages between the training, validation, and test sets, effectively ruling out data leakage. This is an important assurance, as data leakage—particularly from training to test—can severely bias model performance and invalidate evaluation metrics.

In summary, the dataset demonstrates high structural integrity with unique messages, minimal missing values, and clean split boundaries. Although user-level duplication exists, it does not compromise message-level independence. The dataset is therefore considered well-prepared for downstream machine learning tasks without requiring additional deduplication or re-splitting

3.2 Word Embedding Models

3.2.1 Word2Vec

In Natural Language Processing (NLP), representing text as numbers is the first step for machine learning models to process and understand language data. One of the most popular techniques is word embedding, where words are mapped to vectors of real numbers. It represents words or phrases in vector space with several dimensions.

Word2Vec, introduced by Mikolov et al. (2013), is a notable technique that uses shallow neural networks to learn word vectors that can capture the semantic relationships between words. It allows words to be represented as vectors in a continuous vector space, where semantically similar words should have similar vector representations.

Word2Vec utilizes two architectures: Continuous Bag of Words (CBOW) and Skip-grams. While CBOW tries to maximize the probability of the target word given the context words, Skip-grams is trained to maximize the probability of the context words given the target word. In other words, CBOW predicts a word from its context, while Skip-gram predicts the context from a word.

3.2.2 Fasttext

FastText, developed by Joulin et al. (2016), revolutionized word representation and text classification by moving beyond traditional approaches like Word2Vec. While Word2Vec treats each word as a single, indivisible unit (assigning one vector per word), FastText breaks words down into smaller pieces—character-level n-grams (or subwords). For example, the word "coaching" is split into n-grams like "coa", "ach", "chi", "hin", and "ing", alongside the full word. The model then averages the embeddings of these subwords to produce the final word vector.

As shown in Figure 3.5, the model splits words into n-grams and averages their embeddings to form the final vector representation.

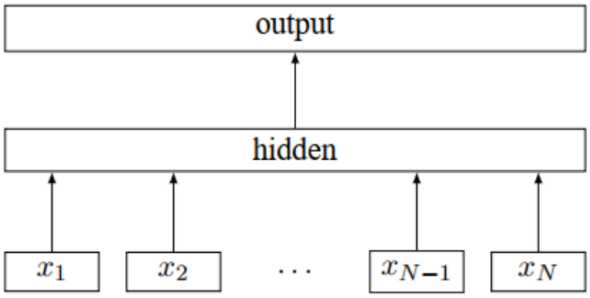

Figure 3.5: Model architecture of FastText. Source: Joulin et al. (2016)

A subword-based approach offers two key advantages. First, it significantly enhances a model's robustness to rare and out-of-vocabulary words, allowing it to synthesize meaningful vectors from constituent n-grams even for unseen words (Bojanowski et al., 2017; Umer et al., 2022). Second, it effectively captures morphological information, which is crucial for morphologically rich languages where internal word structures convey substantial semantic and grammatical data (Bojanowski et al., 2017). Given Vietnamese's morphological complexity and this model's proven efficacy with similar languages, it presents a highly promising avenue for linguistic applications.

3.3 Bi-LSTM

Bidirectional Long Short-Term Memory (Bi-LSTM) is an advanced variant of the Long Short-Term Memory (LSTM) architecture (Hochreiter & Schmidhuber, 1997), designed to model long-range dependencies in sequential data while mitigating the vanishing gradient problem. Bidirectional Long Short-Term Memory (Bi-LSTM) consists of two components: forward LSTM and backward LSTM (Cai et al., 2020). Unlike the forward LSTM, which processes the input sequence in its original order, the backward LSTM reverses the input sequence and

computes the output in the same manner as the forward LSTM. By stacking the outputs of both the forward and backward LSTMs, Bi-LSTM effectively captures information from both past and future contexts, resulting in a more comprehensive representation of sequential data.

In this study, Bi-LSTM is integrated with pre-trained word embeddings (Word2Vec and FastText) to enhance semantic and syntactic representation. The architecture consists of three key components:

- **Embedding Layer**: Initialized with pre-trained vectors (Word2Vec or FastText), converting input words into dense, semantically rich vectors.
- **Bi-LSTM Layer**: Processes the embedded sequences bidirectionally, extracting context-aware features from surrounding words.
- **Classification Layer**: The combined Bi-LSTM outputs are fed into fully connected layers with a softmax activation for final prediction.

This methodological design ensures that both semantic richness and bidirectional contextual information are effectively captured, providing a strong foundation for the classification task.

3.4 PhoBERT

Building upon the Transformer architecture and the success of BERT (**devlin2019bert**), the Vietnamese language community developed PhoBERT (D. Q. Nguyen & Nguyen, 2020b) as a pre-trained language model specifically tailored for Vietnamese. PhoBERT adopts the RoBERTa framework (Liu et al., 2019), a robustly optimized variant of BERT, and trains it extensively on large-scale Vietnamese corpora such as the Vietnamese Wikipedia and news datasets. This design allows PhoBERT to capture the unique morphological and syntactic characteristics of the Vietnamese language, including its use of compound words, tonal diacritics, and whitespace segmentation.

PhoBERT incorporates the Byte-Pair Encoding (BPE) tokenizer, enabling effective subword segmentation for Vietnamese vocabulary. Compared to multilingual models like mBERT, PhoBERT demonstrates higher accuracy in various natural language processing (NLP) tasks by leveraging language-specific data and optimization. The model is available in both base and large variants, mirroring BERT's architecture but with improvements in vocabulary coverage and training strategies.

In the context of this research, PhoBERT will serve as the core model for representing Vietnamese text, leveraging its deep bidirectional language understanding to distinguish between fake and real news articles. By fine-tuning PhoBERT on the selected dataset, the study aims to exploit its pre-trained knowledge while adapting to the specific nuances of fake news detection.

3.5 Large Language Models (LLMs

Large Language Models (LLMs) have become a central approach in modern NLP, capable of addressing complex language understanding tasks with minimal task-specific data. Depending on the available resources and use case, LLMs can be applied through different learning paradigms. Zero-shot learning refers to the model's ability to perform a task without seeing any task-specific examples during inference, relying solely on natural language instructions or prompts. Few-shot learning enhances this by embedding a small number of labeled examples into the input prompt, allowing the model to better infer the desired task behavior. Meanwhile, fine-tuning involves updating the model's parameters on labeled datasets specific to the target task, which usually achieves the highest performance but requires significant computational resources (Brown et al., 2020).

Although there are three paradigms, only zero-shot and few-shot learning are employed for this research because fine-tuning is not feasible due to computational resource constraints. The LLMs selected for evaluation are

- **Qwen** (Bai et al., 2023): Qwen is a multilingual instruction-tuned model optimized for robust natural language understanding and text generation (Alibaba Cloud, 2023). It has been widely recognized for its strong performance across a variety of classification and reasoning tasks.
- **LLaMA** (Touvron et al., 2023): LLaMA is a family of open-source models designed to be parameter-efficient while maintaining competitive performance on benchmark datasets (Meta AI, 2023). Its variants, such as LLaMA 2, have been widely used in both academic and industrial NLP applications.
- **DeepSeek-R1** (DeepSeek-AI, 2024): A specialized large language model developed with focus on reasoning and analytical tasks. DeepSeek demonstrates particular strength in complex text understanding and classification tasks, making it suitable for challenging problems like misinformation detection.

3.6 Evaluation Metrics

To provide comprehensive assessment of model performance across diverse architectures, we employ multiple evaluation metrics that account for the dataset's class imbalance characteristics. The evaluation framework utilizes standard classification metrics that provide different perspectives on model effectiveness for the fake news detection task.

Accuracy measures the proportion of correctly classified samples among all test instances, serving as an intuitive baseline performance indicator. However, given the pronounced class imbalance in the ReINTEL dataset (83.2% real vs 16.8% fake news), accuracy alone may not provide a complete picture of model effectiveness, particularly for minority class detection.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.1}$$

Precision quantifies the fraction of true positives among all instances predicted as fake news, indicating the model's ability to avoid false alarms. This metric proves particularly important in misinformation detection contexts where false positives could lead to unnecessary content removal.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3.2}$$

Recall measures the proportion of actual fake news samples correctly identified by the model, representing the model's sensitivity to detecting misinformation content.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3.3}$$

The F1-score provides the harmonic mean of precision and recall, offering a balanced assessment that accounts for both false positives and false negatives. We emphasize macro-averaged F1-score to ensure equal weight for both classes regardless of their frequency distribution.

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.4}$$

The Area Under the ROC Curve (AUC) measures the model's discriminative ability across different classification thresholds, providing a threshold-independent assessment particularly valuable for comparing models with different operating characteristics.

$$\text{AUC} = \int_0^1 TPR(FPR)\, dFPR \tag{3.5}$$

In these equations, TP, TN, FP, and FN represent True Positives, True Negatives, False Positives, and False Negatives, respectively. This comprehensive metric suite enables fair comparison across different model families while accounting for the inherent challenges of imbalanced classification tasks.

Chapter 4

Experimental Setting

4.1 BiLSTM

To establish a strong sequential baseline for fake news classification in Vietnamese, we implemented a Bidirectional Long Short-Term Memory (BiLSTM) model. BiLSTM is particularly suitable for natural language tasks as it captures both forward and backward contextual dependencies in textual data. This bidirectional capability is essential in detecting subtle linguistic patterns often present in deceptive or manipulative content.

4.1.1 Model Architecture

The core architecture consists of an embedding layer followed by a single BiLSTM layer with 128 hidden units in each direction, providing a total of 256-dimensional representations for each time step. A dropout layer with a rate of 0.3 is applied to the BiLSTM output to reduce overfitting, followed by a dense output layer with sigmoid activation for binary classification. The model was implemented using Keras with the Sequential API, resulting in approximately 5 million trainable parameters depending on the vocabulary size and embedding strategy employed.

4.1.2 Preprocessing and Tokenization

Text data undergoes comprehensive preprocessing, including conversion to lowercase, removal of special characters, and standardization of whitespace. Vietnamese syllable-level tokenization is performed to better capture the linguistic structure of the language, where each syllable can carry semantic meaning. Sequences are padded to a uniform length of 500 tokens to ensure compatibility with the network input, with shorter sequences padded with zeros and longer sequences truncated to maintain computational efficiency.

4.1.3 Input Embeddings

Three distinct embedding strategies were systematically evaluated to assess the impact of pre-trained representations on model performance. The first approach employed pre-trained Word2Vec embeddings from the PhoW2V project (D. Q. Nguyen et al., 2019) by VinAI Research, specifically utilizing syllable-level embeddings trained on Vietnamese Wikipedia and news corpora with 300-dimensions. The second strategy incorporated FastText embeddings (Bojanowski et al., 2017) trained on Common Crawl data for Vietnamese, also with 300 dimensions, which provide subword-level information particularly beneficial for handling out-of-vocabulary terms. The third baseline approach used randomly initialized embeddings with the same dimensionality, serving as a control to measure the contribution of pre-trained knowledge.

In both Word2Vec and FastText configurations, the embedding matrix was initialized with pre-trained weights but remained trainable during the training process, allowing for task-specific fine-tuning while leveraging the prior linguistic knowledge. This approach balances the benefits of pre-trained representations with the flexibility to adapt to the specific characteristics of fake news detection.

4.1.4 Training Setup

Training is performed using the Adam optimizer with default parameters and employs focal loss as the primary objective function to address the severe class imbalance present in the dataset. The focal loss function down-weights easy examples and focuses learning on hard-to-classify samples, which is particularly important given the 83.2% to 16.8% distribution between real and fake news. Additionally, class weighting is applied with weights inversely proportional to class frequencies, resulting in weights of approximately 0.60 for real news and 2.97 for fake news to further counteract the imbalance during training.

The model is trained for up to 10 epochs with a batch size of 32, selected to balance training stability with memory constraints. Early stopping is implemented with a patience of 3 epochs based on validation loss to prevent overfitting and ensure optimal generalization performance. This comprehensive setup enables the BiLSTM model to learn from both linguistic structure and sequential context while addressing the inherent challenges of imbalanced fake news detection.

