HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

MACHINE LEARNING AND DATA MINING - IT3191E

CAPSTONE PROJECT REPORT: FAKE NEWS DETECTION

Instructor: **Ph.D. Nguyen Duc Anh**

Group: 1
Students: Ho Bao Thu - 20226003
Tran Kim Cuong - 20226017
Nguyen Dinh Duong - 20225966
Nguyen My Duyen - 20225967
Ha Viet Khanh - 20225979
Dang Van Nhan - 20225990

Hanoi, June 2025

Contents

Abstract

Automatic fake news detection is a challenging problem in deception detection with significant political and social implications. This study investigates the effectiveness of large-scale pretrained language models – RoBERTa, XLNet, and DeBERTa – for fake news classification across diverse domains. We evaluate three training strategies: (1) Domain-specific fine-tuning; (2) Pooled-domain fine-tuning; (3) Domain-matched ensemble. Experiments are conducted on three prominent datasets – FakeNewsNet, LIAR, and COVID-19 – representing varied content genres and misinformation styles. Results show that domain-specific models perform best within their respective domains, while pooled-domain sacrifices domain precision for generality. The ensemble approach improves accuracy and precision across domains, demonstrating the potential of combining specialized models. This work highlights the strengths and trade-offs of each strategy and provides insight into building reliable fake news detection systems across heterogeneous media environments.

1 Introduction

The pervasive influence of online social media has transformed news dissemination, enabling rapid access to information. However, fake news is increasingly spread across various social media platforms, making it harder to distinguish between reliable information and deceptive content, and posing serious threats to public trust. The fake news contents can be expressed in different formats such as rumors, false statements, satires, and fake advertisements. Today, many researchers aimed to develop an automatic system to prevent fake news on social media. However, the fake news detection task is still a challenging research problem in natural language processing (NLP).

Traditional machine learning methods such as Naive Bayes, Support Vector Machines (SVMs), and Decision Trees have historically been used to address the problem of fake news detection. These methods focus on hand-crafted features, capturing patterns in text. Recent advancements in neural network architectures, particularly Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and transformer-based models, have significantly improved fake news detection by capturing complex semantic and contextual patterns.

Recurrent models like LSTM and GRU process sequences unidirectionally, limiting parallelization and efficiency in handling long-range dependencies. To overcome this, Vaswani et al. [30] introduced the transformer model, which enables parallel sequence learning with an encoder-decoder structure, multi-head attention, feed-forward networks, and masked multi-head attention, allowing for more efficient processing.

Building on the transformer architecture, Devlin et al. [3] introduced BERT, which uses the encoder structure to capture bidirectional context, enabling the model to consider both left and right context simultaneously. This improves the model's ability to handle long-range dependencies.

In this project, we evaluate the performance of RoBERTa, XLNet, and DeBERTa – advanced variants of BERT – on multiple datasets from diverse domains. These datasets include FakeNewsNet, LIAR, and the COVID-19 Fake News Dataset, ensuring that our models are tested on a wide range of misinformation styles and sources. To improve the robustness and generalization of fake news detection, we also adopt two strategies: Pooled-Domain Fine-Tuning, where a

single model is fine-tuned on a combined dataset from all domains, and Domain-Matched Ensemble (DME), which ensembles multiple models. By comparing both individual and ensemble models across these strategies, we aim to enhance the overall accuracy and adaptability of fake news detection systems.

2 Related Work

This section provides a comprehensive overview of existing studies and techniques applied in the field of fake news detection. By reviewing and synthesizing previous research, we outline technical advances, identify research gaps, and propose directions for our work.

2.1 Traditional Methods

Early fake news detection methods primarily relied on traditional machine learning (ML) algorithms such as Support Vector Machines (SVM), Decision Trees, and Logistic Regression, often utilizing basic textual features like Term Frequency-Inverse Document Frequency (TF-IDF) and n-grams. While structurally simple, these approaches proved effective on well-curated datasets; for instance, Patwa et al. [20] achieved an F1-score of 93.32% using SVM on COVID-19-related news. However, these methods rely on shallow features, failing to capture deep semantics, generalize to new fake-news tactics, or scale to large data volumes. [13]

With the advancement of deep learning, single-modal models like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), including Long Short-Term Memory (LSTM) and Bidirectional LSTMs (Bi-LSTM), have significantly enhanced the ability to capture semantic and sequential dependencies in text [34, 22]. These models often incorporate pre-trained word embeddings such as Word2Vec [14] and GloVe [21] to improve textual representation [13]. Nevertheless, these methods remain constrained to textual modalities, frequently struggle with generalization to out-of-domain content, and typically lack the capacity to exploit rich cross-modal signals present in modern misinformation formats. [13]

2.2 Advanced Deep Learning: Transformer-based Approaches

Recent studies have increasingly demonstrated that Transformer-based architectures–particularly models such as BERT (Bidirectional Encoder Representations from Transformers) [4] and its robust variant RoBERTa [12]–have driven a significant paradigm shift in Natural Language Processing (NLP). Characterized by their sophisticated self-attention mechanisms [31], these models represent a fundamental departure from previous sequential models like Recurrent Neural Networks (RNNs) [23] and Convolutional Neural Networks (CNNs) [11] by processing entire input sequences in parallel. This architectural innovation allows them to overcome the limitations of capturing long-range dependencies, which was a notorious challenge for earlier models. By effectively capturing rich contextual information and intricate relationships between distant words within a text, Transformer-based models deliver state-of-the-art performance across a wide range of NLP tasks. This includes not only core tasks like machine translation and text summarization but critically, also extends to complex applications such as textual fake news detection, where their ability to understand nuanced linguistic structures and identify subtle contextual inconsistencies is paramount [13, 8].

Despite recent progress, there is still a lack of direct comparisons between BERT-based models like RoBERTa, XLNet, and DeBERTa under consistent settings. Further research is needed on the generalizability of multimodal attention models and the development of lightweight, efficient architectures for real-time use.

3 Dataset Survey & Selection

3.1 Overview

To construct a reliable fake news detection system, we conducted a comprehensive survey of related datasets, guided by the framework proposed by Kuntur et al. [10], D'Ulizia et al. [5] and Murayama [16]. These references provided a comprehensive overview of dataset characteristics including size, domain, annotation quality, label distribution, and accessibility. Based on these surveys, we limit ourselves to publicly available models and ensure data balance.

