Truth, Lies and Language Models: A New Approach to Detecting AI-Generated Fake News

Candidate Number: 49187

Supervisor: Milena Tsvetkova

06/08/2025

Contents

Abstract

The emergence of powerful Large Language Models (LLMs) has made misinformation more convincing and harder to detect, posing serious risks to public trust and information integrity. This study addresses this challenge through two key contributions: (1) the construction of a large, multi-source dataset combining both human and and LLM-generated news content, and (2) the development of a dual-headed BERT-based model that jointly classifies whether a news item is Fake or Real and whether it was written by a human or an LLM. The model achieves 85% accuracy in Fake-vs-Real classification and 96% in Human-vs-LLM detection. Further the model is evaluated on content from multiple LLM's and various promting strategies, demonstrating strong generalization and robustness in detecting modern AI-generated fake news. For this reason, it seems to outperform existing detectors, since the last fail to generalize across sources.

1 Introduction

The advancement of the internet has brought numerous benefits, particularly in enhancing communication and broadening access to information. Nonetheless, the same digital progress has also facilitated the widespread circulation of false information through social media, news websites, and other digital channels. The speed at which misinformation spreads poses a serious societal concern and contributes to public unease. False narratives can heavily impact public opinion, influence political outcomes, and endanger health and safety, as was clearly demonstrated during the COVID-19 crisis. Although various measures have been implemented to detect and limit the spread of fake news, it remains a persistent and urgent threat to society (Denniss and Lindberg, 2025). As such, continued research in this field is both necessary and highly relevant.

Over the past ten years, Large Language Models (LLMs) have seen rapid and unprecedented development (Naveed et al., 2024), driven by advances in hardware, particularly improvements in Graphical Processing Units (GPUs), and access to vast amounts of training data. These models have become widely adopted, with OpenAI's ChatGPT, for instance, reaching an estimated 100 million monthly users within just two months of its release in early 2023, making it a rapidly-growing application. While LLMs offer considerable advantages across various domains, they also present substantial risks when misused. Their ability to generate large volumes of coherent, contextually relevant, and persuasive text makes them especially well-suited for producing disinformation. Unlike earlier fake news, which often contained grammatical errors or lacked fluency, content generated by LLMs can easily mimic authentic news reports with high credibility and sophistication (Su et al., 2023). This evolution significantly complicates detection efforts, as it necessitates identifying nuanced semantic and stylistic features beyond traditional surface-level cues (Hu et al., 2025b).

As the challenges posed by misinformation continue to evolve, so too have the methods developed to detect it (Hu et al., 2025a). Researchers are increasingly employing sophisticated techniques, with particular emphasis on machine learning, deep learning, and natural language processing. A central question in current studies is whether state-of-the-art detection systems- especially those built on transformer-based architectures (Vaswani et al., 2017)- can reliably differentiate between fake content produced by LLMs and fake content produced by humans.

Thus, the main research question of this study is the following: Can a classification model accurately detect LLM-generated fake news? In order for this to be answered, this study will try to implement the following objectives in the order they are mentioned below:

- It is a hypothesis of this study that most models in the public domain that are fine-tuned for fake news detection, have low generalizability on other sources other than those they were trained on. This points to the need for more varied training datasets. For this reason, the first objective in this study is to construct a robust, high-quality dataset sourced from multiple sources to support effective fake news detection.

- Design and implement a range of deep learning model architectures for the binary classification of news as real or fake, and assess their effectiveness using standard evaluation metrics.

- After finding the best model for fake news detection it will seek to broaden its scope and architecture to also, simultaneously, classify whether content has been authored by a human or generated by an LLM (i.e., Human-vs-LLM binary classification), and assess its effectiveness.

- Finally it will explore the potential of different Large Language Models and different prompting strategies to produce fake news that can bypass the model (classify them wrongly).

2 Related Work

2.1 The model side of misinformation

There are multiple approaches to detecting fake news, as outlined in the comprehensive survey by (Zhou and Zafarani, 2018). Their work evaluates various detection strategies based on four key aspects: (1) the deceptive content itself, (2) stylistic and linguistic features, (3) patterns of dissemination, and (4) the trustworthiness of the information source. The study highlights the importance of combining language-based indicators with behavioral and social cues. Early research in this field relied on traditional machine learning methods such as Support Vector Machines (SVM) and Logistic Regression, utilizing manually engineered features like sentiment analysis, readability scores, and TF-IDF values (Rubin et al., 2016);(Potthast et al., 2018). One noteworthy finding is that clickbait often accompanies false or misleading articles, making it a useful signal for detection (Indurthi et al., 2020). Indurthi et al's research supports integrating clickbait detection (focused on identifying exaggeration language and curiosity) into broader fake news classification frameworks. Further studies emphasize that social network dynamics and content propagation behavior play a critical role in detection accuracy, with some models achieving up to 92.7% ROC AUC. Remarkably, these models can often identify fake news within just a few hours of its initial spread (Monti et al., 2019). Nevertheless, although these earlier models are generally interpretable and relatively efficient, they often struggle with nuanced language patterns and can be resource-intensive compared to modern deep learning methods.