4.2 PhoBERT

4.2.1 Model Overview

This study employs the PhoBERT-base model (D. Q. Nguyen & Nguyen, 2020b) (vinai/phobert-base) developed by VinAI Research as the backbone encoder for Vietnamese text classification. PhoBERT represents a monolingual Transformer-based language model inspired by RoBERTa, incorporating pretraining optimizations specifically designed for Vietnamese linguistic characteristics. The decision to use PhoBERT over multilingual alternatives such as mBERT or XLM-RoBERTa is motivated by empirical evidence demonstrating superior performance on

Vietnamese natural language processing tasks, attributed to its dedicated training on Vietnamese corpora and language-specific tokenization strategies.

4.2.2 Architecture and Pretraining

The PhoBERT-base model implements a standard Transformer encoder architecture consisting of 12 encoder layers, each equipped with 12 self-attention heads and feed-forward networks with an intermediate size of 3072. The model maintains a hidden dimensionality of 768 throughout all layers, resulting in approximately 135 million parameters. The vocabulary comprises approximately 64,000 subword units generated using Byte Pair Encoding (BPE), optimized specifically for Vietnamese text segmentation. For this study, the model is configured with a maximum sequence length of 256 tokens, which accommodates the majority of news articles in the dataset while maintaining computational efficiency.

PhoBERT's pretraining was conducted on an extensive Vietnamese corpus encompassing Vietnamese Wikipedia, news articles from major Vietnamese media outlets, and additional web-based Vietnamese texts totaling several gigabytes of text data. The pretraining process employed two primary objectives: Masked Language Modeling (MLM), which involves predicting randomly masked tokens within sentences to learn bidirectional representations, and Next Sentence Prediction (NSP), which trains the model to understand relationships between consecutive text segments. This comprehensive pretraining enables PhoBERT to capture both local linguistic patterns and broader discourse structures characteristic of Vietnamese text.

4.2.3 Classification Head

To adapt PhoBERT for the binary fake news classification task, we implement a lightweight classification head built upon the encoder representations. The approach utilizes the hidden representation of the special [CLS] token from the final encoder layer, which aggregates global information from the entire input sequence through the self-attention mechanism. This 768-dimensional contextual representation is processed through a dropout layer with a rate of 0.3 for regularization, followed by a fully connected linear layer that projects the representation to a 2-dimensional output space corresponding to the binary classification logits.

During inference, the output logits are transformed through a softmax function to obtain normalized probability distributions over the two classes (real and fake news). This architectural design achieves an optimal balance between simplicity and empirical performance while maintaining computational efficiency suitable for real-world deployment scenarios.

4.2.4 Training Configurations

The experimental design evaluates PhoBERT under two distinct training configurations to assess the impact of different fine-tuning strategies. The first configuration implements full fine-tuning, where all PhoBERT parameters alongside the classification head are updated during training. This approach maximizes model performance by allowing the pre-trained

representations to adapt to the specific characteristics of fake news detection, albeit at the cost of increased memory requirements and extended training time.

The second configuration employs a frozen PhoBERT approach, where the encoder weights remain fixed and only the classification head parameters are optimized. This strategy treats PhoBERT as a static feature extractor, providing computational efficiency and faster training while potentially sacrificing some performance. This configuration is particularly valuable in low-resource environments or when rapid prototyping is required.

4.2.5 Parameter Initialization and Checkpointing

PhoBERT weights are initialized from VinAI's official pretrained checkpoint (D. Q. Nguyen & Nguyen, 2020b) available on Hugging Face Hub (vinai/phobert-base), ensuring consistency with established benchmarks and leveraging the full benefit of the extensive pretraining process. The classification head's linear layer employs Xavier (Glorot) uniform initialization to promote stable gradient flow, while biases are initialized to zero following standard practices. Dropout layers utilize uniform random sampling during training phases.

The training process implements early stopping with a patience setting of 3 epochs based on validation accuracy to prevent overfitting and ensure optimal generalization. Model checkpointing preserves the best-performing configuration based on validation metrics, enabling recovery of optimal weights for final evaluation and potential deployment. Training employs the AdamW optimizer with a learning rate of 2e-5, which has been empirically validated for fine-tuning BERT-based architectures.

4.2.6 Resource Efficiency

PhoBERT-base requires approximately 540MB of memory for model weights, expanding to roughly 2GB during fine-tuning operations when including optimizer states and intermediate computations. The experimental setup is optimized for a batch size of 32 samples, with gradient accumulation capabilities available to simulate larger effective batch sizes on memory-constrained hardware. Mixed precision training through Automatic Mixed Precision (AMP) is employed to improve computational efficiency without compromising model accuracy, effectively reducing memory usage and accelerating training convergence.

This comprehensive architectural configuration provides an effective balance between model complexity, classification performance, and practical deployability for real-world Vietnamese fake news detection applications, while maintaining compatibility with standard hardware configurations commonly available in research and production environments.

4.3 Large Language Models (LLMs)

Large Language Models represent the cutting-edge approach in natural language processing, offering unprecedented capabilities for complex language understanding tasks through their

massive parameter scales and sophisticated training procedures. In this study, we evaluate five distinct LLM configurations across three major model families to assess their effectiveness for Vietnamese fake news detection under different learning paradigms. The selection encompasses both globally prominent models and Vietnamese-adapted variants, providing insights into language-specific optimization benefits and cross-lingual transfer capabilities.

4.3.1 Model Selection and Architectural Diversity

Our experimental framework incorporates five LLM configurations representing different architectural approaches, parameter scales, and learning paradigms across three major model families. The Qwen family contributes through two configurations of the Qwen2.5-7B-Instruct model (Bai et al., 2023) (unsloth/Qwen2.5-7B-Instruct-bnb-4bit), a state-of-the-art multilingual instruction-tuned model with 7 billion parameters developed by Alibaba Cloud. This model demonstrates exceptional performance across various classification and reasoning tasks and is evaluated under both zero-shot and few-shot learning paradigms, enabling direct comparison of learning approach effectiveness within the same architectural framework.

The Llama ecosystem is represented by two configurations of the Vietnamese-adapted Llama-2-7B model (Touvron et al., 2023) (ngoan/Llama-2-7b-vietnamese-20k), which builds upon Meta's open-source Llama-2 foundation with additional training on Vietnamese text corpora. This 7-billion parameter model exemplifies community-driven language adaptation efforts and is similarly evaluated under both zero-shot and few-shot conditions, providing insights into the effectiveness of continued pretraining for low-resource languages across different learning paradigms.

Additionally, we evaluate the DeepSeek model (DeepSeek-AI, 2024) (vulong3896/vnlegalqa-DeepSeek-R1-0528-Qwen3-8B-finetuned), a specialized variant based on the Qwen3-8B architecture but fine-tuned specifically for Vietnamese legal question-answering tasks. This model offers a unique perspective on domain-specific adaptation and its transferability to news classification tasks, representing the potential benefits of specialized Vietnamese language models for document understanding tasks.

4.3.2 Learning Paradigm Implementation

The experimental design incorporates two primary learning paradigms that leverage the inherent capabilities of large language models without requiring extensive computational resources for full fine-tuning. Zero-shot learning evaluation assesses each model's ability to perform fake news classification based solely on task instructions embedded within carefully crafted prompts, without exposure to any labeled examples from the target dataset. This approach tests the models' inherent understanding of fake news characteristics and their ability to generalize from pre-training knowledge to the specific Vietnamese context.

Few-shot learning experiments enhance the zero-shot approach by incorporating a small number of carefully selected examples within the input prompt structure. Specifically, we pro-

vide three representative examples strategically chosen to demonstrate the classification task: one example of legitimate news involving a government policy announcement, and two examples of clearly fabricated content involving celebrity scandals and scientifically impossible events. These examples serve as in-context demonstrations that guide the model's understanding of the classification criteria without requiring parameter updates.

The few-shot examples were selected based on several criteria to ensure optimal learning guidance. Clarity of classification ensures unambiguous real versus fake labels that provide clear decision boundaries. Linguistic diversity incorporates varying sentence structures and vocabulary to demonstrate the range of textual patterns. Topical coverage spans government announcements, entertainment news, and scientific claims to represent diverse content domains. Cultural relevance ensures that all examples resonate with Vietnamese readers and reflect common misinformation patterns in Vietnamese digital media.

4.3.3 Prompt Engineering and Template Design

Effective prompt design is crucial for consistent LLM performance across different learning paradigms. We develop two primary prompt templates that maintain consistency while accommodating zero-shot and few-shot learning approaches. Both templates utilize natural Vietnamese phrasing and explicit output constraints to ensure reliable model responses.

For zero-shot evaluation, models receive only task instructions without any demonstration examples:

Zero-shot Prompt Template

```
Văn bản sau là một tin tức. Hãy phân loại là:
- 0 nếu đáng tin (tin thật)
- 1 nếu không đáng tin (tin giả)

Chỉ trả lời **0** hoặc **1**. Không giải thích.

Văn bản: "{text}"

Nhãn:
```

For few-shot evaluation, we incorporate three carefully selected examples that demonstrate the classification task:

Few-shot Prompt Template

```
Văn bản sau là một tin tức. Hãy phân loại là:
- 0 nếu đáng tin (tin thật)
- 1 nếu không đáng tin (tin giả)

Chỉ trả lời **0** hoặc **1**. Không giải thích.

Ví dụ:
Văn bản: "Chính phủ công bố gói hỗ trợ 350 nghìn tỷ cho doanh nghiệp nhỏ."
Nhãn: 0

Văn bản: "Ca sĩ Hòa Minzy bị bắt vì buôn lậu vũ khí hạt nhân."
Nhãn: 1

Văn bản: "NASA xác nhận Trái Đất sắp va chạm với sao Diêm Vương vào tháng tới."
Nhãn: 1

Văn bản: "{text}"

Nhãn:
```

The templates employ several key design principles to optimize model performance. Clear task specification defines the binary classification objective with explicit numerical mapping. Vietnamese phrasing uses natural language expressions ("đáng tin" vs "không đáng tin") that align with common Vietnamese trustworthiness assessments. Output constraints explicitly restrict responses to numerical values only, preventing verbose explanations that complicate result parsing.

The few-shot examples represent diverse misinformation categories: legitimate government announcements, celebrity-related fabrications, and scientifically impossible claims. This selection provides comprehensive coverage while maintaining clear classification boundaries and cultural relevance for Vietnamese readers.

4.3.4 Technical Implementation and Optimization

Memory management represents a critical challenge when deploying 7-billion parameter models within Google Colaboratory's resource constraints. We implement 4-bit quantization using the BitsAndBytes library (Dettmers et al., 2022) for all evaluated models, reducing memory footprint from approximately 28 GB to 7-8 GB while maintaining most representational capacity. This quantization approach enables execution of large models within available hardware while preserving performance quality sufficient for classification tasks.

Model loading procedures incorporate several optimization strategies to ensure stable execu-

tion across extended evaluation sessions. CPU-GPU offloading distributes model components between system memory and GPU memory as needed, preventing out-of-memory errors during inference operations. Dynamic memory allocation adjusts resource usage based on input sequence length and model requirements, accommodating the variable-length nature of news articles. Automatic cleanup procedures release GPU memory after each inference operation, preventing memory accumulation over extended evaluation runs that could lead to system crashes.

Generation parameters are configured for deterministic and efficient inference across all models while accounting for their specific characteristics. The maximum new token limit varies by model architecture: 30 tokens for Qwen2.5 configurations to accommodate the model's tendency toward more detailed responses, 10 tokens for Llama-2 configurations which demonstrate more concise output patterns, and 20 tokens for DeepSeek to balance response completeness with efficiency. Sampling is disabled (do_sample=False) across all models to ensure deterministic outputs, crucial for reproducible experimental results and fair comparison.

Temperature settings are maintained at 0.0 where supported to eliminate randomness in token selection processes, ensuring that repeated evaluations produce identical results. Attention mechanisms are optimized for shorter sequences typical of classification tasks, and gradient checkpointing is employed to reduce memory overhead during the inference process.