Concerning dataset accessibility, several datasets were excluded from consideration because they are either no longer publicly available or require complex scraping procedures, which would hinder reproducibility – a key principle in modern machine learning research. These include Verification Corpus [2], MisInfoText [28], BuzzFace [24], FacebookHoax [27] and r/Fakeddit [17]. Label distribution imbalance can significantly hinder the model's learning capability and lead to misleading evaluation results. Therefore, we excluded certain datasets such as the CREDBANK dataset [15] suffers from extreme skewness: 99.35% of events are labeled as "credible" based on majority vote annotations. The disproportionate distribution undermines the dataset's suitability for binary fake news detection tasks, where model robustness depends on relatively balanced class representation.

We also examined whether the dataset's focus aligns with our specific objective. Since our task centers on **text-based** veracity classification of individual news articles, we excluded datasets with divergent goals or annotation levels. For example, PHEME [35] analyzes rumor propagation in conversation threads, Yelp [1] assesses review credibility in user feedback, and NELA-GT-2018 [18] assigns credibility at the source level. These differences risk introducing domain mismatch and annotation noise that could compromise model generalization.

After a thorough survey, we retained three datasets that best satisfy our requirements in terms of annotation quality, accessibility, label balance, and domain alignment. These are LIAR [32], FakeNewsNet [26] and The COVID-19 Fake News Dataset [19]. LIAR [32] offers fine-grained labels with a 6-point scale for political claims, balanced, and manually verified. FakeNewsNet [26] includes real/fake articles across political and entertainment domains. The COVID-19 Fake News Dataset [19] was specially curated for COVID-19 misinformation, manually verified, and balanced.

These datasets provide reliable, diverse, and task-aligned data sources for both training and evaluation. A more detailed exploration of these datasets is presented in Section 3.2, 3.3, 3.4.

3.2 The COVID-19 Fake News Dataset

The COVID-19 Fake News Dataset [19] (COVID-19 Dataset) contains 10,700 labeled samples from both news media and social media, annotated as either fake or real. Collected during the

pandemic, it captures a wide range of public discourse and was designed to support misinformation detection in a high-stakes health crisis. Its diverse sources and textual structures make it a valuable benchmark for training and evaluating models under real-world, high-variance conditions.

This dataset consists of three subsets: Train, Validation, and Test Set. Each sample includes an `id` as the identifier for the `tweet` field, and a binary `label` indicating whether the information is real or fake. The dataset is designed to reflect the format of social media content, particularly tweets related to the COVID-19 pandemic.

Table 1: COVID-19 Dataset Partition and Size

Subset	Samples	Real	Fake
Train	6,420	3,360	3,060
Validation	2,140	1,120	1,020
Test	2,140	1,120	1,020
Total	**10,700**	**5,600**	**5,100**

The dataset is clean and well-structured, with all three subsets sharing the same format: `id` (*int64*), `tweet` (*object*), and `label` (*object*), with no missing values. It also exhibits a nearly balanced label distribution – 52.3% real (5,600 samples) and 47.7% fake (5,100 samples) – which is beneficial for training binary classifiers and is detailed in Table 1.

3.3 FakeNewsNet

The FakeNewsNet dataset is a large-scale benchmark for fake news detection, comprising two distinct domains: GossipCop, which focuses on celebrity and entertainment news, and PolitiFact, which contains political statements and claims. Each sample contains four fields: `id`, `news_url`, `title`, and `tweet_ids`, allowing for both textual and social propagation analysis. Across both domains, every entry has a non-null `title` field. The dataset remains structurally consistent and suitable for text-based analysis.

Table 2 presents the number of samples in each domain and split, along with the corresponding label distributions.

Table 2: FakeNewsNet Dataset Partition and Size

Subset	Samples	Real	Fake	Domain
Train	15,498	11,772	3,726	GossipCop
Validation	3,321	2,523	798	GossipCop
Test	3,321	2,522	799	GossipCop
Train	739	437	302	PolitiFact
Validation	158	93	65	PolitiFact
Test	159	94	65	PolitiFact
Total	**23,196**	**16,441**	**6,755**	Both

3.4 LIAR

The LIAR dataset [32] comprises 13,791 fact-checked political statements from PolitiFact, each annotated with one of six fine-grained truthfulness labels. Alongside the core `statement` text, each sample includes metadata fields such as `speaker`, `subject`, `party`, and `context`. For our purposes, we focus exclusively on the `statement` field and convert the original labels into binary classes for fake news detection.

Specifically, `true` and `mostly-true` are grouped into the `real` class, while `half-true`, `barely-true`, `false`, and `pants-fire` are mapped to `fake`. This reduction aligns the dataset with others used in our study and reflects the semantic distinction between broadly factual claims and those exhibiting varying degrees of misinformation, from partial distortion to outright fabrication. The binarization facilitates a consistent evaluation framework while preserving the core distinction between credible and misleading content.

The label distribution after binarization is summarized in Table 3. Across all splits, approximately 65% of the statements are labeled as `fake`, reflecting a moderate class imbalance. The training set contains 6,602 fake and 3,638 real samples, while the validation and test sets show similar proportions.

Table 3: LIAR Dataset Partition and Size

Subset	Samples	Real	Fake
Train	10,240	3,638	6,602
Validation	1,284	420	864
Test	1,267	449	818

3.5 Summary

Based on the selection criteria discussed above, we retain three datasets for downstream experimentation: COVID-19 Dataset, FakeNewsNet, and LIAR. Table 4 summarizes the key characteristics of these datasets that justify their inclusion in our experimental pipeline.

Table 4: Summary of Selected Datasets

Dataset	Domain	Total Samples	Label Type
COVID-19 [19]	Health	10,700	Binary
FakeNewsNet [26]	Politics & Entertainment	23,196	Binary
LIAR [32]	Politics (Statements)	13,791	6-class $\rightarrow$ Binary

These datasets will serve as the foundation for the experiments described in Section 5, where we analyze their lexical characteristics, train transformer-based classifiers, and evaluate performance across domains.

4 Methodology

In this section, we analyze three transformer-based models, "XLNet", "RoBERTa, and "DeBERTa", focusing on their architectural and training improvements over BERT. Both models aim to overcome BERT's limitations in context modeling, masking strategy, and pretrain–finetune mismatch.

4.1 XLNet

XLNet builds upon the Transformer architecture but introduces several architectural innovations that address BERT's limitations in context modeling and pretraining.