Deep learning approaches (especially those utilizing transformer-based architectures) have significantly advanced the field of fake news detection. Models like BERT (Devlin et al., 2019), which generate contextualized word embeddings and support transfer learning, have set new standards for performance in this domain. BERT is designed to pre-train deep bidirectional representations by simultaneously considering both the left and right context of a word across all layers. This allows the model to be fine-tuned on domain-specific tasks even with relatively few labeled examples, often resulting in high levels of accuracy. For example, (Kaliyar et al., 2021) demonstrated that BERT-based models significantly outperformed conventional approaches when applied to real-world news classification tasks. In addition, further enhancement has been shown through bidirectional training methods. One such study introduced Deep Fake BERT, a model that incorporates multiple parallel components with varying kernel sizes and strides into a single-layer Diffusion-Convolutional Neural Network (DCNN), which leverages BERT for deep representation learning (processing DCNN'S outputs) (Kanchana et al., 2023).

Incorporating visual elements such as images and attachments alongside textual content has proven effective in enhancing fake news detection using Convolutional Neural Networks (CNNs). One notable approach is the Text and Image-based Convolutional Neural Network (TI-CNN), which simultaneously processes both textual and visual data. By mapping both explicit (i.e. sentence length, punctuation counts etc) and latent features (i.e. hidden patterns learned automatically via convolutional layers applied to text embeddings and image data) into a unified feature space, TI-CNN achieves strong performance in identifying fake news (Yang et al., 2018). While TI-CNN fused handcrafted and learned features across modalities via simple concatenation, new state-of-the-art transformer-based models like CFNN (Li et al., 2023) use a cross-attention layer that aligns semantic features between text and image, then identifies hidden mismatches, and this is far more nuanced than simple stacking.

Recurrent Neural Networks (RNNs) have also been leveraged, particularly for their ability to model sequential dependencies in text or patterns in user behavior, thus enabling context-aware learning that improves over time. A prime example is the CSI model, which integrates user interaction data with article content. It evaluates the reliability of sources through user behavior patterns and employs an RNN to model temporal dynamics. Results indicate that CSI outperforms existing techniques and produces informative representations of both users and articles when these components are jointly analyzed (Ruchansky et al., 2017). Alternative strategies focus on the source of the information rather than its linguistic or visual characteristics. For instance, fake news shared on Facebook can be accurately identified by analyzing the profiles that engage with such content. Research shows that using data from user interactions (such as 'likes') allows for high-accuracy classification, even when trained on less than 1% of the total dataset. One study achieved 99% accuracy on a dataset consisting of 15,500 Facebook posts and over 900,000 user profiles (Tacchini et al., 2017). These results highlight the potential of integrating dissemination patterns into automated misinformation detection systems.

2.2 The data side of misinformation

To support advancements in fake news detection, numerous datasets have been curated. One of the most prominent is the LIAR dataset, which serves as a benchmark for assessing the truthfulness of short political statements. It provides detailed labeling based on expert fact-checking, along with metadata such as the speaker's identity, the context of the statement, and the topic discussed (Wang, 2017). This rich contextual information allows classification models to leverage both textual and metadata features. The LIAR dataset played a critical role in training more sophisticated models like GROVER, which achieved 73% accuracy in differentiating between human-written news and content generated by neural networks (Zellers et al., 2019). A more comprehensive version, LIAR2, was released in 2024 to expand its scope and relevance (Xu and Kechadi, 2024). On the other hand, the Fakeddit dataset (Nakamura et al., 2019) offers a much larger and multi-modal collection of Reddit posts, covering diverse topics and formats. This dataset emphasizes both scale and modality, enabling experiments that incorporate textual and visual inputs.

Despite the value of such datasets, for fake news detection models to perform reliably, they must be trained on datasets that are both large in scale and rich in diversity. This diversity enables models to effectively capture patterns and generalize across a wide range of topics, languages, and media sources. However, much of the existing research relies on single-source datasets, which often suffer from limitations in terms of size, modality, and level of detail (Xu and Kechadi, 2024) (Nakamura et al., 2019) . In this work, I argue that integrating a broader variety of input data can significantly improve the resilience and scalability of fake news detection systems.

Although many models in the literature have demonstrated strong performance (above 95% accuracy), they are often trained and evaluated on relatively small or narrowly focused datasets. This limits their ability to generalize to real-world scenarios, especially when faced with content from unfamiliar domains or writing styles. Table 1 presents the testing performance of a selection of three famous fake news detection models [1] on a combined dataset from various sources (the one which is going to be used in this study and described in detail in the next section). These models were initially trained on smaller datasets, and, although they achieved over 95% accuracy in their respective datasets, in Table 1 we see that, interestingly, they exhibit far worse accuracy on unseen data than their originally stated one.