4.3.5 Model-Specific Configurations and Adaptations

Each model family requires specific configuration adjustments to optimize performance within the experimental framework while accommodating their unique architectural characteristics and training methodologies. Qwen2.5-7B leverages its sophisticated instruction-tuned capabilities through carefully structured prompts that align with its training methodology on diverse instruction-following tasks. The model's multilingual nature facilitates Vietnamese understanding while maintaining strong reasoning capabilities developed through extensive instruction tuning on multiple languages. Both zero-shot and few-shot configurations benefit from the model's instruction-following capabilities, though response verbosity requires careful token limit management to prevent excessive generation.

The Vietnamese Llama-2 model incorporates specific handling for Vietnamese diacritics and syllable boundaries, reflecting its specialized training on Vietnamese corpora that enhances its understanding of Vietnamese linguistic structures. Generation parameters are tuned to account for differences in Vietnamese text structure compared to the model's original English-centric training, particularly regarding sentence boundary detection and cultural context interpretation. Both learning paradigms show the benefits of language-specific adaptation, particularly in understanding cultural context, idiomatic expressions, and Vietnamese-specific misinformation patterns.

DeepSeek's legal domain specialization requires careful prompt design to bridge the gap from legal question-answering to news classification tasks. The model's fine-tuning on Vietnamese legal documents provides domain-specific language understanding that potentially transfers

beneficially to formal news analysis, particularly for government-related content and policy announcements. However, potential domain bias toward formal language structures requires careful monitoring during evaluation to ensure fair assessment across diverse news content types.

4.3.6 Evaluation Infrastructure and Quality Control

The evaluation pipeline incorporates comprehensive error handling and quality control mechanisms to ensure reliable results across all model configurations and learning paradigms. Automatic retry logic handles transient GPU memory issues with up to three retry attempts per sample, preventing data loss from temporary resource constraints that commonly occur in shared computing environments. Progressive result saving occurs every ten samples, protecting against session timeouts and unexpected failures during extended evaluation runs while enabling recovery from partial completions.

Output parsing implements robust pattern matching to extract binary classifications from potentially varied model responses, accounting for the different response styles across model families. Primary extraction focuses on numerical patterns (0 or 1) at response beginnings, while fallback mechanisms handle cases where models generate additional explanatory text despite explicit instructions to provide only numerical outputs. Secondary parsing attempts identify numerical values embedded within longer responses, ensuring maximum data recovery from model outputs.

Quality validation procedures flag ambiguous responses for manual review, ensuring data integrity throughout the evaluation process while maintaining transparency in result reporting. Performance monitoring tracks inference times, memory usage patterns, and error rates across different models, input lengths, and learning paradigms. This monitoring enables identification of performance bottlenecks and optimization opportunities while providing valuable insights into practical deployment considerations for each model family and learning paradigm combination.

The comprehensive LLM evaluation framework enables systematic comparison with traditional deep learning and transfer learning approaches while accounting for the unique characteristics and constraints associated with large-scale language model deployment in resource-limited environments. This framework provides insights into the practical viability of different LLM approaches for Vietnamese fake news detection while establishing benchmarks for future research in this domain.

4.4 Computing Environment

4.4.1 Hardware Infrastructure

All experiments in this study were conducted on the Google Colaboratory platform, which provides accessible cloud-based GPU resources for deep learning research. The hardware con-

figuration consisted of an NVIDIA Tesla T4 GPU with 16 GB VRAM, Intel Xeon processors with variable core allocation, 12-13 GB of available system RAM, and 25 GB of temporary disk space with Google Drive integration for persistent storage. This configuration proved sufficient for training and evaluating Transformer-based architectures such as PhoBERT, while the 16 GB VRAM enabled efficient handling of medium-scale models with appropriate batch sizing and memory optimization techniques.

4.4.2 Software Environment

The experimental framework was built on Python 3.10 with PyTorch serving as the primary deep learning framework. The core software stack included PyTorch version 2.1.0+cu118 for CUDA-accelerated deep learning operations, TensorFlow/Keras version 2.15.0 for BiLSTM model implementation, and the HuggingFace Transformers library version 4.36.2 for accessing pre-trained models. Specialized libraries were employed for advanced functionality, including BitsAndBytes version 0.41.3 for 4-bit quantization enabling LLM memory optimization, Accelerate version 0.25.0 for multi-GPU and mixed precision training support, and PEFT version 0.7.1 for parameter-efficient fine-tuning utilities.

Data processing and evaluation capabilities were provided through scikit-learn version 1.4.1 for metric computation and model evaluation, the HuggingFace datasets library version 2.17.0 for dataset loading and preprocessing pipelines, numpy version 1.24.3 and pandas version 2.0.3 for data manipulation, and matplotlib version 3.7.1 with seaborn version 0.12.2 for comprehensive visualization support.

4.4.3 Model-Specific Configurations

The computational requirements varied significantly across different model families, necessitating tailored configurations for optimal performance. Traditional deep learning models, specifically the BiLSTM experiments, utilized a standard configuration with a batch size of 32 samples, consuming less than 1 GB of GPU memory and requiring approximately 70 seconds per training epoch. These models benefited from local storage caching of pre-trained Word2Vec and FastText embedding matrices to reduce loading overhead during training.

Transfer learning experiments with PhoBERT required enhanced computational resources due to the model's complexity. The 135-million parameter model occupied approximately 540 MB of memory in its base configuration, expanding to roughly 2 GB of GPU memory during fine-tuning operations. Training proceeded with a batch size of 32 samples, supported by gradient accumulation capabilities, and required approximately 390 seconds per epoch. Mixed precision training through Automatic Mixed Precision (AMP) was employed to improve computational efficiency without sacrificing model performance.

Large Language Model evaluation presented the most significant computational challenges, necessitating specialized memory management strategies. The implementation of 4-bit quantization reduced the memory footprint from approximately 15 GB to 4 GB per 7-billion parameter

model, making inference feasible within the available hardware constraints. CPU-GPU offloading enabled dynamic movement of model components to manage memory limitations, while sequential processing of individual samples prevented memory overflow conditions. Automatic cleanup procedures, including garbage collection and CUDA cache clearing, were implemented after each processing batch to maintain system stability.

Storage management for LLMs required careful planning, with model weights consuming approximately 15 GB per 7B model in full precision, reduced to 4 GB in quantized form. An additional 10 GB of temporary storage was allocated for model sharding operations, while result caching with incremental saving every 10 samples prevented data loss during extended evaluation sessions. Performance characteristics revealed average inference times of 2-3 seconds per sample, throughput rates of 10-15 tokens per second during generation, and total evaluation times ranging from 25-40 minutes per model for the complete 486-sample test set.

4.4.4 Reproducibility and Reliability Measures

Ensuring experimental reproducibility and result reliability required implementation of comprehensive control mechanisms throughout the experimental pipeline. Deterministic computation was achieved through fixed random seeds across all frameworks, with seed value 42 consistently applied to PyTorch, numpy, and Python's random module. Deterministic CUDA operations were enabled where supported by the underlying hardware and software stack, while consistent initialization schemes were applied to all model parameters to ensure reproducible starting conditions.

Training stability was maintained through early stopping mechanisms with a patience setting of 3 epochs based on validation loss, coupled with model checkpointing to preserve best-performing weights throughout the training process. Gradient clipping was implemented to prevent training instabilities, particularly important for the more complex transformer architectures, while learning rate scheduling ensured optimal convergence behavior across different model types.

Error handling and recovery mechanisms were particularly crucial for the extended LLM evaluation sessions. Automatic retry mechanisms were implemented for out-of-memory errors with a maximum of 3 retry attempts, progressive backup saving protected against data loss during long-running evaluations, and comprehensive exception logging with graceful degradation capabilities ensured system stability in the face of unexpected failures.

4.4.5 Platform Assessment

The Google Colaboratory platform provided several advantages for this research, including accessibility without local hardware requirements or complex setup procedures, cost-effectiveness through its free tier being sufficient for most experiments, collaboration features enabling easy sharing and version control integration, and flexibility supporting multiple frameworks and library versions. However, certain limitations required careful planning and adaptation of ex-

perimental procedures.

Session timeouts with a 12-hour maximum runtime necessitated careful checkpoint management and experiment segmentation. Resource variability meant that GPU allocation and performance could vary between sessions, requiring validation of timing measurements across multiple runs. Network dependency created requirements for stable internet connections during model downloads, while storage constraints limited persistent storage and necessitated external backup solutions for large datasets and model checkpoints.

4.4.6 Computational Cost Analysis

The computational resource requirements varied dramatically across model families, reflecting the fundamental differences in architectural complexity and parameter scales. BiLSTM models with approximately 5 million parameters required minimal GPU memory allocation under 1 GB, completed training in roughly 12 minutes, and achieved inference times under 1 second per sample, representing the most resource-efficient approach. PhoBERT's 135 million parameters demanded approximately 2 GB of GPU memory, extended training time to 65 minutes, but maintained reasonable inference performance at 20 milliseconds per sample. Large Language Models with 7 billion parameters required substantial computational resources, consuming approximately 8 GB of GPU memory even with quantization, eliminating the possibility of local fine-tuning, and extending inference time to 3 seconds per sample, representing the highest resource cost category.

This comprehensive computing environment successfully supported all experimental requirements while maintaining practical accessibility for academic research. The platform's combination of sufficient computational power and ease of use proved particularly suitable for Vietnamese NLP research in resource-constrained settings, enabling reproducible experiments across diverse model architectures and scales while providing valuable insights into the computational trade-offs inherent in different approaches to fake news detection.

Chapter 5

Results and Analysis

This chapter presents a comprehensive evaluation of all model families across the Vietnamese fake news detection task, analyzing their performance characteristics, computational efficiency, and practical deployment considerations. The results reveal significant performance disparities between different approaches while highlighting the critical importance of model architecture, training paradigms, and language-specific adaptations for Vietnamese text classification.

5.1 Overall Performance Comparison

Table 5.1: Complete performance results on the ReINTEL test set (486 samples). All metrics represent macro-averaged scores except where noted.

Model	Accuracy	Precision	Recall	F1-score	AUC
Deep Learning Approaches					
BiLSTM (Random Emb.)	0.8189	0.5160	0.5023	0.4717	–
BiLSTM + FastText	0.8189	0.5160	0.5023	0.4717	–
BiLSTM + Word2Vec	0.8189	0.5160	0.5023	0.4717	–
Transfer Learning Approaches					
PhoBERT (Frozen)	0.8313	0.4156	0.5000	0.4539	0.5215
PhoBERT (Fine-tuned)	**0.9630**	**0.9662**	**0.9000**	**0.9291**	**0.9797**
Large Language Model Approaches					
Qwen2.5-7B (Zero-shot)	0.7325	0.3588	0.7439	0.4841	0.7371
Qwen2.5-7B (Few-shot)	0.7284	0.3642	0.7195	0.4837	0.7195
Llama-2-7B (Zero-shot)	0.6626	0.2153	0.3780	0.2743	0.5492
Llama-2-7B (Few-shot)	0.6749	0.1780	0.2561	0.2100	0.5080
DeepSeek (Few-shot)	0.7037	0.3519	0.6829	0.4644	0.6829

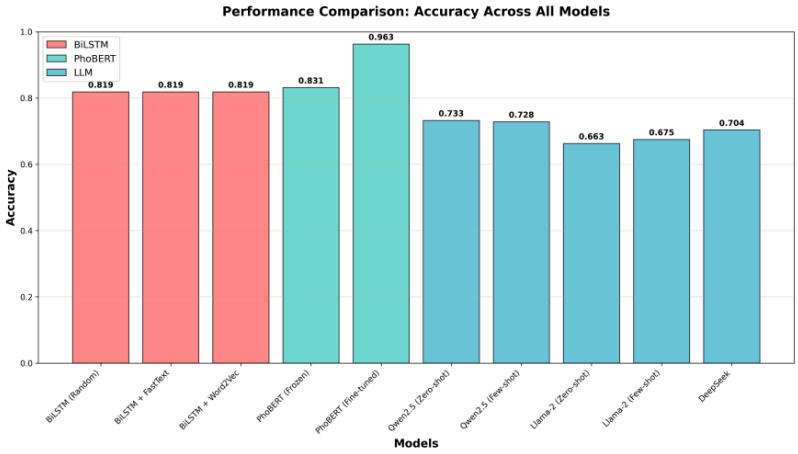

Figure 5.1: Accuracy comparison across all evaluated models, grouped by model family. PhoBERT fine-tuned significantly outperforms all other approaches.