- **Segment-level Recurrence:** XLNet inherits the segment-level recurrence mechanism from Transformer-XL, where hidden states from previous segments are cached and reused as key/value pairs in the self-attention computation of subsequent segments. This allows each token in the current segment to attend not only to earlier tokens in the same segment, but also to contextual representations from preceding segments, effectively extending the model's receptive field beyond fixed-length boundaries.
- **Relative Positional Encoding:** Instead of absolute position embeddings like BERT, XLNet employs a relative encoding scheme. This allows seamless integration of memory from previous segments without positional conflicts.
- **Two-Stream Self-Attention:** To support permutation-based autoregressive learning, XLNet introduces a dual representation per token: a content stream for contextual representation and a query stream for prediction without self-leakage. During training, the content stream attends to both previous and current positions to encode full context, while the query stream attends only to positions before the target token (based on the sampled permutation), ensuring that the model does not access the token it is predicting. This decoupling allows XLNet to maintain autoregressive consistency while learning bidirectional context in expectation. The attention mechanism is illustrated in Figure 1.

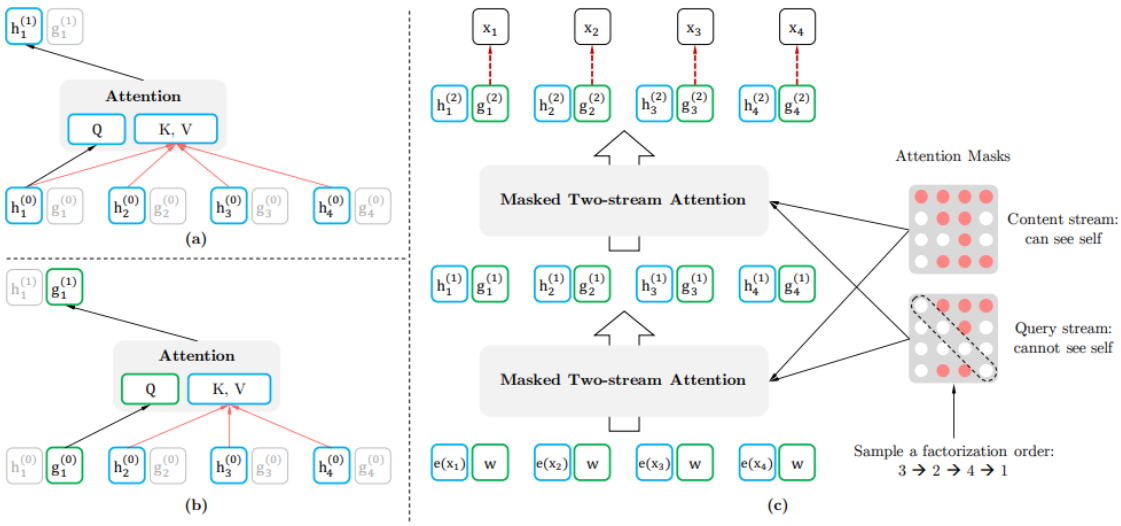

Figure 1: XLNet [33] architecture with two-stream attention and permutation-based factorization.

- **Permutation-Based Factorization:** XLNet maximizes the expected log-likelihood over all possible permutations of the input sequence, allowing it to capture bidirectional context while preserving the autoregressive property. To reduce the optimization complexity introduced by the permutation space, XLNet applies *partial prediction*, where only a subset of target positions–typically the last tokens in the permutation–are predicted. This improves training efficiency by reducing redundant computation while maintaining sufficient contextual coverage.
- **Larger Training Corpus:** Compared to BERT, XLNet is pretrained on a significantly larger and more diverse corpus, including BookCorpus, Wikipedia, Giga5, ClueWeb09, and Common Crawl, totaling approximately 126GB of text. This broader data coverage allows XLNet to learn more robust language representations and contributes to its improved performance on downstream tasks.

Together, these modifications enable XLNet to model bidirectional dependencies, handle long-range context via segment-level recurrence, and avoid the pretrain–finetune discrepancy inherent in BERT. However, these improvements come at the cost of increased computational complexity due to two-stream attention and dynamic permutation-based training, which lead to slower training and inference speeds compared to BERT and RoBERTa.

4.2 RoBERTa

RoBERTa (Robustly Optimized BERT Pretraining Approach), introduced by Liu et al. [12], is a robustly optimized version of the original BERT model [4]. It has demonstrated superior performance across various natural language processing (NLP) tasks through more extensive pretraining and fine-tuning of critical hyperparameters.

RoBERTa inherits the multi-layer Transformer encoder architecture from BERT. Each Transformer layer comprises a multi-head self-attention mechanism to learn contextual relationships and a feed-forward network to transform these representations. The key differences and optimizations of RoBERTa over the original BERT model lie in its pretraining process and hyperparameter refinements, which significantly enhance its performance and stability. Notable improvements, applied concurrently during RoBERTa's pretraining, include:

- **Dynamic Masking:** While the original BERT uses static masking, RoBERTa employs a dynamic masking strategy [12]. This means that masked tokens change randomly each time an input sequence is fed into the model. This technique exposes the model to a wider variety of contexts for the same original sentence, thereby reducing overfitting and encouraging the learning of more robust, generalized language representations, while also saving storage costs.
- **Elimination of Next Sentence Prediction (NSP) Task:** Empirical studies by Liu et al. (2019) [12] indicated that removing the NSP task during BERT's pretraining yielded better results on certain downstream tasks. Beside, RoBERTa focuses solely on the Masked Language Model (MLM) task using full-length sentences (FULL-SENTENCES) to maximize internal contextual learning.
- **Larger Batch Sizes and Higher Learning Rates:** RoBERTa is pretrained with significantly larger batch sizes (up to 8K) combined with higher learning rates. The use of large batch sizes has been shown to improve performance and optimize stability. To mitigate

initial instability when using high learning rates, RoBERTa applies a "warm-up" [31] strategy, gradually increasing the learning rate from a small value to its maximum during the initial training steps.