Furthermore, Table 2 breaks down these models' accuracy across individual datasets, clearly demonstrating their inconsistency and limited generalizability when applied to data distributions they were not explicitly trained on. While the results in Table 1 show that none of the evaluated models exceed 51% accuracy on the full combined dataset, suggesting a significant struggle when faced with heterogeneous data, Table 2 reveals deeper weaknesses. For instance, although performance on LIAR 2 appears moderate (around 58%), the same models drop significantly on other datasets such as Kaggle 2 or Kaggle 3, with some accuracies falling below 40%. This performance variance suggests that these models may be overfitting to the style, structure, or topic distribution of the

[1] Example 1, Example 2, Example 3

dataset they were originally trained on, and therefore lack the robustness needed for real-world deployment. The inconsistency across datasets, despite being from the same domain (news), highlights the critical need for constructing more diverse training corpora, encompassing a variety of formats, sources, and linguistic patterns. These findings justify the central methodological decision of this study: to develop and train on a composite dataset that integrates posts and news articles from multiple sources, with the aim of achieving better generalization and more reliable classification across varied contexts. The study's dataset is described in detail in the next section (3.1).

Dataset	**Model**	**Accuracy**
ALL	Fake-News-Bert-Detect	0.50290
ALL	fake-news-classification-distilbert-fine-tuned	0.47860
ALL	albert-base-v2-fakenews-discriminator	0.48685

Table 1: Performance of Pretrained Fake News Detection Models on Combined Dataset

Dataset	**Model**	**Accuracy**
Fakeddit	Fake-News-Bert-Detect	0.45975
Fakeddit	fake-news-classification-distilbert-fine-tuned	0.45725
Fakeddit	albert-base-v2-fakenews-discriminator	0.48750
Kaggle 1	Fake-News-Bert-Detect	0.58250
Kaggle 1	fake-news-classification-distilbert-fine-tuned	0.52650
Kaggle 1	albert-base-v2-fakenews-discriminator	0.54175
Kaggle 2	Fake-News-Bert-Detect	0.44500
Kaggle 2	fake-news-classification-distilbert-fine-tuned	0.44325
Kaggle 2	albert-base-v2-fakenews-discriminator	0.45525
Kaggle 3	Fake-News-Bert-Detect	0.44625
Kaggle 3	fake-news-classification-distilbert-fine-tuned	0.38600
Kaggle 3	albert-base-v2-fakenews-discriminator	0.40600
LIAR 2	Fake-News-Bert-Detect	0.58100
LIAR 2	fake-news-classification-distilbert-fine-tuned	0.58000
LIAR 2	albert-base-v2-fakenews-discriminator	0.54375

Table 2: Performance of Pretrained Fake News Detection Models by Dataset

2.3 LLM's and misinformation

The rise of advanced LLMs has intensified concerns about AI-driven misinformation, prompting a surge of research into detecting machine-generated fake news. Recent studies highlight that text generated by LLMs can be more deceptive and harder to spot than human-written text (Chen and Shu, 2024) (Sallami et al., 2024). For example, Chen and Shu (2024) found that LLM-crafted misinformation often adopts a highly convincing style that fools both human readers and automated detectors more often than equivalent human-written fake news (Chen and Shu, 2024). Similarly, Sallami et al. (2024) report that LLM-generated fake news is less likely to be detected than human-created ones, underscoring the unique challenge posed by AI-authored disinformation (Sallami et al., 2024). These findings suggest that traditional cues for fake news (e.g. grammatical errors or incoherence) are no longer reliable, as modern language models produce fluent and contextually plausible narratives. Detecting such content requires more nuanced analysis of semantic and stylistic patterns beyond surface-level anomalies.

In response, researchers have developed a spectrum of detection approaches tailored to AI-generated text. A comprehensive survey by (Wu et al., 2024) categorizes LLM-output detection methods into four broad strategies: (1) watermarking, (2) statistical analysis, (3) neural network classifiers, and (4) human-in-the-loop techniques. Watermarking methods embed hidden signals in LLM outputs to later verify their origin, offering a proactive defense if such markers can be widely implemented. In contrast, statistical detectors look for distributional irregularities in text. For instance, the GLTR tool (2019) flags passages that are "too predictable" under a language model's probability distribution (Gehrmann et al., 2019). Gehrmann et al. showed that highlighting tokens with unusually high model likelihood helped humans boost their detection accuracy from 54% to 72%, by revealing the telltale overuse of common words in AI-written text. Neural network-based detectors form another key line of defense. These involve fine-tuning Transformer-based classifiers (e.g. BERT or RoBERTa) on labeled datasets of human and AI text, enabling the model to automatically learn subtle differentiating features. Such learned detectors have achieved strong results, especially when trained on outputs from the latest generators. A seminal example is Grover (Zellers et al. 2019), a large Transformer originally built to produce fake news. Grover's authors demonstrated that the same model could be used to detect neural fake news with up to 92% accuracy, far outperforming earlier discriminators that achieved 73% on the same task (Zellers et al., 2019). This showed that powerful generative models can double as effective discriminators when provided with sufficient training data. Finally, human-in-the-loop approaches integrate human judgment or feedback into the detection pipeline. Tools like GLTR fall in this category, as do recent frameworks where human experts post-edit or verify AI-flagged content. While human insight remains vital, these methods acknowledge that purely manual detection is impractical at scale and often error-prone (e.g. unassisted people barely perform better than chance on GPT-generated news).