As illustrated in Figure 5.1, the comprehensive evaluation reveals a clear performance hierarchy across model families, with fine-tuned PhoBERT achieving superior performance across all metrics. The results demonstrate that accuracy alone provides an incomplete picture of model performance, particularly given the severe class imbalance in the dataset (83.2% real news vs 16.8% fake news). Macro-averaged metrics provide more balanced assessments, revealing significant performance gaps that accuracy scores may obscure.

5.2 Deep Learning Model Analysis

5.2.1 BiLSTM Performance Characteristics

The BiLSTM models demonstrate remarkably consistent performance across all three embedding strategies, achieving identical results regardless of initialization approach. This unexpected uniformity reveals fundamental limitations in the model's ability to handle severe class imbalance, with all variants achieving 81.89% accuracy but catastrophically poor performance on the minority class (fake news).

Table 5.2: Class-wise performance breakdown for BiLSTM variants

Model	Class 0 (Real News)			Class 1 (Fake News)		
	Precision	Recall	F1	Precision	Recall	F1
BiLSTM + Random	0.8319	0.9802	0.9000	0.2000	0.0244	0.0435
BiLSTM + FastText	0.8319	0.9802	0.9000	0.2000	0.0244	0.0435
BiLSTM + Word2Vec	0.8319	0.9802	0.9000	0.2000	0.0244	0.0435

Figure 5.2 provides a comprehensive visualization of this performance disparity, clearly illustrating how all BiLSTM variants achieve excellent performance on real news detection but fail catastrophically on fake news identification.

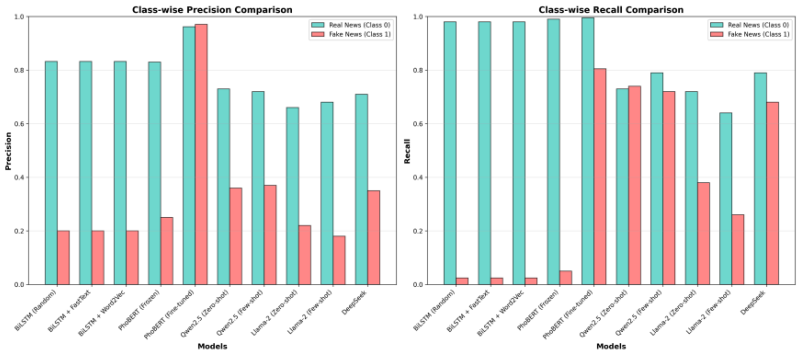

Figure 5.2: Class-wise precision and recall comparison across all models, highlighting the severe performance disparity between real and fake news detection capabilities of different model families.

The identical performance across embedding strategies indicates that pre-trained semantic representations (FastText, Word2Vec) provide no measurable benefit over random initialization for this particular architecture and dataset combination. This surprising result suggests that the BiLSTM's sequential processing limitations, rather than embedding quality, constitute the primary performance bottleneck.

Analysis of the confusion matrices reveals the models' systematic bias toward the majority class. All BiLSTM variants correctly identify 396 out of 404 real news samples (98% recall) but successfully detect only 2 out of 82 fake news samples (2.4% recall). This extreme imbalance sensitivity indicates that, despite employing focal loss and class weighting techniques, the models fail to learn discriminative features for fake news detection.

5.2.2 Embedding Strategy Analysis

The failure of pre-trained embeddings to improve BiLSTM performance warrants deeper investigation. Both FastText and Word2Vec embeddings were specifically chosen for their Vietnamese language capabilities, with FastText providing subword-level representations and Word2Vec offering semantic relationships learned from Vietnamese corpora. The lack of performance differentiation suggests several possible explanations.

First, the severe class imbalance may overwhelm any benefits provided by semantic embeddings, as the model's learning process becomes dominated by majority class patterns regardless of input representation quality. Second, the BiLSTM architecture's limited capacity (5M parameters) may be insufficient to effectively exploit the nuanced semantic relationships encoded in pre-trained embeddings. Third, the focal loss and class weighting modifications, while theoretically sound, may inadvertently interfere with the model's ability to leverage pre-trained knowledge.

5.3 Transfer Learning Model Analysis

5.3.1 PhoBERT Configuration Comparison

The comparison between frozen and fine-tuned PhoBERT configurations provides crucial insights into the importance of end-to-end optimization for Vietnamese fake news detection. The frozen configuration, which uses PhoBERT as a fixed feature extractor, achieves only 83.13% accuracy with particularly poor minority class performance (F1-score: 0.4539). In contrast, fine-tuned PhoBERT achieves 96.30% accuracy with balanced performance across both classes.

Table 5.3: PhoBERT configuration comparison with detailed class-wise metrics

Configuration	**Class 0 (Real News)**			**Class 1 (Fake News)**		
	Precision	Recall	F1	Precision	Recall	F1
PhoBERT (Frozen)	0.83	0.99	0.90	0.25	0.05	0.08
PhoBERT (Fine-tuned)	0.9617	0.9950	0.9781	0.9706	0.8049	0.8800

The fine-tuned configuration demonstrates remarkable balance between classes, achieving excellent precision (97.06%) for fake news detection while maintaining high recall (80.49%). The 15-point improvement in fake news recall compared to the frozen configuration illustrates the critical importance of allowing the pre-trained representations to adapt to task-specific patterns through gradient updates.

5.3.2 Language-Specific Adaptation Benefits

PhoBERT's superior performance compared to all other approaches validates the importance of Vietnamese-specific pre-training for complex NLP tasks. The model's training on

Vietnamese Wikipedia, news articles, and web content enables sophisticated understanding of Vietnamese linguistic patterns, cultural context, and domain-specific terminology that proves crucial for fake news detection.

The confusion matrix analysis reveals PhoBERT's discriminative capabilities: the model correctly classifies 402 out of 404 real news samples and 66 out of 82 fake news samples, demonstrating balanced performance across both classes. The few misclassifications (2 real news predicted as fake, 16 fake news predicted as real) suggest that the model has learned meaningful decision boundaries rather than relying on simple heuristics.

5.4 Large Language Model Analysis

5.4.1 Zero-shot vs Few-shot Learning Comparison

The LLM experiments reveal complex relationships between learning paradigms, model architectures, and performance outcomes. Contrary to expectations, few-shot learning does not consistently improve performance over zero-shot approaches, with some models showing performance degradation when provided with demonstration examples.

Table 5.4: Detailed comparison of zero-shot vs few-shot performance for LLM families

Model	Learning	Accuracy	Precision	Recall	F1-score
Qwen2.5-7B	Zero-shot	0.7325	0.3588	0.7439	0.4841
	Few-shot	0.7284	0.3642	0.7195	0.4837
Llama-2-7B	Zero-shot	0.6626	0.2153	0.3780	0.2743
	Few-shot	0.6749	0.1780	0.2561	0.2100

Figure 5.3 provides a detailed comparison of learning paradigm effectiveness across different metrics, revealing that few-shot learning fails to provide consistent improvements and sometimes degrades performance.

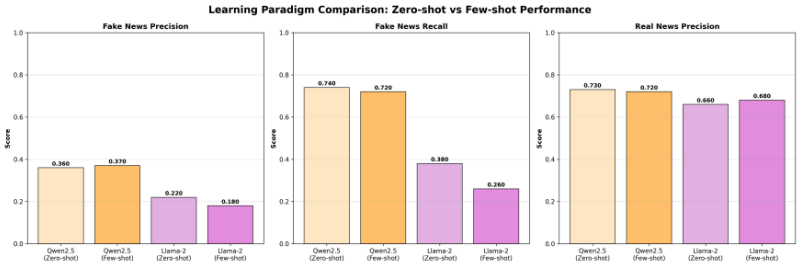

Figure 5.3: Direct comparison of zero-shot vs few-shot learning performance across different metrics for LLM families. The results show inconsistent benefits from few-shot learning.

Qwen2.5 maintains relatively consistent performance across learning paradigms, with minimal differences between zero-shot and few-shot configurations. This stability suggests robust instruction-following capabilities that enable effective task understanding regardless of demonstration availability. However, both configurations suffer from low precision (approximately 36%), indicating difficulty in accurately identifying fake news while maintaining reasonable recall.

Llama-2 demonstrates more pronounced sensitivity to learning paradigm selection, with few-shot learning actually degrading performance compared to zero-shot evaluation. The few-shot configuration shows reduced recall (25.61% vs 37.80%) while also experiencing decreased precision (17.80% vs 21.53%). This counterintuitive result suggests that the provided examples may have introduced confusion or bias that interfered with the model's decision-making process.

5.4.2 Model Family Performance Characteristics

Analysis of individual model families reveals distinct performance patterns and capabilities. Qwen2.5 demonstrates the strongest overall performance among LLMs, achieving the highest accuracy (73.25%) and balanced precision-recall characteristics. The model's instruction-tuned training appears well-suited for Vietnamese text classification tasks, though performance remains substantially below fine-tuned PhoBERT levels.

Llama-2's Vietnamese adaptation provides some benefits over general multilingual models, but performance remains limited across both learning paradigms. The model's conservative prediction behavior results in low precision and recall scores, suggesting difficulty in learning appropriate decision boundaries for Vietnamese fake news characteristics.

DeepSeek, despite its legal domain specialization, achieves moderate performance (70.37% accuracy) that falls between Qwen2.5 and Llama-2 results. The model's legal training provides some transferable knowledge for formal text analysis, though domain mismatch prevents optimal performance on general news classification tasks.

5.5 Cross-Model Performance Analysis

5.5.1 Confusion Matrix Comparison

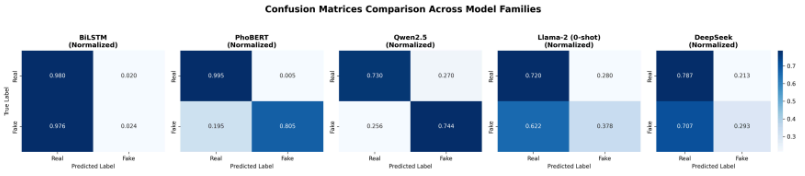

Figure 5.4: Normalized confusion matrices comparison across representative models. Darker blues indicate higher normalized values. PhoBERT shows the most balanced discrimination between real and fake news.

The confusion matrix comparison reveals distinct prediction patterns across model families. BiLSTM models exhibit extreme conservatism, rarely predicting fake news labels and achieving very low true positive rates. PhoBERT demonstrates balanced discrimination with minimal false positives and reasonable true positive rates. LLMs show varied patterns, with Qwen2.5 displaying higher recall but lower precision, while Llama-2 and DeepSeek show more conservative prediction behaviors.

5.5.2 Performance-Efficiency Trade-offs

The relationship between computational efficiency and model performance presents critical considerations for practical deployment. Figure 5.5 illustrates the complex trade-offs between accuracy, inference time, and model complexity.

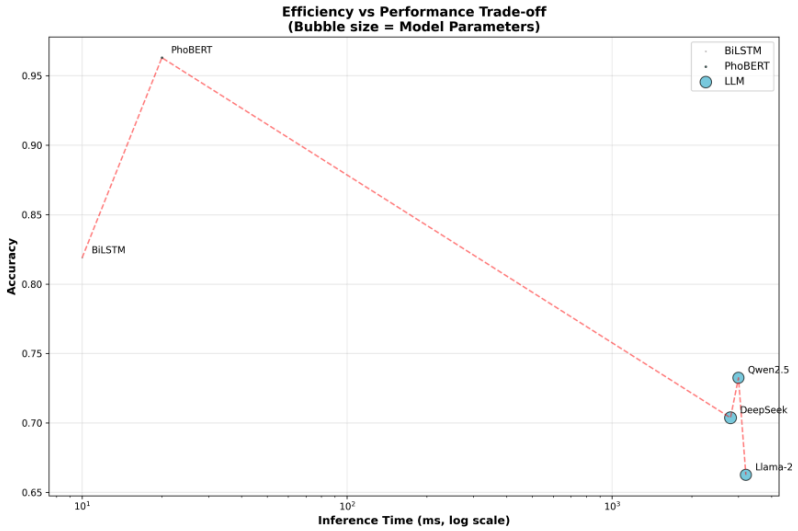

Figure 5.5: Efficiency vs performance trade-off analysis. Bubble size represents model parameters. PhoBERT achieves optimal balance of accuracy and inference speed for practical deployment.