- **Larger Training Data and Extended Training Time:** RoBERTa is pretrained on a significantly larger and more diverse dataset than BERT, totaling approximately 160GB of uncompressed text. This corpus additionally encompasses CC-News, OpenWebText, and Stories. This combination allows the model to learn from a wide range of domains and writing styles. Furthermore, RoBERTa is trained with longer schedules, larger mini-batches, and higher learning rates. These enhancements collectively enable it to better capture complex linguistic structures and significantly improve downstream task performance.
- **Byte-Pair Encoding (BPE) for Text Encoding**: RoBERTa uses a byte-level BPE algorithm [25] instead of BERT's WordPiece. This enables encoding any text sequence without generating an '[UNK]' token, effectively handling rare, novel, or foreign words, thus enhancing scalability and stability despite a minor trade-off in encoding efficiency.

In our Fake News Detection study, the RoBERTa-base model was selected due to its proven capability in modeling complex contextual relationships across long textual sequences. This is particularly important given the nature of fake news, which often contains nuanced claims, context-dependent semantics, and non-standard vocabulary. Moreover, RoBERTa's robust handling of out-of-vocabulary (OOV) tokens enhances its adaptability in dynamic and linguistically diverse domains.

4.3 DeBERTa

DeBERTa (Decoding-enhanced BERT with Disentangled Attention), introduced by He et al. [7], enhances the BERT and RoBERTa models by addressing limitations in their attention mechanisms and positional encoding strategies. The key innovations in DeBERTa include:

- **Disentangled Attention Mechanism:** Unlike BERT and RoBERTa, where each token is represented by a single vector combining content and positional information, DeBERTa represents each token using two separate vectors: one for content and one for position. The attention weights between tokens are computed using disentangled matrices that separately consider content-to-content, content-to-position, position-to-content, and position-to-position interactions. This separation allows the model to capture more nuanced dependencies between tokens.
- **Enhanced Mask Decoder (EMD):** In the pretraining phase, DeBERTa incorporates absolute positional information at the decoding stage rather than at the input. This approach allows the model to benefit from relative positional encoding throughout the Transformer layers while still leveraging absolute positions during the prediction of masked tokens, leading to improved performance in masked language modeling tasks.
- **Virtual Adversarial Training (VAT):** To improve generalization during fine-tuning, DeBERTa employs a virtual adversarial training technique. This method introduces small perturbations to the input embeddings, encouraging the model to produce consistent predictions and enhancing its robustness to input variations.

- **Efficient Pretraining:** DeBERTa achieves superior performance compared to RoBERTa-Large while being pretrained on half the amount of data. For instance, it improves accuracy on the MNLI task by +0.9% (90.2% vs. 91.1%), on SQuAD v2.0 by +2.3% (88.4% vs. 90.7%), and on RACE by +3.6% (83.2% vs. 86.8%).
- **Scalability:** The DeBERTa architecture scales effectively to larger model sizes. A version with 1.5 billion parameters surpasses human performance on the SuperGLUE benchmark, achieving a macro-average score of 89.9 compared to the human baseline of 89.8.

These architectural enhancements enable DeBERTa to capture richer linguistic representations and achieve state-of-the-art results across various natural language understanding tasks.

5 Experiments

5.1 Data Preprocessing & Exploratory Data Analysis (EDA)

Prior to model training, we conducted comprehensive preprocessing and exploratory analysis to understand the structural characteristics of our selected datasets (Section 3). Our analysis focused on examining length distributions within training sets, comparing patterns across different labels, and understanding overall text characteristics. These insights proved crucial for making informed decisions about tokenization strategies and appropriate sequence lengths.

Our preprocessing approach was deliberately minimal to preserve the natural characteristics of the text while ensuring consistency. We standardized inputs by filtering non-alphanumeric characters and normalizing whitespace. This approach eliminated potential noise in our length-based statistics and tokenization processes while maintaining the integrity of the original content. Importantly, we retained stopwords to preserve natural sentence structure, which we considered essential for subsequent modeling phases.

5.1.1 The COVID-19 Fake News Dataset

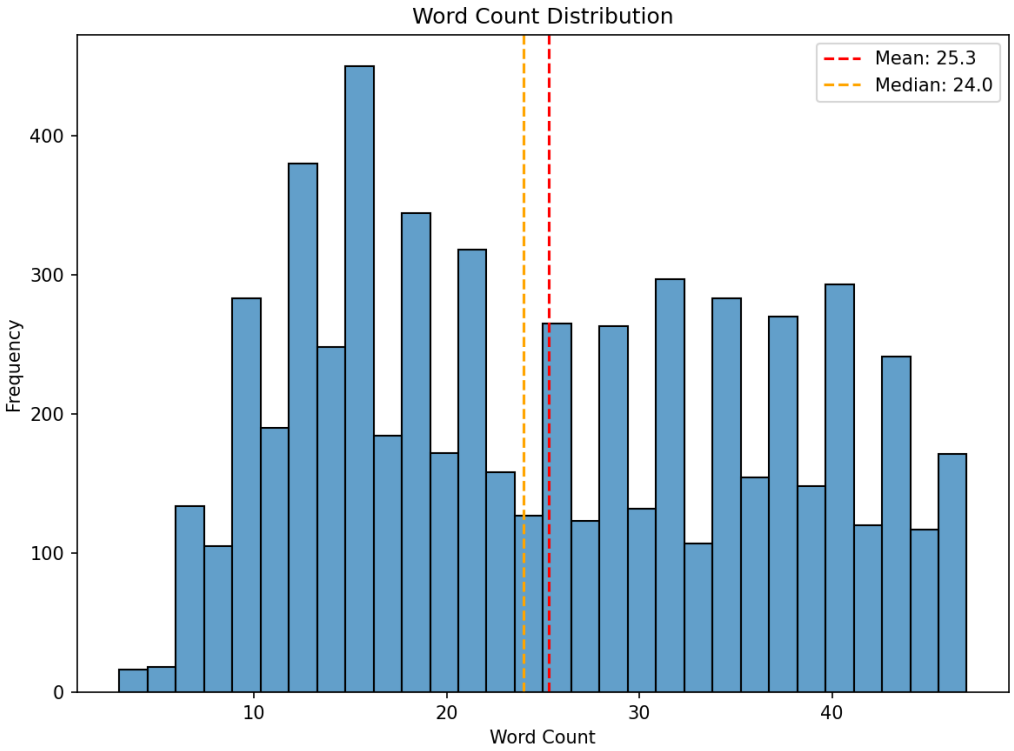

Figure 2: Word count distribution for COVID-19 Fake News dataset. Dashed lines show mean and median.