Beyond developing detectors, researchers have also started simulating LLM-driven misinformation to study its characteristics. For example, Guo et al. (2024) introduced the MegaFake dataset of AI-generated news, leveraging a theory-driven framework (LLM-Fake Theory) to produce fake articles via four strategies: content manipulation, integration of truth and lies, entirely fictional stories, and style imitation (Wang et al., 2024). Intriguingly, their experiments found that traditional natural language understanding (NLU) models (i.e. classifiers) significantly outperformed natural language generation (NLG) models in detecting these AI-crafted fake news. However, they also observed that detectors trained on human-written fake news did not generalize well to LLM-generated fake news and vice-versa, reflecting a distribution shift between human and AI distributions. This aligns with broader concerns that detectors can struggle on unseen domains or novel model outputs.

2.4 Literature summary and main contributions of the present study

The literature on fake news detection has evolved from traditional machine learning using handcrafted features to deep learning, especially transformer-based models like BERT. These models have significantly improved detection accuracy by capturing contextual semantics. However, many are trained on narrow datasets (e.g., LIAR, Fakeddit), limiting their generalization to real-world scenarios. Tables 1 and 2 confirm that popular pretrained models underperform when evaluated on a multi-source dataset, revealing the risk of overfitting to specific formats or domains.

Simultaneously, the emergence of Large Language Models (LLMs) has introduced a new threat: highly fluent, machine-generated fake news. Recent studies show that LLM-generated misinformation is harder to detect than human-written content. While existing solutions include watermarking, statistical analysis, and fine-tuned transformers, most struggle to generalize across different LLMs or generation strategies.

This thesis addresses these limitations with four main contributions: (1) constructing a diverse, balanced dataset from six sources, (2) designing a dual-head BERT model for joint Fake-vs-Real and Human-vs-LLM classification, (3) demonstrating the model's strong generalization across LLMs and misinformation strategies (e.g., MegaFake), and (4) achieving state-of-the-art accuracy and F1 scores on both tasks. Together, these results affirm the central hypothesis of this study: LLM-generated fake news can be accurately detected when models are trained on sufficiently diverse data and tasked with learning both factual accuracy and whether the content was human- or machine- generated.

3 Methodology

3.1 Data Description and Processing

The datasets utilised in this study fall into two categories:

1. Human-authored content with binary true/fake classification labels

   - **Fakeddit:** The Fakeddit dataset (Nakamura et al., 2019) contains over 1,000,000 Reddit posts collected from 22 distinct subreddits, covering a time span from March 2008 to October 2019. This extensive, decade-long dataset offers a wide variety of linguistic and cultural content, making it highly valuable for training fake news detection models. Each post is annotated using a hierarchical labeling system that supports binary (true vs. fake), ternary (3-way), and 6-class classification, allowing for both broad and nuanced analysis. Labels are determined based on the originating subreddit, for example, categories like "us-news" and "mildlyinteresting" are considered reliable (true). For my work, I focus specifically on the binary (2-way) classification scheme.

- **LIAR 2:** The LIAR dataset (Wang, 2017) is widely recognized and frequently used within the misinformation detection research community. It comprises approximately 13,000 short political statements related to U.S. politics, each annotated with contextual metadata and labeled by professional fact-checkers from PolitiFact. These labels are assigned based on a thorough verification process and fall into six categories: true, mostly true, half-true, barely true, false, pants on fire. The LIAR2 dataset (Xu and Kechadi, 2024) builds on this foundation, expanding the collection to around 23,000 samples. For this study's purposes, I transform the original six-class scheme into a binary classification one. Specifically, the ambiguous half-true category is excluded, while statements labeled mostly true and true are grouped as true, and those labeled false and pants on fire are grouped as fake.
- **Kaggle:** In addition, three separate datasets sourced from Kaggle were incorporated into the study (Abaghyan, 2025) (Kokiantonis, 2022) (Jruvika, 2018). These collections include a variety of news headlines and full-length articles drawn from media outlets across the English-speaking world, each annotated with binary true/false labels. Combined, these Kaggle datasets contribute approximately 54,000 additional samples to the overall dataset used in the analysis.

2. Content generated by LLM's

- **LLM-Generated News Dataset:** This dataset was specifically developed to explore how different content manipulation strategies interact with the capabilities of various LLMs (Ayoobi et al., 2024). It consists of 3,000 articles sourced from credible news sources, spanning six distinct subject areas. Each original article (marked as true) was altered using three types of manipulation techniques -rephrasing, content expansion, and summary elaboration- across four language models: Llama2-13B, Llama2-7B, Mistral-7B and GPT-3.5. The modified outputs produced by these models are labeled as fake. Both the original human-written articles and their manipulated counterparts are included in my analysis.