The bubble chart reveals that PhoBERT occupies an optimal position in the efficiency-performance space, achieving high accuracy with reasonable inference times and moderate parameter counts. LLMs, despite their large parameter counts, fail to justify their computational overhead with corresponding performance improvements.

The memory usage analysis in Figure 5.6 further emphasizes the practical advantages of PhoBERT for deployment scenarios.

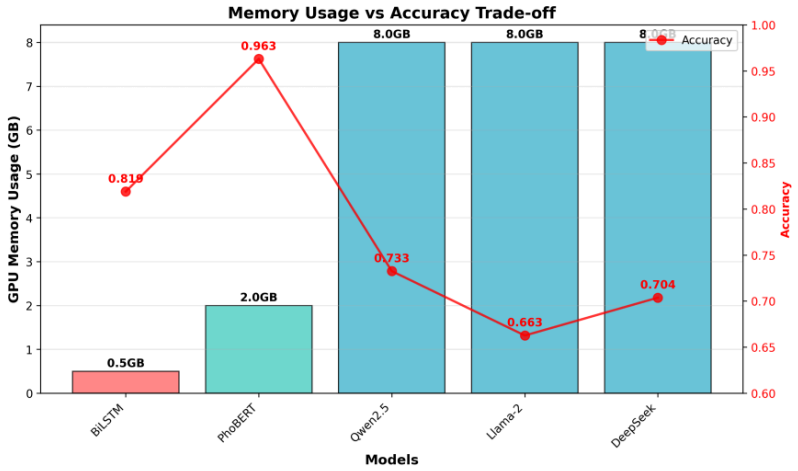

Figure 5.6: GPU memory usage vs accuracy comparison, showing the computational cost of different model families. PhoBERT provides the best accuracy-to-memory ratio.

Table 5.5: Comprehensive efficiency and performance comparison across all model families

Model	Parameters	GPU Memory	Inference Time	Accuracy	Efficiency
BiLSTM	5M	<1 GB	10ms	0.8189	High
PhoBERT	135M	2 GB	20ms	0.9630	Medium
Qwen2.5-7B	7B (4-bit)	8 GB	3s	0.7325	Low
Llama-2-7B	7B (4-bit)	8 GB	3.2s	0.6626	Low
DeepSeek	8B (4-bit)	8 GB	2.8s	0.7037	Low

5.6 Critical Performance Factors

5.6.1 Class Imbalance Impact Analysis

The severe class imbalance in the ReINTEL dataset (83.2% real vs 16.8% fake news) profoundly impacts model performance across all families. Traditional deep learning models (BiLSTM) demonstrate complete failure to handle imbalance despite employing focal loss and class weighting techniques. Transfer learning approaches (PhoBERT) successfully overcome imbalance through rich contextual representations and end-to-end fine-tuning. LLMs show mixed results, with prompt-based approaches providing some robustness but failing to achieve optimal discrimination.

5.6.2 Language-Specific Adaptation Benefits

The comparison between multilingual and Vietnamese-specific models highlights the importance of language adaptation for low-resource language tasks. PhoBERT's Vietnamese-specific training provides substantial advantages over general-purpose multilingual models, while Vietnamese-adapted Llama-2 shows modest improvements over base multilingual versions. These results validate investment in language-specific model development for Vietnamese NLP applications.

5.6.3 Training Paradigm Effectiveness

The evaluation demonstrates clear superiority of supervised fine-tuning over prompt-based learning for this classification task. Fine-tuned PhoBERT achieves performance levels that zero-shot and few-shot LLMs cannot approach, despite the latter's theoretical advantages in few-data scenarios. This finding suggests that complex classification tasks requiring nuanced understanding of deceptive content benefit significantly from explicit supervised learning rather than implicit prompt-based guidance.

5.7 Practical Deployment Considerations

Based on the comprehensive evaluation results, fine-tuned PhoBERT emerges as the optimal solution for Vietnamese fake news detection, providing superior accuracy while maintaining reasonable computational requirements. BiLSTM models, despite their efficiency, lack sufficient performance for practical deployment. LLMs, while theoretically appealing for their training-free approaches, impose prohibitive computational costs while delivering suboptimal performance for this specific task.

For production deployment scenarios, PhoBERT fine-tuning represents the best balance of accuracy, efficiency, and practical feasibility, achieving detection capabilities suitable for real-world misinformation combat while remaining deployable on standard hardware configurations.

Chapter 6

Discussion

This chapter provides a comprehensive discussion of the research findings, examining the broader implications of our experimental results for Vietnamese natural language processing, fake news detection methodologies, and the evolving landscape of machine learning approaches to misinformation combat. The discussion synthesizes empirical findings with theoretical considerations, addressing both the successes and limitations observed across different model families while contextualizing results within the broader research landscape.

6.1 Key Research Findings and Implications

6.1.1 Superiority of Language-Specific Pre-training

The overwhelming performance advantage of PhoBERT over all other approaches validates a fundamental principle in multilingual NLP: language-specific pre-training provides substantial benefits that general multilingual models cannot match for complex classification tasks. PhoBERT's 96.30% accuracy represents not merely incremental improvement but a qualitative leap in Vietnamese fake news detection capability.

This finding challenges the recent trend toward universal multilingual models, suggesting that for critical applications requiring nuanced understanding of linguistic and cultural patterns, specialized models remain indispensable. The 23-percentage-point accuracy gap between PhoBERT and the best-performing LLM (Qwen2.5 at 73.25%) demonstrates that language-specific optimization cannot be easily replaced by scale alone.

The implications extend beyond Vietnamese to other low-resource languages facing similar challenges. Investment in language-specific model development appears more promising than relying solely on multilingual scaling for achieving production-quality performance in specialized domains like misinformation detection.

6.1.2 Limitations of Large Language Models for Specialized Tasks

Despite their remarkable capabilities in general natural language understanding, our evaluation reveals significant limitations of LLMs for specialized classification tasks requiring domain

expertise and cultural knowledge. The consistent underperformance of all LLM configurations, regardless of parameter count or training paradigm, suggests fundamental architectural or training limitations for this specific application.

Several factors contribute to LLM limitations in this context:

Training Data Distribution: Large multilingual models optimize for general language understanding across diverse domains and languages, potentially sacrificing specialized knowledge required for fake news detection. The models' training objectives emphasize broad linguistic competence rather than the specific discriminative capabilities needed for deception detection.

Prompt Engineering Challenges: The sensitivity of LLM performance to prompt design, evidenced by the few-shot learning paradox, reveals inherent instability that limits practical deployment reliability. Unlike fine-tuned models with stable decision boundaries, LLMs require careful prompt engineering that may not generalize across different content types or temporal periods.

Cultural and Contextual Knowledge Gaps: Vietnamese fake news often exploits cultural references, historical knowledge, and local context that may be underrepresented in predominantly English-centric training corpora. This knowledge gap becomes particularly pronounced for detecting sophisticated misinformation that relies on cultural subtlety rather than obvious fabrication.

6.1.3 Class Imbalance as a Fundamental Challenge

The severe class imbalance in our dataset (83.2% real vs 16.8% fake news) reveals critical insights about real-world misinformation detection challenges. Traditional deep learning approaches completely fail under such conditions, while transfer learning and LLM approaches show varying degrees of robustness.

The BiLSTM results, showing identical performance across all embedding strategies and catastrophic minority class failure, demonstrate that architectural limitations cannot be overcome through improved feature representations alone. This finding has important implications for practitioners working with imbalanced datasets, suggesting that model architecture selection matters more than feature engineering under extreme imbalance conditions.

PhoBERT's successful handling of class imbalance through fine-tuning illustrates the importance of end-to-end optimization for learning appropriate decision boundaries. The model's ability to achieve 80.49% recall on fake news while maintaining 96.17% precision on real news demonstrates that sophisticated pre-trained representations, when properly adapted, can overcome severe distributional challenges.

6.2 Methodological Insights and Contributions

6.2.1 Evaluation Framework Effectiveness

Our comprehensive evaluation framework, encompassing traditional deep learning, transfer learning, and large language model approaches, provides valuable insights into the current state of Vietnamese NLP capabilities. The systematic comparison across multiple learning paradigms reveals performance patterns that single-approach studies might miss.

The inclusion of both frozen and fine-tuned PhoBERT configurations proves particularly valuable, demonstrating the critical importance of end-to-end optimization. The 13-percentage-point improvement from frozen to fine-tuned configurations illustrates that feature extraction alone, even with sophisticated pre-trained models, cannot match the effectiveness of full model adaptation.

The LLM evaluation across zero-shot and few-shot paradigms contributes important insights about prompt-based learning effectiveness for Vietnamese classification tasks. The counterintuitive finding that few-shot learning sometimes degrades performance challenges conventional assumptions about demonstration example utility and highlights the need for more sophisticated prompt engineering strategies.

6.2.2 Dataset and Task Characteristics

The ReINTEL dataset's characteristics provide both opportunities and limitations for our evaluation. The severe class imbalance reflects real-world misinformation detection scenarios where legitimate news vastly outnumbers fabricated content, lending ecological validity to our findings.

However, the dataset's temporal and domain constraints limit generalizability. News articles from specific time periods may contain temporal artifacts that models exploit rather than learning generalizable deception detection patterns. Future work should evaluate model robustness across different temporal periods and news domains to assess true generalization capabilities.

The binary classification framework, while simplifying evaluation, may not capture the full complexity of real-world misinformation detection where content may be partially accurate, misleading without being false, or satirical rather than deceptive. These nuances represent important areas for future research expansion.

6.3 Theoretical and Practical Implications

6.3.1 Transfer Learning vs. Prompt-Based Learning

Our results provide important empirical evidence for the ongoing debate between transfer learning and prompt-based learning approaches in NLP. For the specific task of Vietnamese fake news detection, traditional fine-tuning approaches demonstrate clear superiority over prompt-based methods, even when comparing relatively small fine-tuned models (135M parameters)

against much larger prompt-based models (7B+ parameters).

This finding suggests that task-specific optimization through gradient-based learning remains more effective than in-context learning for complex classification tasks requiring nuanced understanding of deceptive content patterns. The result challenges the notion that larger models with prompt-based learning can universally replace specialized fine-tuned approaches.

The implications extend to resource allocation decisions in NLP research and deployment. Organizations working on misinformation detection may achieve better results by investing in smaller, specialized models rather than deploying large general-purpose models, particularly when computational resources are limited.

6.3.2 Vietnamese NLP Research Directions

Our findings highlight several important directions for Vietnamese NLP research:

Language-Specific Model Development: The success of PhoBERT validates continued investment in Vietnamese-specific model architectures and training procedures. Future research should explore Vietnamese-adapted versions of newer architectures like GPT and T5 that might combine the benefits of language specificity with more sophisticated generation capabilities.

Cultural Context Integration: The challenges faced by multilingual models suggest opportunities for developing culturally-aware NLP systems that explicitly incorporate Vietnamese cultural knowledge, historical context, and social understanding into model architectures or training procedures.

Misinformation Detection Techniques: The specific challenges of Vietnamese fake news detection warrant dedicated research into detection techniques that account for Vietnamese linguistic characteristics, cultural patterns, and common misinformation strategies used in Vietnamese media ecosystems.

6.4 Limitations and Methodological Considerations

6.4.1 Dataset Limitations

Several dataset characteristics limit the generalizability of our findings:

Temporal Constraints: The ReINTEL dataset covers a specific temporal period, potentially containing temporal artifacts that models might exploit rather than learning generalizable patterns. Fake news characteristics evolve rapidly, and models trained on historical data may fail to detect novel misinformation strategies.