Our analysis of the COVID-19 dataset revealed insightful distributional characteristics. The text lengths clustered around an average of 25 words with a median of 24, indicating moderately short inputs overall. While we observed some degree of variability–particularly among fake news samples–the overall token count remained within a manageable range for most instances. Specifically, 95% of samples fell below the 128-token threshold when accounting for subword tokenization. This observation suggests that a maximum sequence length of 128 tokens is sufficient to capture the vast majority of content in this dataset without truncation.

5.1.2 GossipCop (FakeNewsNet)

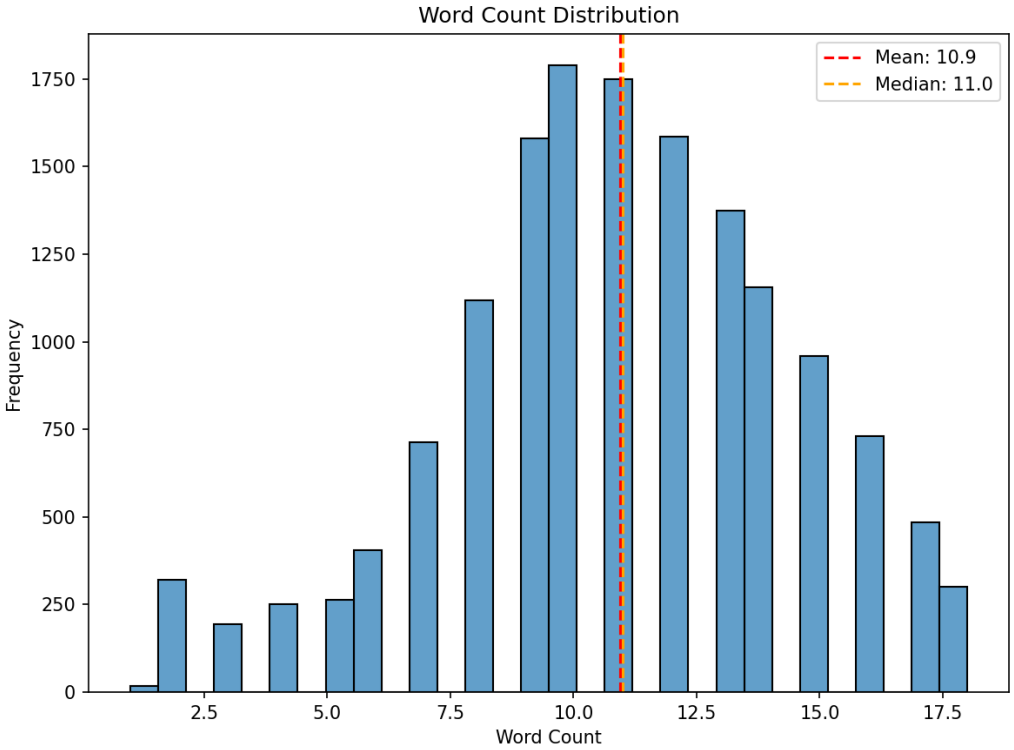

Figure 3: Word count distribution for GossipCop.

The GossipCop dataset presents a markedly different pattern from COVID-19 data. Here, we find a remarkably compact and symmetric distribution centered tightly around 11 words for both mean and median. What's particularly noteworthy is the consistency between fake and real news samples – both categories exhibit nearly identical length characteristics. This uniformity suggests that in the entertainment news domain, the length of content doesn't serve as a distinguishing feature between genuine and fabricated stories. Given this tight distribution, we determined that 128 tokens would provide adequate coverage for virtually all samples in this dataset.

5.1.3 PolitiFact (FakeNewsNet)

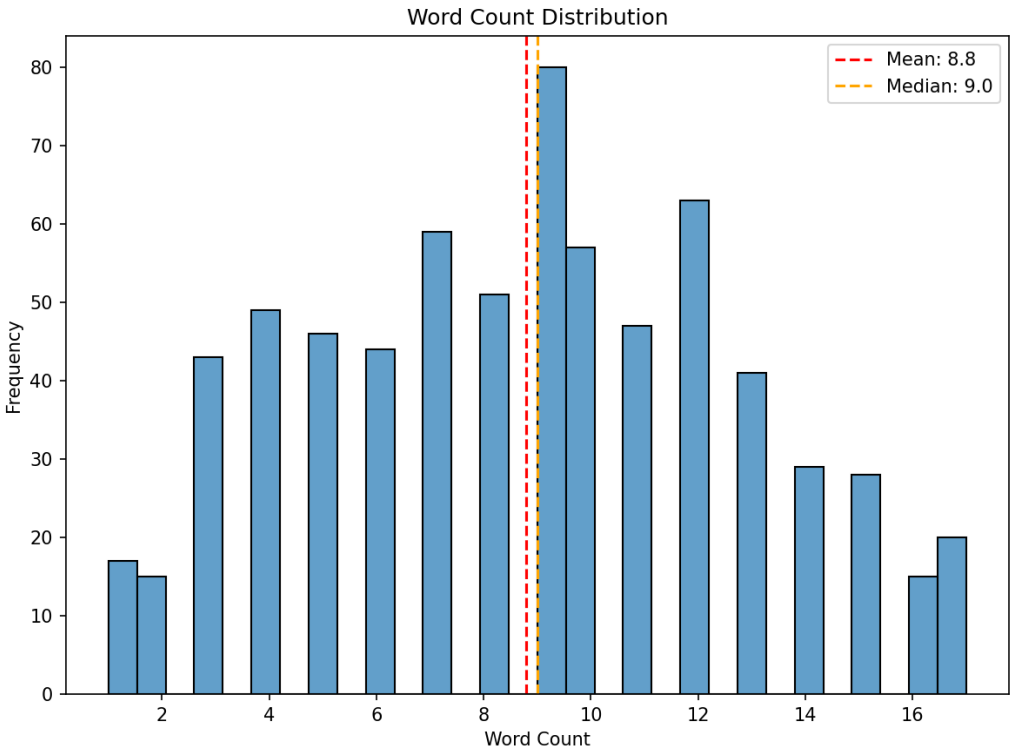

Figure 4: Word count distribution for PolitiFact.