- **MegaFake:** The MegaFake dataset features four distinct categories of fake news alongside two types of authentic news content, all produced by LLMs. The generation process is guided by a conceptual model referred to as LLM-Fake Theory (Wang et al., 2024). Fake news content is generated automatically using carefully crafted prompt engineering pipelines, which are designed to simulate various misinformation strategies. The dataset includes news generated with four such prompting techniques: Story-Based, Style-Based, Content-Based and Integration-Based Fake News. Altogether, there are approximately 46,000 fake news articles to this dataset. In my analysis, I utilize this dataset in order to evaluate the model's performance on completely unseen data.

Once the datasets were gathered, several preprocessing steps were applied. To accommodate the GPU memory limitations of the training environment, each text sample was segmented into chunks containing no more than 60 tokens. This fixed token limit was a practical constraint based on available computational resources. The GPU graphics card that was used is an NVIDIA Geforce RTX 4050, which had just 6 GB VRAM. Future improvements to the preprocessing pipeline could explore more sophisticated chunking methods that better retain sentence boundaries and contextual flow. Tables 3 and 4 provide overviews of the human-generated and LLM-generated datasets, respectively.

Dataset	Samples	Real (%)	Avg String Length
Fakeddit	707,590	53.8	8.7
LIAR 2	22,936	42.3	17.6
Kaggle 1 - Fake News	300,913	52.7	53.1
Kaggle 2 - News Project	85,316	55.8	57.4
Kaggle 3 - Fake News Detection	28,763	62.2	55.3

Table 3: Human-generated data summary

Dataset	Samples	Real (%)	Avg String Length
LLM-Generated Fake News	477,809	29.5	58.6
MegaFake	203,530	0	56.7

Table 4: LLM-generated data summary

For the final dataset I combined all the aforementioned ones into a single dataset, except for the Megafake dataset (for reasons explained below). After that, I split it into a training and testing set. The test set was constructed by randomly selecting 4,000 samples from each dataset, ensuring a balanced representation. This deliberate design aimed to assess the model's capacity to generalize effectively across multiple and varied data sources. I assume that the model's "real" efficiency, is more accurately reflected this way, since there are not "easy" or "hard" data sources that overwhelm the test set. In addition, alongside the primary classification task of distinguishing "Fake" from "Real" news, I also introduced an additional target variable, "Human-vs-LLM," which identifies whether a given text was authored by a person or generated by a language model.

The class distributions for both classification targets in the training set are shown in Table 5, under the "Original" column. Notably, the "Human-vs-LLM" target exhibited a significant class imbalance. To address this, I applied oversampling to the minority class prior to model training. The adjusted class proportions following this balancing step are displayed under the Post-Oversampling column in the same table.

Finally, the MegaFake dataset (Wang et al., 2024), was excluded from the training phase and used exclusively for inference evaluation, in order to test whether the final model can perform equally well on completely unseen data, and across various prompting strategies

Label	Original (%)	Post-Oversampling (%)
Fake	52.6	63.8
Real	47.4	36.2
Human	81.5	60.5
LLM	18.5	39.5

Table 5: Label Distributions for Fake-vs-Real and Human-vs-LLM Targets Before and After Oversampling (Training Set)

3.2 Architecture

This study explored two primary categories of neural network architectures for misinformation detection: recurrent neural networks (RNNs) and transformer-based models. Although, as highlighted in the literature review, fine-tuned transformer models should be more suitable for fake news detection since they are the new state-of-the-art on this domain, I also explored other architectures (RNN) to ensure transparency and perform robustness checks. After trying several models, I came up with the final one which achieved very high accuracy. Below, I discuss the architectures of all the models used in this study.

3.2.1 RNN Architecture

For the RNN-based method, I developed and trained a bidirectional Long Short-Term Memory (BiLSTM) network. The architecture began with an embedding layer initialized using pre-trained GloVe vectors, allowing the model to capture semantic relationships in the text. To preserve the semantic quality of the embeddings, the embedding layer was frozen during training. This was followed by a BiLSTM layer (with 64 hidden units), which processed sequences in both forward and backward directions to better understand contextual dependencies. In addition, in order to make the model capture complex non-linear patterns I included two fully connected (dense) layers with 64 and 1 units, respectively. Each dense layer is preceded by a dropout layer with a rate of 0.3 to reduce overfitting. The first dense layer uses a ReLU activation function, while the final output layer uses a sigmoid activation function for binary classification.

3.2.2 Transformer-based Models Architecture

For the transformer-based approach, I chose to fine-tune a pre-trained language model. Specifically, I employed the bert-base-uncased version of BERT, which is composed of 12 transformer layers, each containing 12 self-attention heads and featuring an embedding dimension of 768. The model comprises a total of 109,482,240 trainable parameters. BERT is particularly effective for classification tasks due to its use of bidirectional attention, allowing each token in the input to consider the entire sequence during training, both forward and backward words. It is pre-trained on two objectives: masked language modeling and next sentence prediction. As a result, the special [CLS] token, which is placed at the beginning of every input sequence, learns a rich, context-aware embedding that encapsulates both local (surrounding words) and global (overall sequence) semantic information, making it ideal for downstream classification tasks.