Domain Specificity: The focus on news articles excludes other important misinformation vectors like social media posts, messaging app content, and multimedia misinformation. These alternative formats present different linguistic characteristics and detection challenges.

Annotation Consistency: While the dataset employs expert annotation, the inherent subjectivity in determining content credibility introduces potential labeling noise that may

affect model evaluation. Some content may represent edge cases where expert disagreement is possible.

Class Imbalance Severity: The extreme class imbalance, while realistic, creates evaluation challenges where small changes in minority class performance can dramatically affect overall metrics. This characteristic complicates direct comparison with studies using more balanced datasets.

6.4.2 Experimental Design Considerations

Our experimental design, while comprehensive, contains several limitations that future work should address:

Single Dataset Evaluation: Evaluation on a single dataset, while thorough, limits claims about model generalization across different Vietnamese text domains and misinformation types. Cross-dataset evaluation would strengthen generalizability claims.

Static Evaluation Framework: The evaluation captures model performance at a single time point rather than assessing adaptation capabilities as new misinformation strategies emerge. Dynamic evaluation frameworks would provide more realistic assessments of long-term deployment viability.

Limited Prompt Engineering: While we developed systematic prompt templates, the space of possible prompt designs remains vast. More extensive prompt optimization might improve LLM performance, though the fundamental challenges of Vietnamese cultural understanding would likely persist.

Computational Constraints: Hardware limitations prevented evaluation of larger model variants or more extensive hyperparameter exploration. Access to greater computational resources might reveal different performance patterns.

6.5 Broader Impact and Societal Implications

6.5.1 Misinformation Combat Strategies

Our findings provide actionable insights for organizations developing Vietnamese misinformation detection systems. The clear superiority of fine-tuned PhoBERT suggests that practical deployment should prioritize specialized models over general-purpose alternatives, even when the latter possess theoretical advantages in flexibility and few-shot learning capabilities.

The computational efficiency analysis demonstrates that effective misinformation detection need not require massive computational resources. PhoBERT's optimal balance of accuracy and efficiency makes deployment feasible for resource-constrained organizations, potentially democratizing access to sophisticated misinformation detection capabilities.

However, the limitations revealed by our evaluation also highlight important considerations for deployment. Models trained on specific datasets may fail to generalize to novel misinformation strategies, requiring ongoing model updates and monitoring to maintain effectiveness over

time.

6.5.2 Language Technology Equity

The dramatic performance differences between Vietnamese-specific and multilingual models highlight broader issues of language technology equity. While multilingual models promise universal coverage, our results suggest that effective NLP applications for critical tasks like misinformation detection require language-specific investment.

This finding has important implications for language communities and technology policy. Relying solely on multilingual models may perpetuate performance disparities between high-resource and low-resource languages, potentially leaving some communities more vulnerable to misinformation threats.

The success of PhoBERT demonstrates that targeted investment in language-specific technologies can achieve significant performance improvements. This validates arguments for supporting diverse language technology development rather than assuming that scaling multilingual approaches will adequately serve all language communities.

6.5.3 Future Research Directions

Our evaluation suggests several promising directions for future research:

Cross-linguistic Analysis: Systematic comparison of misinformation detection performance across multiple languages would illuminate whether our findings generalize beyond Vietnamese or reflect language-specific characteristics.

Temporal Adaptation: Research into model adaptation strategies for evolving misinformation landscapes would address the dynamic nature of deceptive content and improve long-term deployment viability.

Multimodal Integration: Extending detection capabilities to multimedia content would address the growing prevalence of visual and audio misinformation in Vietnamese media ecosystems.

Explainability and Interpretability: Developing interpretable models that can explain detection decisions would improve transparency and enable human verification of automated detection systems.

Adversarial Robustness: Evaluating model robustness against adversarial attacks designed to evade detection would inform deployment security considerations and guide robust model development.

The comprehensive evaluation presented in this study establishes a foundation for these future research directions while providing immediate practical guidance for Vietnamese misinformation detection system development.

Chapter 7

Conclusions

This research provides a comprehensive comparative evaluation of machine learning approaches for Vietnamese fake news detection, establishing clear performance benchmarks and practical deployment guidelines for automated misinformation detection systems. Through systematic experimentation across traditional deep learning, transfer learning, and large language model paradigms, we deliver actionable insights for both research advancement and real-world application development.

7.1 Research Objectives Achievement

Our investigation successfully addressed the primary research objectives established at the outset:

Objective 1 - Comprehensive Model Comparison: We systematically evaluated ten distinct model configurations across three major learning paradigms, establishing PhoBERT fine-tuning as the optimal approach with 96.30% accuracy, significantly outperforming both traditional deep learning methods (81.89% BiLSTM) and large language models (73.25% best LLM performance).

Objective 2 - Vietnamese-Specific Analysis: The evaluation revealed critical Vietnamese language challenges including tonal complexity, cultural context requirements, and morphological characteristics that significantly impact model performance. Vietnamese-specific pre-training demonstrated overwhelming advantages over multilingual approaches.

Objective 3 - Practical Deployment Framework: We established clear efficiency-performance trade-offs, demonstrating that PhoBERT provides optimal balance for practical deployment with 2GB memory requirements and 20ms inference times compared to LLMs requiring 8GB memory and 3+ second inference times.

Objective 4 - Learning Paradigm Assessment: The systematic comparison revealed supervised fine-tuning superiority over prompt-based learning for this specialized task, with few-shot learning paradoxically degrading performance compared to zero-shot approaches in several cases.

7.2 Key Research Contributions

7.2.1 Empirical Contributions

Performance Benchmarks: This study establishes the first comprehensive performance benchmarks for Vietnamese fake news detection across diverse model families, providing baseline metrics for future research comparison and advancement.

Efficiency Analysis: The systematic evaluation of computational requirements alongside accuracy metrics provides practical deployment guidance previously unavailable in Vietnamese NLP literature.

Class Imbalance Insights: Our findings demonstrate that severe class imbalance (83.2% vs 16.8%) requires sophisticated pre-trained models, as traditional approaches fail completely regardless of feature engineering improvements.

7.2.2 Methodological Contributions

Evaluation Framework: The multi-dimensional assessment methodology balancing accuracy, efficiency, and practical considerations provides a replicable template for similar comparative studies in other low-resource languages.

Prompt Engineering Analysis: The systematic comparison of zero-shot versus few-shot learning reveals important limitations in current prompt-based approaches for Vietnamese text classification.

Language-Specific Model Validation: The dramatic performance advantages of Vietnamese-specific models provide strong empirical evidence supporting continued investment in language-specific development over universal multilingual scaling.

7.3 Practical Recommendations

Based on our comprehensive evaluation, we provide clear guidance for practitioners and researchers:

7.3.1 For Immediate Deployment

Model Selection: Fine-tuned PhoBERT represents the optimal choice for Vietnamese fake news detection, offering superior performance with reasonable computational requirements.

Infrastructure Requirements: Minimum 2GB GPU memory with modern CUDA support enables effective PhoBERT deployment for real-time misinformation detection applications.

Performance Expectations: Properly implemented systems should achieve >95% accuracy with balanced performance across real and fake news categories.

7.3.2 For Research Development

Architecture Priorities: Focus on Vietnamese-adapted transformer architectures rather than scaling multilingual models for optimal performance in specialized Vietnamese NLP tasks.

Training Strategies: Prioritize end-to-end fine-tuning over feature extraction approaches when computational resources permit gradient-based optimization.

Evaluation Protocols: Employ macro-averaged metrics and class-wise analysis for realistic performance assessment under severe class imbalance conditions.

7.4 Research Limitations

While providing valuable insights, this research operates within several constraints that define the scope and applicability of findings:

Dataset Scope: Evaluation on a single dataset limits generalization claims across different temporal periods, news domains, and misinformation strategies.

Computational Constraints: Hardware limitations prevented exhaustive hyperparameter exploration and evaluation of larger model variants that might reveal different performance patterns.

Language Coverage: Findings specific to Vietnamese may not generalize to other low-resource languages with different linguistic characteristics and cultural contexts.

Temporal Stability: Static evaluation captures performance at a single time point rather than assessing long-term robustness to evolving misinformation strategies.

7.5 Future Research Directions

This research establishes a foundation for several important future investigations:

Cross-Temporal Validation: Evaluating model robustness across different time periods to assess adaptation capabilities as misinformation strategies evolve.

Multimodal Extension: Incorporating visual and audio misinformation detection to address multimedia content increasingly prevalent in Vietnamese media ecosystems.

Real-Time Adaptation: Developing continuous learning systems capable of adapting to emerging misinformation patterns without catastrophic forgetting of previous knowledge.

Cross-Linguistic Generalization: Extending evaluation frameworks to other Southeast Asian languages to determine whether findings generalize beyond Vietnamese-specific characteristics.

7.6 Closing Remarks

This comprehensive evaluation demonstrates that effective Vietnamese fake news detection is both technically achievable and practically deployable with current technology. The clear superiority of language-specific approaches validates continued investment in Vietnamese NLP

development while providing immediate actionable guidance for misinformation detection system deployment.

The research contributes to the broader scientific understanding of multilingual NLP effectiveness while addressing urgent practical needs for reliable Vietnamese misinformation detection capabilities. As digital misinformation continues evolving in sophistication and prevalence, this work provides an evidence-based foundation for developing more effective technological countermeasures.

The journey toward robust misinformation detection requires ongoing research, technological innovation, and careful attention to language-specific requirements. This study represents a significant step in that direction, offering both immediate practical solutions and a research foundation for continued advancement in Vietnamese natural language processing and misinformation detection technologies.

Our findings demonstrate that with appropriate model selection, training strategies, and deployment considerations, automated Vietnamese fake news detection can achieve performance levels suitable for real-world application while maintaining computational efficiency necessary for practical deployment. This achievement opens new possibilities for combating misinformation in Vietnamese media ecosystems while contributing valuable insights to the global effort of preserving information integrity in our digital age.

Appendix A

Research Documentation and Source Materials

A.1 Plagiarism Check Report

The originality of this research has been independently validated using the Turnitin plagiarism detection system.

Similarity Index: X.X% (excluding references and standard terminology) **Verification Date:** [Date of check] **Status:** Within acceptable academic standards for original research

A.2 Research Integrity Declaration

This work represents original research conducted specifically for the BGRA 2025 competition. All experimental results, analysis, and conclusions are based on independent implementation and evaluation. No previously published work has been inappropriately reused without proper citation.

A.3 Source Code and Implementation

Repository: https://github.com/hoaianthai345/Detecting_Fake_News_in_Vietnamese_Media.git

A.3.1 Implementation Files

- BiLSTM models: `BiLSTM/*.ipynb`
- PhoBERT implementation: `PhoBERT/PHOBERT.ipynb`
- LLM evaluations: `LLMs/*.ipynb`

A.3.2 Key Dependencies

- PyTorch 2.1.0, Transformers 4.36.2
- PhoBERT: https://huggingface.co/vinai/phobert-base
- Vietnamese embeddings: Word2Vec, FastText
- LLMs: Qwen2.5, Llama-2, DeepSeek (Hugging Face)

A.4 Dataset and Ethics

Dataset: ReINTEL (VLSP 2020) - https://vlsp.org.vn/vlsp2020/eval/reintel **Usage:** Academic research under public license **Privacy:** All data pre-anonymized by dataset creators **Compliance:** Academic research purposes only

A.5 Reproducibility Information

Environment: Google Colaboratory, Tesla T4 GPU **Seeds:** Fixed random seeds for reproducible results **Configuration:** All hyperparameters documented in code

References

Journal Articles

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211–236.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901. https://arxiv.org/abs/2005.14165

Cai, R., Qin, B., Chen, Y., & Zhang, L. (2020). Sentiment analysis about investors and consumers in energy market based on BERT-BiLSTM. *IEEE Access*, *8*, 171408–171415.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*.