PolitiFact data exhibits the most constrained length distribution among our datasets, with an average of just 8.8 words, and median is 9.0. This brevity is entirely consistent with the dataset's focus on discrete political claims and fact-check statements rather than full articles. The tight clustering around such short lengths reflects the nature of political fact-checking, where statements are typically extracted as concise, verifiable claims. This characteristic brevity made 128 tokens more than sufficient for encoding the entirety of most PolitiFact samples.

5.1.4 LIAR

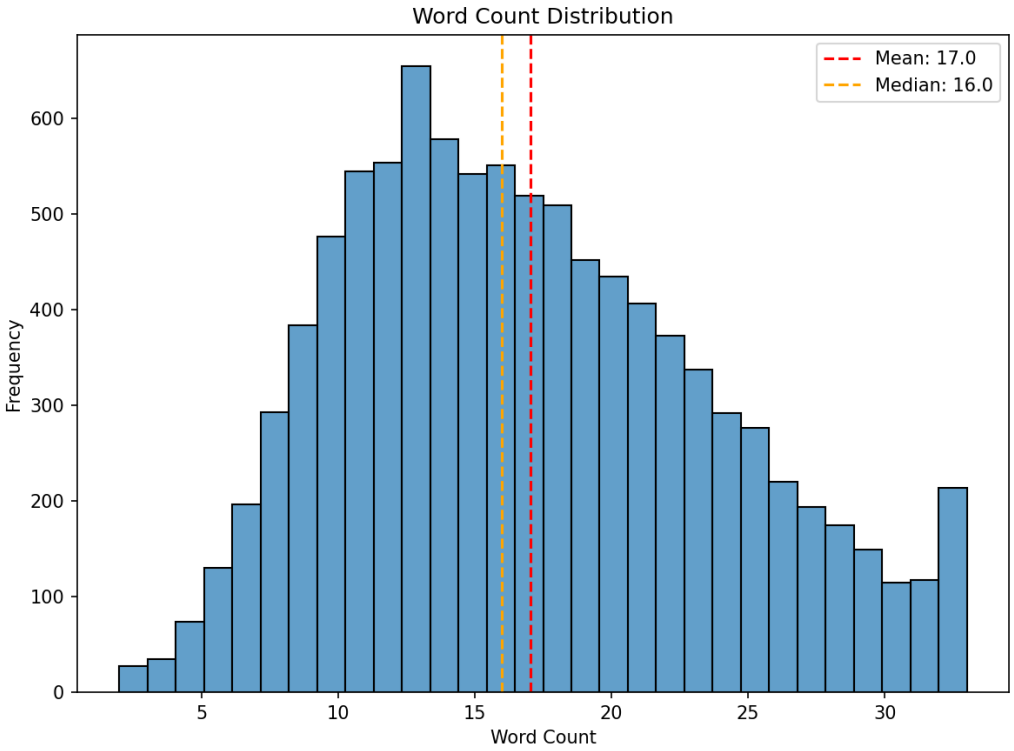

Figure 5: Word count distribution for LIAR dataset.

The LIAR dataset occupies a middle ground between the extremes we observed in other datasets. With a mean of 17 words, and median of 16, it represents moderately-sized statements that are longer than PolitiFact's brief claims but shorter than COVID-19's more variable content. The distribution appears well-balanced across different truthfulness labels, suggesting that statement length isn't a primary indicator of veracity in this political context. The symmetric nature of the distribution and moderate length led us to adopt 128 tokens as our sequence limit, which proves adequate for capturing the vast majority of LIAR samples without unnecessary padding.

5.2 Evaluation Metrics

In our binary classification setting, we label fake news as class 1 and real news as class 0. To evaluate model performance, we use standard classification metrics based on the confusion matrix:

- **True Positive (TP)**: The model correctly predicts a news piece as fake when it is actually fake.
- **True Negative (TN)**: The model correctly predicts a news piece as real when it is actually real.
- **False Positive (FP)**: The model incorrectly predicts a news piece as fake when it is actually real.

- **False Negative (FN)**: The model incorrectly predicts a news piece as real when it is actually fake.

Given these definitions, we compute the following evaluation metrics:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

5.3 General Configuration

To optimize training efficiency and resource usage, we adopt a maximum sequence length of 128 tokens uniformly across the COVID-19, GossipCop, PolitiFact, and LIAR datasets. This threshold ensures that over 95% of all samples remain untruncated after subword tokenization, preserving essential content while enabling efficient mini-batching.

Tokenization is performed using the pre-trained RoBERTa and XLNet tokenizers from the HuggingFace Transformers library. We apply subword tokenization with tail truncation ("truncation=True") and dynamic padding ("padding='longest"), ensuring optimal memory allocation and the retention of prefix-critical information.

Training is conducted with a batch size of 64, incorporating shuffling when datasets are combined across domains to encourage domain-invariant learning. We use the AdamW optimizer with a weight decay of 0.01 to prevent overfitting. The learning rate is set to `2e-5`, and training is stabilized using early stopping and warmup steps equal to 10% of the total training steps, promoting smoother convergence and preventing premature divergence in early iterations.

5.4 Pretrained Models

We employed pre-trained language models from the HuggingFace Transformers library, including `"roberta-base"` for RoBERTa, `"xlnet-base-cased"` for XLNet, and `"deberta-v3-base"` for DeBERTa. All models are available through the HuggingFace Transformers library and provide strong contextualized representations suited for downstream classification tasks using a classification head on top of the pretrained encoder.

5.5 Results

To comprehensively evaluate model performance, we conducted three experimental setups, all targeting binary fake news classification but differing in training strategy and model architecture.

5.5.1 Domain-Specific Fine-Tuning (DSFT)

In the first setting, each model was fine-tuned separately on individual domain-specific datasets, including COVID-19, LIAR, PolitiFact, and Gossip. This approach allows each model to specialize in the linguistic and topical characteristics of its corresponding domain. The results are shown in Table 6. This evaluation setting aims to measure how well each model adapts to domain-specific traits and to establish a performance baseline before investigating cross-domain or unified approaches.

Although CT-BERT currently achieves state-of-the-art performance on the CONSTRAINT2021 COVID-19 dataset, its effectiveness is limited to COVID-related contexts. CT-BERT is pretrained exclusively on COVID-19 tweets, making it less suitable for cross-domain fine-tuning on datasets like LIAR or PolitiFact, which differ significantly in language style, topic distribution, and source characteristics. Moreover, the top-performing CT-BERT ensemble relies on adversarial training and heated-up softmax loss, which substantially increases computational overhead. Due to resource constraints, we were unable to reproduce or extend these training procedures across other domains.