When fine-tuning transformer models for classification tasks, a standard approach involves using the embedding of the [CLS] token as a condensed representation of the entire sequence. This embedding is typically passed through a linear layer followed by a non-linear activation function. The transformation can be represented mathematically as:

$$\alpha(XW + b), \tag{1}$$

where $X \in \mathbb{R}^{n\times d}$ denotes the matrix of [CLS] embeddings, $W \in \mathbb{R}^{d\times c}$ is the weight matrix, and $b \in \mathbb{R}^{1\times c}$ represents the bias term. Here, c corresponds to the number of output classes, n is the number of input sequences, d is the embedding dimension and α is the activation function. For this study here, the activation function is sigmoid since it deals with binary classification.

In addition, I fine-tuned DistilBERT, a more compact and efficient alternative to the original BERT model. DistilBERT preserves much of BERT's structural design and effectiveness, but with significantly reduced computational overhead—approximately 40% fewer parameters and up to 60% faster in execution. This efficiency is achieved by using only 6 transformer layers instead of 12 and by omitting the next sentence prediction task during pre-training.

Finally, the last model, which is one of this study's main contributions, is a modification of the BERT-base architecture by introducing a second classification head, enabling the model to learn two tasks simultaneously: Fake-vs-Real classification and Human-vs-LLM classification. This design choice aligns with the structure of my dataset, which contains both real and fake content, as well as text authored by humans and generated by LLMs. Also, it is the most suitable to help answer the main question of this study, which is if we can accurately detect LLM-generated fake news. Each classification head processes the [CLS] token independently using the transformation method previously described and applies a sigmoid activation to produce binary outputs.

Both heads are trained using the Binary Cross-Entropy loss function. During training, the model jointly optimizes the combined objective, with the total loss defined as the sum of the individual task losses. The gradient with respect to any model parameter θ is computed as:

$$\frac{\partial L_{total}}{\partial \theta} = \frac{\partial (L_1 + L_2)}{\partial \theta} = \frac{\partial L_1}{\partial \theta} + \frac{\partial L_2}{\partial \theta} \tag{2}$$

where L_1 and L_2 correspond to the loss values from the Fake-vs-Real and Human-vs-LLM classification tasks, respectively. Thanks to the linearity of derivatives, as seen in Equation 2, this setup allows the model to update its parameters based on feedback from both learning objectives simultaneously. Also, each task contributes equally to the total loss. Finally, a shared dropout layer with a rate of 0.1 is applied to the [CLS] token embedding before it is passed to the two classification heads, in order to prevent overfitting. The final architecture contains 109,483,778 trainable parameters. An overview of the model structure is illustrated in Figure 1.

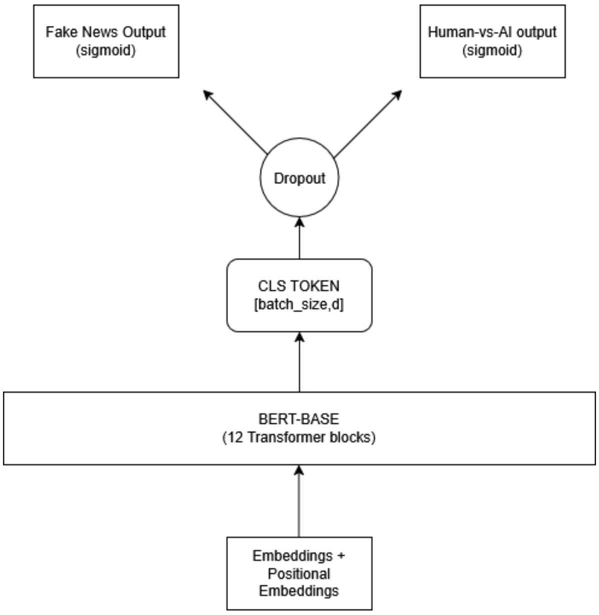

Figure 1: Architecture of the Dual-Head BERT-base Model

3.3 Training

Before training the models, all input text was tokenized using each model's specific tokenizer. In addition, sequences were standardized to a fixed length of 60 tokens in order to stay within the limits of available GPU memory, as mentioned above.

For the RNN-based implementation, I employed the Keras tokenizer to process the text data and convert it into integer sequences. The vocabulary was limited to the 20,000 most frequent words, with an out-of-vocabulary (OOV) token specified to handle any unseen or rare words. Each sequence was padded and truncated to ensure a consistent length of 60 tokens. The model was trained using the Adam optimizer with a learning rate of 10^{-3}, a batch size of 32, and binary cross-entropy as the loss function. Training was conducted for a maximum of 10 epochs, with early stopping based on validation loss (with 2 epochs patience) to mitigate overfitting. In other words, if the validation loss was not decreasing for 2 epochs then the training automatically stoped.

For the transformer models, I utilized the bert-base and distilbert-base tokenizers, respectively, configuring all input sequences to a fixed length of 60 tokens. Truncation and padding were applied as needed to ensure consistent length of 60 tokens. They were fine-tuned using the Adam optimizer with a learning rate of 2×10^{-5} and a batch size of 16. These hyperparameters were selected based on their frequent and successful application in downstream tasks, including those in the widely recognized GLUE benchmark suite. Each model was trained for a single epoch, aligning with findings from prior studies, which suggest that 1 to 4 epochs are typically sufficient for effective adaptation of pre-trained transformers to downstream tasks (Komatsuzaki, 2019).