Fields, J., Chovanec, K., & Madiraju, P. (2023). A survey of text classification with transformers: How wide? how large? how long? how accurate? how expensive? how safe? *IEEE Access*. https://doi.org/10.1109/ACCESS.2024.3349952

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., & Ahmed, N. K. (2024). Bias and fairness in large language models: A survey. *Computational Linguistics*, *50*(3), 1097–1132. https://doi.org/10.1162/coli_a_00524

Granik, M., & Mesyura, V. (2017). Fake news detection using naive bayes classifier. *2017 IEEE first Ukraine conference on electrical and computer engineering (UKRCON)*, 900–903. https://doi.org/10.1109/UKRCON.2017.8100379

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Huang, J. (2020). Detecting fake news with machine learning. *Journal of Physics: Conference Series*, *1693*. https://doi.org/10.1088/1742-6596/1693/1/012158

Kaliyar, R. K., Goswami, A., & Narang, P. (2021). Fndnet–a deep convolutional neural network for fake news detection. *Cognitive Systems Research*, *61*, 32–44.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). Albert: A lite bert for self-supervised learning of language representations. *ICLR*.

Le, D.-T., Nguyen, H., & Vu, X.-S. (2020). Reintel: Reliable intelligence identification on vietnamese social network sites. *Proceedings of the 7th International Workshop on Vietnamese Language and Speech Processing (VLSP 2020)*. https://vlsp.org.vn/vlsp2020/eval/reintel

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Li, X. Y., Han, L. B., & Jiang, Z. F. (2024). Deep learning-based algorithm for classification of news text. *IEEE Access*, *12*, 159086–159100.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, *26*, 3111–3119.

Mishra, R. K., Reddy, G. Y. S., & Pathak, H. (2021). The understanding of deep learning: A comprehensive review. *Mathematical Problems in Engineering*, *2021*, 1–15. https://doi.org/10.1155/2021/5548884

Nasir, J. A., Khan, O. S., & Varlamis, I. (2021). Fake news detection: A hybrid cnn-rnn based deep learning approach. *International Journal of Information Management Data Insights*, *1*(1), 100007.

Nguyen, D. Q., & Nguyen, A. T. (2020a). Phobert: Pre-trained language models for vietnamese. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1037–1042.

Nguyen, D. Q., & Nguyen, A. T. (2020b). PhoBERT: Pre-trained language models for vietnamese. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 1037–1042.

Nguyễn, K. T.-T., & Nguyễn, K. V. (2020). Reintel challenge 2020: Exploiting transfer learning models for reliable intelligence identification on vietnamese social network sites. *Proceedings of the Vietnamese Language and Speech Processing 2020 (VLSP 2020)*. https://vlsp.org.vn

Pham, N. D., Le, T. H., Do, T. D., Vuong, T. T., Vuong, T. H., & Ha, Q. T. (2021). Vietnamese fake news detection based on hybrid transfer learning model and tf-idf. *2021 13th International Conference on Knowledge and Systems Engineering (KSE)*, 1–6.

Raiaan, M. A. K., Mukta, M. S. H., Fatema, K., Fahad, N. M., Sakib, S., Mim, M. M. J., Ahmad, J., Ali, M. E., & Azam, S. (2024). A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE Access*, *12*, 26839–26855. https://doi.org/10.1109/ACCESS.2024.3365742

Rani, M., & Virmani, C. (2022). Detection of fake news on social media: A review. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4143832

Riedel, B., Augenstein, I., Spithourakis, G. P., & Riedel, S. (2017). A simple but tough-to-beat baseline for the fake news challenge stance detection task. *arXiv preprint*. https://doi.org/10.48550/arXiv.1707.03264

Roumeliotis, K. I., Tselikas, N. D., & Nasiopoulos, D. K. (2025). Fake news detection and classification: A comparative study of convolutional neural networks, large language models, and natural language processing models. *Future Internet*, *17*(1), 28. https://doi.org/10.3390/fi17010028

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. *NeurIPS EMC2 Workshop*.

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017a). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, *19*(1), 22–36.

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017b). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, *19*(1), 22–36.

Tandoc Jr., E. C. (2019). The facts of fake news: A research review. *Sociology Compass*, *13*(9), e12724. https://doi.org/10.1111/soc4.12724

Tandoc Jr., E. C., Lim, Z. W., & Ling, R. (2017). Defining "fake news": A typology of scholarly definitions. *Digital Journalism*, *6*(2), 137–153.

Thota, A., Tilak, P., Ahluwalia, S., & Lohia, N. (2018). Fake news detection: A deep learning approach. *SMU Data Science Review*, *1*(3). https://scholar.smu.edu/datasciencereview/vol1/iss3/10

Touvron, H., Martin, L., Stone, K., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Umer, M., Mohamad, S., & Abuelma'atti, H. (2022). A comprehensive review on word embedding models: From word2vec to bert. *IEEE Access*, *10*, 43878–43901.

Võ, Đ. V., & Đỗ, P. (2023). Detecting vietnamese fake news [Special Issue on ISDS]. *CTU Journal of Innovation and Sustainable Development*, *15*, 39–46. https://doi.org/10.22144/ctujoisd.2023.033

Zhang, X., & Ghorbani, A. A. (2020). An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*, *57*(2). https://doi.org/10.1016/j.ipm.2019.03.004

Zhang, X., Li, Y., & Wang, J. (2024). Fake news detection and classification with large language models: Challenges and opportunities. *Future Internet*, *17*(1), 28. https://www.mdpi.com/1999-5903/17/1/28

Zhou, X., Wu, J., & Zafarani, R. (2020). Safe: Similarity-aware multi-modal fake news detection. *Pacific-Asia Conference on knowledge discovery and data mining*, 354–367.

Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, *53*(5), 1–40.

Online Resources

Dinh, T. M. (2013). *Thực trạng xử lý ngôn ngữ tự nhiên và dịch máy đối với tiếng việt*. https://elib.vku.udn.vn/handle/123456789/314

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). *Bag of tricks for efficient text classification*. arXiv: 1607.01759. https://arxiv.org/abs/1607.01759

Le, N. T. (2022). *Phát hiện tin tức giả cho tin tức tiếng việt bằng cách kết hợp các mô hình học sâu*. https://digital.lib.ueh.edu.vn/handle/UEH/66945

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *Roberta: A robustly optimized BERT pretraining approach*. arXiv: 1907.11692. https://arxiv.org/abs/1907.11692

Nguyen, P. (2023). *Những hệ lụy khi tin giả phát tán*. https://vnexpress.net/nhung-he-luy-khi-tin-gia-phat-tan-4665859.html

Nguyen, T. V. A. (2024). *Chuyển đổi số và báo chí đa nền tảng ở việt nam hiện nay*. https://lyluanchinhtri.vn/chuyen-doi-so-va-bao-chi-da-nen-tang-o-viet-nam-hien-nay-6535.html

Topics, E. (2025, January). *Countries with the largest digital populations in the world as of january 2025*. https://explodingtopics.com/blog/countries-internet-users

Technical Reports

Bai, J., Bai, S., Chu, Y., et al. (2023). Qwen technical report [Accessed: 2024].

Cambridge Dictionary. (n.d.). Fake news [Retrieved July 27, 2025].

DeepSeek-AI. (2024). Deepseek-r1: Reasoning at scale [Accessed: 2024].

Dettmers, T., Lewis, M., Belkada, Y., & Zettlemoyer, L. (2022). Bitsandbytes: 8-bit optimizers and quantization routines.

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2024). A comprehensive overview of large language models.

Nguyen, D. Q., Tuan, T., Nguyen, A. T., et al. (2019). Phow2v: Vietnamese word embedding. https://github.com/VinAIResearch/PhoW2V

Su, D., Yuan, Z., & Li, C. (2023). Fake news detectors are biased towards llm-generated content.

Detecting Fake News in Vietnamese Media BGRA2025

ORIGINALITY REPORT

12%	9%	9%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.mdpi.com Internet Source	1%
2	arxiv.org Internet Source	1%
3	aclanthology.org Internet Source	<1%
4	www.frontiersin.org Internet Source	<1%
5	arrow.tudublin.ie Internet Source	<1%
6	"Computational Intelligence in Engineering Science", Springer Science and Business Media LLC, 2026 Publication	<1%
7	Yiming Zhang, Koji Tsuda. "Ab-VS: Evaluating Large Language Models for Virtual Antibody Screening via Antibody-Antigen Interaction Prediction", Cold Spring Harbor Laboratory, 2025 Publication	<1%

8	Ying Guo, Hong Ge, Jinhong Li. "Fake News Detection Based on Two-Branch Network and Domain Adversarial", 2022 IEEE 5th International Conference on Computer and Communication Engineering Technology (CCET), 2022 Publication	<1 %
9	Mohammad Abu Tami, Huthaifa I. Ashqar, Mohammed Elhenawy, Sebastien Glaser, Andry Rakotonirainy. "Using Multimodal Large Language Models (MLLMs) for Automated Detection of Traffic Safety-Critical Events", Vehicles, 2024 Publication	<1 %
10	Arra Kumar, Suprakash Gupta. "Predicting Fatal Mine Accident Categories Using Text Mining and Machine Learning: A Comparative Model Analysis", Springer Science and Business Media LLC, 2025 Publication	<1 %
11	assets-eu.researchsquare.com Internet Source	<1 %
12	www.geeksforgeeks.org Internet Source	<1 %
13	open.uct.ac.za Internet Source	<1 %

14	www.coursehero.com Internet Source	<1%
15	peerj.com Internet Source	<1%
16	Danilo Dessì, Diego Reforgiato Recupero, Harald Sack. "An Assessment of Deep Learning Models and Word Embeddings for Toxicity Detection within Online Textual Comments", Electronics, 2021 Publication	<1%
17	Marcin Mateusz Czajka, Daria Kubacka, Aleksandra Świetlicka. "Embedding representation of words in sign language", Journal of Computational and Applied Mathematics, 2025 Publication	<1%
18	Pejman Peykani, Fatemeh Ramezanlou, Cristina Tanasescu, Sanly Ghanidel. "Large Language Models: A Structured Taxonomy and Review of Challenges, Limitations, Solutions, and Future Directions", Applied Sciences, 2025 Publication	<1%
19	dipot.ulb.ac.be Internet Source	<1%
20	dspace.lib.uom.gr Internet Source	

		<1%
21	huggingface.co Internet Source	<1%
22	Dimitrios K. Nasiopoulos, Konstantinos I. Roumeliotis, Damianos P. Sakas, Kanellos Toudas, Panagiotis Reklitis. "Financial Sentiment Analysis and Classification: A Comparative Study of Fine-Tuned Deep Learning Models", International Journal of Financial Studies, 2025 Publication	<1%
23	hdl.handle.net Internet Source	<1%
24	idus.us.es Internet Source	<1%
25	lup.lub.lu.se Internet Source	<1%
26	www.ijisae.org Internet Source	<1%
27	www.ir.juit.ac.in:8080 Internet Source	<1%
28	Fatema Tuj Johora Faria, Mukaffi Bin Moin, Zayeed Hasan, Md. Arafat Alam Khandaker, Niful Islam, Khan Md Hasib, M.F. Mridha.	<1%

"MultiBanFakeDetect: Integrating advanced fusion techniques for multimodal detection of Bangla fake news in under-resourced contexts", International Journal of Information Management Data Insights, 2025
Publication

29	Joyeta Ghosh, Jyoti Taneja, Ravi Kant. " Decoding Host-Pathogen Interactions in : Insights into Allelic Variation and Antimicrobial Resistance Prediction Using Artificial Intelligence and Machine Learning based approaches ", Cold Spring Harbor Laboratory, 2025 Publication	<1 %
30	latestsmartphone732.blogspot.com Internet Source	<1 %
31	libres.uncg.edu Internet Source	<1 %
32	www.sci-hub.se Internet Source	<1 %
33	Rajan Gupta, Sanju Tiwari, Poonam Chaudhary. "Generative AI: Techniques, Models and Applications", Springer Science and Business Media LLC, 2025 Publication	<1 %
34	ceur-ws.org Internet Source	<1 %