Meanwhile, the BERT-Base + CNN model achieves state-of-the-art performance on the LIAR dataset by integrating additional metadata such as emotion, specific features, and sentiment. This enriched input allows the CNN to capture nuanced contextual signals, leading to improved accuracy (70%) and F1-score (0.637).

RoBERTa yields state-of-the-art results on both GossipCop and PolitiFact, demonstrating strong domain adaptability. Notably, even when trained on the same combined dataset and using identical model architecture, slight differences in performance are still observed across experimental runs. These variations are likely due to factors such as random initialization, shuffling of training batches, and non-deterministic behavior in GPU computation. Such stochasticity in the training process can lead to minor fluctuations in evaluation metrics, even under controlled conditions.

Dataset	Metric	XLNet	RoBERTa	DeBERTa	SOTA
COVID-19	Accuracy	96.07	97.85	97.35	98.70 [6]
	Precision	95.97	98.80	99.32	98.91 [6]
	Recall	95.78	96.67	95.10	98.33 [6]
	F1-score	95.88	97.72	97.16	98.62 [6]
PolitiFact	Accuracy	84.91	89.94	85.14	86.16 [9]
	Precision	91.84	92.98	77.14	85.91 [9]
	Recall	69.23	81.54	90.00	86.16 [9]
	F1-score	78.95	86.89	83.08	78.56 [9]
LIAR	Accuracy	68.35	60.93	65.14	70.00 [29]
	Precision	72.54	75.84	70.47	–
	Recall	82.03	57.95	79.02	–
	F1-score	76.99	65.70	74.50	63.00 [29]
GossipCop	Accuracy	86.27	86.36	86.61	86.16 [9]
	Precision	73.92	71.68	76.27	85.91 [9]
	Recall	66.33	71.59	64.30	86.16 [9]
	F1-score	69.92	71.63	69.77	78.56 [9]

Table 5: Performance of Domain-Specific Fine-Tuning (DSFT) on Individual Datasets

5.5.2 Pooled-Domain Fine-Tuning (PDFT)

In the second setting, we combined all domain-specific datasets into a single dataset that includes samples from multiple domains. Each model (RoBERTa, XLNet, and DeBERTa) was then trained on this combined dataset to encourage learning of more generalized representations across heterogeneous data. While performance improved on underrepresented domains, it slightly decreased on domains with strong stylistic uniqueness, suggesting a trade-off between generality and specialization. The final results are shown in Table 6, highlighting the trade-off between generalization and domain-specific performance.

Dataset	Metric	XLNet-PDFT	RoBERTa-PDFT	DeBERTa-PDFT
COVID-19	Accuracy	94.16	96.78	97.06
	Precision	94.88	98.27	97.90
	Recall	92.75	94.90	95.88
	F1-score	93.80	96.56	96.88
PolitiFact	Accuracy	79.25	84.91	81.76
	Precision	74.24	83.61	83.33
	Recall	75.38	78.46	69.23
	F1-score	74.81	80.95	75.63
LIAR	Accuracy	65.75	63.77	64.40
	Precision	70.47	75.18	73.68
	Recall	80.81	65.53	69.80
	F1-score	75.28	70.02	71.69
GossipCop	Accuracy	85.94	86.78	86.51
	Precision	76.43	76.24	75.47
	Recall	60.08	65.46	65.08
	F1-score	67.27	70.44	69.89

Table 6: Generalization Performance from PDFT Models Across Domains

5.5.3 Domain-Matched Ensemble (DME)

In the third setting, we employed a score-level ensemble strategy that combines the predictions of three different models (BERT, RoBERTa, and DeBERTa), all fine-tuned on the same pooled-domain dataset.

Each expert independently processes the input using its own fine-tuned backbone, and the final prediction is derived by aggregating their outputs. This ensemble approach leverages the complementary strengths of different domain experts, enhancing robustness and improving predictive performance across diverse input styles. Experimental results demonstrated that the ensemble consistently outperformed individual experts as well as the joint-training baseline, as summarized in Table 7.

Experimental results demonstrated that the ensemble consistently improved accuracy and precision over individual experts as well as the joint-training baseline, indicating better reliability in predicting fake news when flagged. However, the recall and F1-score remained comparable to the base models, suggesting that while the ensemble is more confident in its predictions, it may still miss certain fake news instances. These findings highlight the trade-off between precision and recall, and suggest that the ensemble approach is particularly suitable for applications where minimizing false positives is prioritized.

Dataset	Metric	DME
COVID-19	Accuracy	97.48
	Precision	98.59
	Recall	96.08
	F1-score	97.32
PolitiFact	Accuracy	86.16
	Precision	89.09
	Recall	75.38
	F1-score	81.67
LIAR	Accuracy	67.40
	Precision	74.49
	Recall	75.31
	F1-score	74.89
GossipCop	Accuracy	87.53
	Precision	80.51
	Recall	63.58
	F1-score	71.05

Table 7: Performance from DME Model Across Domains

6 Conclusions

Our results show that domain-specific models achieve the best performance within their respective domains, highlighting the benefit of specialization when domain boundaries are clear. In contrast, multi-domain training offers moderate generalization but often underperforms in domain-specific evaluation due to conflicting patterns across datasets. The ensemble approach effectively combines specialized knowledge, yielding notable gains in accuracy and precision, especially in complex or ambiguous samples. These findings suggest that no single strategy universally dominates; instead, the choice depends on the deployment scenario. For high-precision use cases, ensemble models are preferable, while for cross-domain scalability, joint training offers a practical compromise. Future work will explore adaptive routing mechanisms and knowledge distillation to unify domain-specific expertise under a single, efficient model.

A lightweight GUI has also been developed to support model evaluation and inference. For implementation details, source code, and extended documentation, please visit **our GitHub repository**.