Finally, all the models were trained on an NVIDIA RTX 4050 Laptop GPU. The training process for the Dual-Head BERT lasted approximately 4 hours and 18 minutes.

4 Results

Initially, I trained and tested three models, in order to find the best model for fake news detection, and then continued from there, in order to address the main question, namely whether we can accurately detect LLM-generated fake news. Tables 6 and 7 summarize the training and testing performance metrics, accordingly, for each model. Additionally, Table 8 offers a more granular view of test results by reporting precision, recall, and F1-score for both the Fake and Real news labels across all models.

Model	Loss	Accuracy (%)
RNN (BiLSTM)	0.49	76.7
DistilBERT	0.40	80.9
BERT-base	**0.31**	**86.7**

Table 6: Model training performance

Model	Loss	Accuracy (%)
RNN (BiLSTM)	0.76	59.7
DistilBERT	0.40	81.3
BERT-base	**0.28**	**88.5**

Table 7: Model testing performance

Model	Class	Precision	Recall	F1-Score
RNN (BiLSTM)	Fake	59	71	65
	Real	60	48	53
DistilBERT	Fake	78	65	71
	Real	72	83	77
BERT-base	Fake	83	88	86
	Real	88	83	85

Table 8: Classification Metrics on the Test Set for All Models

As shown in Table 7, BERT-base achieved the highest accuracy. In addition to that, it proved to have the lowest loss from all models (both training and testing), which indicates that it learns better from the data, as well as the only one with smaller testing loss (0.28) than training (0.31) which indicates that it did not overfit as much as the other models did. Also, according to Table 8 the results show that BERT-base outperforms both DistilBERT and the RNN (BiLSTM) model, achieving the highest scores across all metrics. While the RNN model shows moderate performance, particularly struggling with recall on the Real class, both transformer-based models demonstrate more balanced and robust results, with BERT-base showing the most consistent classification capability, as well as the highest performance compared to all other models.

Given its superior performance in fake news detection, I proceeded with BERT for the following analyses and I took a step further in order to address this study's main research question, namely if a model can accurately classify LLM-generated fake news. Recognizing the fundamental differences in the data generating processes among the datasets (i.e. whether the text is human-written or LLM-written), there is a need to check if the model can distinguish accurately between human-generated and LLM-generated content. For that reason, I extended the model architecture by adding a second classification head dedicated to distinguishing between human and LLM-generated content (see Figure 1 for the model architecture and chapter 3.2.2 for the description of the model's architecture). This enabled the model to perform dual tasks simultaneously: identifying whether a piece of text is fake or real (Fake-vs-Real), and determining its origin: human or machine (Human-vs-LLM). Accordingly, I included a second target variable in the training process, except for the fake news one, indicating whether a text was created by humans or LLM's. To that extent, all the tables that follow, present results from the dual-headed BERT-model, which was trained on two targets.

Table 9 shows the training and validation losses and accuracies for each target variable. The results suggest that the model does not suffer from overfitting, as validation accuracy slightly surpasses training accuracy, and the validation loss is marginally lower than the training loss.

Evaluation	Loss	Accuracy (%)
Training/Fake-vs-Real	0.22	90.5
Training/Human-vs-LLM	0.06	97.8
Validation/Fake-vs-Real	0.21	91.8
Validation/Human-vs-LLM	0.06	98.1

Table 9: Training and Validation Accuracy and Loss for Each Classification Task (Dual-Head BERT Model)

When evaluated on the test set, the model achieved an overall accuracy of 85% for the Fake-vs-Real task and 96% for the Human-vs-LLM task. Detailed performance metrics -including recall, precision and F1-score- for both targets are reported in Table 10.

Target Variables	Accuracy	Precision	Recall	F1-Score
Fake (%)	85	88	83	86
Real (%)		81	86	84
Human (%)	96	95	100	97
LLM (%)		99	74	85

Table 10: Performance Metrics for Fake-vs-Real and Human-vs-LLM Classification (Dual-Head BERT Model)

Table 10 presents the performance metrics of the dual-head BERT model for both classification tasks: Fake-vs-Real and Human-vs-LLM. For each target class, the table reports overall accuracy, along with precision, recall, and F1-score. The model achieves strong and balanced performance in the Fake-vs-Real task, with an F1-score of 86% for fake news and 84% for real news. In the Human-vs-LLM classification task, the model demonstrates high accuracy (96%), with near-perfect recall for human-generated text and strong precision for both classes. The lower recall for LLM-generated content indicates that a small portion of machine-generated samples were misclassified as human. In other words, the model captured 74% of all existing LLM-generated samples. However, the overall F1-score remains high at 85% and the precision as well (99%). This means that every time the model is presented with an LLM-generated text, it classifies at correctly as LLM-generated 99 out of 100 times. To that extent, from tables 9 and 10 we can conclude that the dual-headed BERT model that this study introduces, proved successful in accurately classifying fake news as well as identifying whether a text is human or LLM-generated, both with very high accuracy, and on the same time, being not prone to overfitting.