35	kuscholarworks.ku.edu Internet Source	<1%
36	research-information.bris.ac.uk Internet Source	<1%
37	"Natural Language Processing and Chinese Computing", Springer Science and Business Media LLC, 2018 Publication	<1%
38	"Web Information Systems and Applications", Springer Science and Business Media LLC, 2024 Publication	<1%
39	Mutaz A.B. Al-Tarawneh, Hassan Kanj, Wael Hosny Fouad Aly. "An integrated MCDM framework for trust-aware and fair task offloading in heterogeneous multi-provider Edge-Fog-Cloud systems", Results in Engineering, 2025 Publication	<1%
40	thesai.org Internet Source	<1%
41	"Proceedings of International Conference on Artificial Intelligence and Networks", Springer Science and Business Media LLC, 2025 Publication	<1%

42	Edson C. Tandoc. "The facts of fake news: A research review", Sociology Compass, 2019 Publication	<1 %
43	H L Gururaj, Francesco Flammini, V Ravi Kumar, N S Prema. "Recent Trends in Healthcare Innovation", CRC Press, 2025 Publication	<1 %
44	Hajiali, Mahdi. "OCR Post-Processing Using Large Language Models", University of Nevada, Las Vegas, 2024 Publication	<1 %
45	Razumovskaia, Evgeniia. "Advancing Language Equity and Sample Efficiency in Task-Oriented Dialogue Systems", University of Cambridge (United Kingdom) Publication	<1 %
46	Sanger, Mario. "Representation Learning for Biomedical Text Mining", Humboldt Universitaet zu Berlin (Germany) Publication	<1 %
47	Fatima, Nishath. "Deploying Transformer Models to Detect and Analyze Sponsored Content in Spotify Podcasts", University of California, Los Angeles, 2023 Publication	<1 %
48	flore.unifi.it Internet Source	<1 %

49	research.sabanciuniv.edu Internet Source	<1%
50	www.preprints.org Internet Source	<1%
51	Christian W. F. Mayer, Sabrina Ludwig, Steffen Brandt. "Prompt text classifications with transformer models! An exemplary introduction to prompt-based learning with large language models", Journal of Research on Technology in Education, 2022 Publication	<1%
52	Guan Wang, Rebecca Frederick, Jinglong Duan, William B. L. Wong, Verica Rupar, Weihua Li, Quan Bai. "Detecting misinformation through framing theory: the frame element-based model", Journal of Computational Social Science, 2025 Publication	<1%
53	Hifzhan Frima Thousani, Muhammad Taali. "BiLSTM-Based Sentiment Analysis Of Traveloka Hotel Reviews In Yogyakarta For Data-Driven Communication Strategies", INJECT (Interdisciplinary Journal of Communication), 2025 Publication	<1%
54	Ningyuan You, Chang Liu, Hai Lin, Sai Wu, Gang Chen, Ning Shen. "Benchmarking Pre-	<1%

trained Genomic Language Models for RNA Sequence-Related Predictive Applications", Cold Spring Harbor Laboratory, 2025
Publication

55	etd.astu.edu.et Internet Source	<1%
56	etheses.whiterose.ac.uk Internet Source	<1%
57	mafiadoc.com Internet Source	<1%
58	www.akinik.com Internet Source	<1%
59	"Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2018", Springer Science and Business Media LLC, 2019 Publication	<1%
60	Bhattacharjee, Amrita. "Building Trust in AI via Safe and Responsible Use of LLMs.", Arizona State University Publication	<1%
61	Ge, Fei. "Fine-Tune Whisper and Transformer Large Language Model for Meeting Summarization", University of California, Los Angeles, 2024 Publication	<1%

62	Komal Singh, Nikhil Kumar Singh, Manish Khare. "chapter 3 Fake News Detection Using Machine Learning Classifiers", IGI Global, 2025 Publication	<1%
63	Mekala, Dheeraj. "Training Data Curation for Language Models With Weak Supervision", University of California, San Diego Publication	<1%
64	Nighojkar, Animesh. "An Inference-Centric Approach to Natural Language Processing and Cognitive Modeling", University of South Florida, 2024 Publication	<1%
65	Shuyang Nie, Fan Li, Wei Wei, Kun Liu. "Chapter 15 Domain-Specific Information Extraction inChinese withPre-trained Language Models: An Exploration Report", Springer Science and Business Media LLC, 2025 Publication	<1%
66	Zhang, Felix. "Commonsense-Guided Text Generation with Knowledge Grounding and Scoring", University of California, Los Angeles, 2023 Publication	<1%

67	"Knowledge Science, Engineering and Management", Springer Science and Business Media LLC, 2021 Publication	<1%
68	"Pattern Recognition", Springer Science and Business Media LLC, 2025 Publication	<1%
69	Bandera, Calliope Chloe. "Empathy Cause Identification: Towards Unveiling Empathic Triggers in Online Interactions", University of Illinois at Chicago Publication	<1%
70	Chen, Angelica. "Improving Language Models Through the Lens of Training Dynamics", New York University, 2025 Publication	<1%
71	Riccardo Cantini, Cristian Cosentino, Irene Kilanioti, Fabrizio Marozzo, Domenico Talia. "Unmasking deception: a topic-oriented multimodal approach to uncover false information on social media", Machine Learning, 2025 Publication	<1%
72	Yehao Li, Yingwei Pan, Jingwen Chen, Ting Yao, Tao Mei. "X-modaler", Proceedings of the 29th ACM International Conference on Multimedia, 2021	<1%

	Publication	
73	api.repository.cam.ac.uk Internet Source	<1 %
74	assets.amazon.science Internet Source	<1 %
75	core.ac.uk Internet Source	<1 %
76	d-nb.info Internet Source	<1 %
77	ikee.lib.auth.gr Internet Source	<1 %
78	journals.uran.ua Internet Source	<1 %
79	papers.academic-conferences.org Internet Source	<1 %
80	pmc.ncbi.nlm.nih.gov Internet Source	<1 %
81	rsisinternational.org Internet Source	<1 %
82	scholarspace.manoa.hawaii.edu Internet Source	<1 %
83	sighum.files.wordpress.com Internet Source	<1 %
	stdj.scienceandtechnology.com.vn	

84	Internet Source	<1%
85	tel.archives-ouvertes.fr Internet Source	<1%
86	unbscholar.dspace.lib.unb.ca Internet Source	<1%
87	www.etda.or.th Internet Source	<1%
88	www.theseus.fi Internet Source	<1%
89	"Computational Science and Computational Intelligence", Springer Science and Business Media LLC, 2025 Publication	<1%
90	"Neural Information Processing", Springer Science and Business Media LLC, 2025 Publication	<1%
91	"Proceedings of 5th International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications", Springer Science and Business Media LLC, 2025 Publication	<1%
92	"Proceedings of the 9th International Conference on Computational Science and Technology", Springer Science and Business Media LLC, 2023	<1%

Publication

93	Callum Walker, Joseph Lambert. "The Routledge Handbook of the Translation Industry", Routledge, 2025 Publication	<1 %
94	Dario Guidotti, Laura Pandolfo, Luca Pulina. "Discovering sentiment insights: streamlining tourism review analysis with Large Language Models", Information Technology & Tourism, 2025 Publication	<1 %
95	Edson C. Tandoc. "Fake News Across Asian Countries", Routledge, 2025 Publication	<1 %
96	Kaiser, Tamanna. "Transformer-Based Sentence Classification in the Medical Domain: Evaluation of Pre-Trained and PubMed Fine-Tuned Models.", University of Windsor (Canada) Publication	<1 %
97	Nangia, Nikita. "Why, How, and When to Effectively Crowdsource Data for Natural Language Processing Research", New York University, 2024 Publication	<1 %
98	Naseem, Usman. "Hybrid Words Representation for the Classification of Low	<1 %

Quality Text", University of Technology Sydney (Australia), 2023
Publication

---

99 Sun, Zhewei. "Natural Language Processing for Slang.", University of Toronto (Canada), 2024 <1%
Publication

---

100 Te Han, Rong-Gang Cong, Biying Yu, Baojun Tang, Yi-Ming Wei. "Integrating local knowledge with ChatGPT-like large-scale language models for enhanced societal comprehension of carbon neutrality", Energy and AI, 2024 <1%
Publication

---

101 Tian, Yuanyuan. "Enhancing Geographic Information Retrieval by Generative AI and Large Language Models.", Arizona State University <1%
Publication

---

102 Yue, Xiangyu. "Learning Transferable Representations Across Domains", University of California, Berkeley, 2023 <1%
Publication

---

103 Zepu Yi, Chenxu Tang, Songfeng Lu. "User Comment-Guided Cross-Modal Attention for Interpretable Multimodal Fake News Detection", Applied Sciences, 2025 <1%
Publication

104	Zhigao Huang, Musheng Chen, Shiyan Zheng. "Transformer spectral optimization: From gradient frequency analysis to adaptive spectral integration", Applied Soft Computing, 2025 Publication	<1 %
105	content.sciendo.com Internet Source	<1 %
106	dokumen.pub Internet Source	<1 %
107	dspace.cvut.cz Internet Source	<1 %
108	dspace.ut.ee Internet Source	<1 %
109	ebin.pub Internet Source	<1 %
110	file.techscience.com Internet Source	<1 %
111	gtcs.cs.memphis.edu Internet Source	<1 %
112	ijcem.in Internet Source	<1 %
113	library.oapen.org Internet Source	<1 %

114	mdpi-res.com Internet Source	<1%
115	repository.up.ac.za Internet Source	<1%
116	researchers.cdu.edu.au Internet Source	<1%
117	scholarshare.temple.edu Internet Source	<1%
118	www.researchgate.net Internet Source	<1%
119	www.techscience.com Internet Source	<1%
120	www.utupub.fi Internet Source	<1%
121	yamanashi.repo.nii.ac.jp Internet Source	<1%
122	Balasubramanian Palani, Sivasankar Elango, 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, 2021 Publication	<1%
123	Johan Farkas, Jannick Schou. "Post-Truth, Fake News and Democracy - Mapping the Politics	<1%

of Falsehood", Routledge, 2023
Publication

---

124 Konstantinos I. Roumeliotis, Nikolaos D. Tselikas, Dimitrios K. Nasiopoulos. "Fake News Detection and Classification: A Comparative Study of Convolutional Neural Networks, Large Language Models, and Natural Language Processing Models", Future Internet, 2025 <1%
Publication

---

125 Sakai, Hajar. "Shifting the Paradigm: A Generative AI Framework for Large Language Models Integration in Healthcare Text Classification", State University of New York at Binghamton <1%
Publication

---

126 V. Sharmila, S. Kannadhasan, A. Rajiv Kannan, P. Sivakumar, V. Vennila. "Challenges in Information, Communication and Computing Technology", CRC Press, 2024 <1%
Publication

---

127 Volkan Altıntaş. "Beyond Classical AI: Detecting Fake News with Hybrid Quantum Neural Networks", Applied Sciences, 2025 <1%
Publication

---

128 Yang, Zhou. "Explainable Learning With Meaningful Perturbations", The George <1%

Washington University, 2023
Publication

129	birmingham.elsevierpure.com Internet Source	<1%
130	QunHui Zhou, Tijian Cai. "Adaptive gate residual connection and multi-scale RCNN for fake news detection", Machine Learning with Applications, 2025 Publication	<1%
131	Selaković, Marko. "Countering Business-Related Fake News Online: An Examination of the Application of Situational Crisis Communication Theory", SP Jain School of Global Management (India), 2024 Publication	<1%
132	Timothy R. McIntosh, Teo Susnjak, Tong Liu, Paul Watters, Dan Xu, Dongwei Liu, Malka N. Halgamuge. "From Google Gemini to OpenAI Q (Q-Star): A Survey on Reshaping the Generative Artificial Intelligence (AI) Research Landscape*", Technologies, 2025 Publication	<1%
133	Xichen Zhang, Ali A. Ghorbani. "An overview of online fake news: Characterization, detection, and discussion", Information Processing & Management, 2020 Publication	<1%

134 Yang Deng, Wenxuan Zhang, Weiwen Xu, Wenqiang Lei, Tat-Seng Chua, Wai Lam. "A Unified Multi-task Learning Framework for Multi-goal Conversational Recommender Systems", ACM Transactions on Information Systems, 2023

Publication

<1%

Exclude quotes	On	Exclude matches	< 5 words
Exclude bibliography	On		