7 Contribution

To ensure effective collaboration and leverage individual strengths, each member of the team was assigned specific tasks aligned with their expertise. The contributions are summarized as follows:

- **Ho Bao Thu - 20226003 (Team Leader)**
  - Conduct survey on models and architectures.
  - Analyze selected models in detail.
  - Fine-tune and evaluate models.
  - Develop and showcase final project report.
- **Nguyen Dinh Duong - 20225966**
  - Conduct survey on datasets for fake news detection.
  - Analyze selected datasets in detail.
  - Handled data preprocessing pipeline.
  - Developed and integrated the GUI application.
- **Nguyen My Duyen - 20225967**
  - Conduct survey on models and architectures.
  - Analyze selected models in detail.
  - Develop and showcase final project report.
- **Tran Kim Cuong - 20226017**
  - Conduct survey on models and architectures.
  - Fine-tune and evaluate models.
  - Develop and showcase final project report.
- **Ha Viet Khanh - 20225979**
  - Conduct survey on models and architectures.
  - Analyze selected models in detail.
  - Fine-tune and evaluate models.
  - Develop and showcase final project report.
- **Dang Van Nhan - 20225990**
  - Conduct survey on datasets for fake news detection.
  - Performed exploratory data analysis (EDA).
  - Fine-tune and evaluate models.
  - Develop and showcase final project report.

References

[1] Rodrigo Barbado, Oscar Araque, and Carlos A. Iglesias. A framework for fake review detection in online consumer electronics retailers. *Information Processing & Management*, 56(4):1234–1244, 2019.

[2] Christina Boididou, Symeon Papadopoulos, Markos Zampoglou, Lazaros Apostolidis, Olga Papadopoulou, and Ioannis Kompatsiaris. Detection and visualization of misleading content on twitter. *International Journal of Multimedia Information Retrieval*, 7, 03 2018.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186. Association for Computational Linguistics, 2019.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT*, 2019.

[5] Alessio D'Ulizia, Anna Maria Rinaldi, and Massimo Mecella. Fake news detection: a survey of evaluation datasets. *PeerJ Computer Science*, 7:e603, 2021.

[6] Anna Glazkova, Maksym Del, Artem Shelmanov, and Aleksandr Panchenko. g2tmn at constraint@aaai2021: Exploiting ct-bert and ensembling learning for covid-19 fake news detection. In *Proceedings of the First Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations (CONSTRAINT)*, Punta Cana, Dominican Republic, 2021. Springer. arXiv preprint arXiv:2012.11967.

[7] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*, 2021.

[8] Rajesh Kumar Kaliyar, Ashish Goswami, and Prerna Narang. Fakebert: Fake news detection in social media with a bert-based deep learning approach. *Multimedia Tools and Applications*, 80(8):11765–11788, 2021.

[9] Sai Venkatesh Kuntur, Geetha Parthasarathy, et al. Comparative analysis of graph neural networks and transformers for robust fake news detection. *Electronics*, 13(23):4784, 2024.

[10] Soveatin Kuntur, Anna Wróblewska, Marcin Paprzycki, and Maria Ganzha. Fake news detection: It's all in the data! *arXiv preprint arXiv:2407.02122*, 2024.

[11] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, volume 86, pages 2278–2324. IEEE, 1998.

[12] Yinhan Liu, Danqi Chen, Jingfei Du, Mandar Joshi Li, Xiao Liu, Yang Ma, Myle Ott, Naman Goyal, Omer Levys, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[13] Manav S Lohabade, Yash W Borde, Rutwik N Kharwadkar, and Aniket M Dhakite. A literature survey on different methods for fake news detection. *International Journal of Creative Research Thoughts (IJCRT)*, 12:551–560, 2024.

[14] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2013.

[15] Tanushree Mitra, Graham Wright, and Eric Gilbert. Credbank: A large-scale social media corpus with associated credibility annotations. In *Proceedings of the 9th International AAAI Conference on Web and Social Media (ICWSM)*, pages 258–267, 2015.

[16] Taichi Murayama. Dataset of fake news detection and fact verification: A survey. *arXiv preprint arXiv:2111.03299*, 2021.

[17] Kai Nakamura, Sharon Levy, and William Yang Wang. r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 6085–6093. European Language Resources Association (ELRA), 2020.

[18] Jeppe Norregaard, Benjamin D Horne, and Sibel Adali. Nela-gt-2018: A large multi-labelled news dataset for the study of misinformation in news articles. In *Proceedings of the 13th International AAAI Conference on Web and Social Media (ICWSM 2019)*, pages 630–638. Association for the Advancement of Artificial Intelligence (AAAI), 2019.

[19] Parth Patwa, Shivam Sharma, Srinivas Pykl, Vineeth Guptha, Gitanjali Kumari, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. Fighting an infodemic: Covid-19 fake news dataset. In *Proceedings of the First Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations (CONSTRAINT)*, pages 21–29. Association for Computational Linguistics, 2021.

[20] Parth Patwa, Shrey Sharma, Snehal Sakhare, Satnam Singh, Amrita Sakhare, and Varsha Solanki. Fighting covid-19 fake news: Using machine learning to detect misinformation on social media. In *2020 5th International Conference on Smart Solutions for Urban Evacuation and Emergency Responses (SSUER)*, pages 1–6. IEEE, 2020.

[21] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[22] Natali Ruchansky, Sungyong Seo, and Yan Liu. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 797–806, 2017.

[23] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

[24] Giovanni Luca Ciampaglia Santia and Jesse Leigh Williams. Buzzface: A news veracity dataset with facebook user commentary and egos. In *Proceedings of the 12th International AAAI Conference on Web and Social Media (ICWSM)*, pages 531–540, 2018.

[25] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725, 2016.

[26] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. Fakenewsnet: A data repository with news content, social context and spatialtemporal information for studying fake news on social media. *Big Data*, 8(3):171–188, 2020.

[27] Eugenio Tacchini, Gabriele Ballarin, Marco L. Della Vedova, Stefano Moret, and Luca de Alfaro. Some like it hoax: Automated fake news detection in social networks. In *Proceedings of the Second Workshop on Data Science for Social Good (SoGood 2017)*, volume 1960 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2017.

[28] Fatemeh Torabi Asr and Maite Taboada. Big data and quality data for fake news and misinformation detection. *Big Data & Society*, 6(1):1–14, 2019.

[29] Bibek Upadhayay and Vahid Behzadan. Sentimental liar: Extended corpus and deep learning models for fake claim classification. *arXiv preprint arXiv:2009.01047*, 2020.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008, 2017.

[32] William Yang Wang. Liar: A benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 422–426, 2017.

[33] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[34] Wenya Zhang and Yongtao Li. An overview of deep learning based fake news detection. *Journal of Physics: Conference Series*, 1533(4):042079, 2020.

[35] Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLOS ONE*, 11(3):e0150989, 2016.