As I mentioned above, on chapter 3.1, the test set was created by randomly drawing 4000 samples from each data source, so as to test if the model can generalize equally among various sources. Given that, on tables 11 and 12 I present the model's accuracy for each task, broken down by data source.

Table 11 shows that the model performs particularly well on the Kaggle 1 and LLM News datasets, with accuracies of 94.8% and 94.0%, respectively. Performance is more modest on the LIAR 2 dataset (70.7%). Several factors likely contribute to this outcome. First, the original six-category labeling scheme was simplified into a binary classification, which, despite excluding the most ambiguous labels (such as "half-true" and "half-false"), inevitably resulted in a loss of nuanced information. Second, unlike datasets that rely on automated annotation, LIAR 2 was manually labeled by domain experts, introducing subtleties that are more difficult for a model to detect and generalize. Finally, the relatively short length of the LIAR 2 entries means the model has access to limited contextual data, making it more challenging to discern patterns or cues indicative of truthfulness. However, overall the model maintains good performance across most sources, indicating strong generalization for the Fake-vs-Real task.

Dataset	Accuracy (%)
Fakeddit	85.9
Kaggle 1 - Fake News	94.8
Kaggle 2 - News Project	78.5
Kaggle 3 - Fake News Detection	81.6
LIAR 2	70.7
LLM News Dataset	94.0
MegaFake	89.0

Table 11: Test Accuracy for Fake News Detection by Dataset (Dual-Head BERT Model)

In Table 12, the model demonstrates excellent performance on distinguishing human and LLM-generated content, achieving near-perfect accuracy (above 90%) across nearly all human-generated datasets and the "LLM News Dataset". However, performance is slightly lower on the MegaFake dataset (84.0%), which is expected due to the complex, diverse strategies used for generating its content (I elaborate more on that later). Still, the high overall accuracy suggests the model effectively distinguishes between human-

and machine-written text across various data sources.

Dataset	Accuracy (%)
Fakeddit	99.9
Kaggle 1 - Fake News	97.5
Kaggle 2 - News Project	94.1
Kaggle 3 - Fake News Detection	90.3
LIAR 2	97.6
LLM News Dataset	94
MegaFake	84.0

Table 12: Test Accuracy for Human-vs-LLM Detection by Dataset (Dual-Head BERT Model)

Overall, the slight performance gap between the accuracies for the two targets (85% for fake news detection and 96% for human-vs-LLM detection, as seen from table 10) may be explained by the inherent characteristics of LLM-generated text. Unlike human writing, which often includes inconsistencies and stylistic variation, text produced by language models tends to be more structured, and grammatically clean. This is a result of how LLMs generate language. They use sampling techniques like top-p (nucleus) sampling or adjusting the temperature parameter to produce outputs. These generation methods follow specific probabilistic patterns, which, while sophisticated, introduce detectable regularities in phrasing, coherence, and syntactic structure. As a result, the model is better able to identify LLM-generated content.

To evaluate the robustness and generalization capability of my model, I conducted an additional test using the MegaFake dataset, an LLM-generated dataset that -as said in chapter 3.1- had not been included in the training phase. As seen from Tables 11 and 12 the model achieved an impressive accuracy of 89% in fake news detection (and 84% on human-vs-LLM detection), indicating strong generalization to previously unseen LLM-generated content. This outcome reinforces the model's adaptability across a wide range of synthetic text sources.

The MegaFake dataset itself was constructed using four distinct generation strategies (Wang et al., 2024), each designed to simulate a different mode of misinformation:

- **Content-Based:** Alters factual elements such as dates, figures, and named entities within existing articles.

- **Integration-Based:** Combines real and fabricated information to produce hybrid narratives that appear credible.

- **Story-Based:** Generates entirely fictional news stories that are not grounded in real events.

- **Style-Based:** Imitates stylistic elements of legitimate journalism to make false articles more convincing.

Generation Technique	Fake-vs-Real Accuracy (%)	Human-vs-LLM Accuracy (%)
Content based	80.3	70.8
Integration based	87.5	81.9
Story based	98.6	97.6
Style based	90.4	86.0

Table 13: Model Accuracy for Fake-vs-Real and Human-vs-LLM Classification by Generation Technique (MegaFake Dataset Test Results)

Table 13 presents the model's performance on the MegaFake dataset, broken down by the four distinct text generation strategies used to create fake news. The results demonstrate that the model performs best on Story-Based fake news, achieving 98.6% accuracy in Fake-vs-Real classification and 97.6% in Human-vs-LLM detection. This is likely due to the clear fictional structure and semantic divergence from real news articles, making such content easier to identify. In contrast, Content-Based manipulation, where only small factual details are altered, proved more challenging for the model, resulting in the lowest accuracies for both tasks (80.3% and 70.8%, respectively).This outcome is expected, as content produced through this method closely mirrors the human-written material it was originally based on. Style-Based and Integration-Based generations fall between these extremes, suggesting that blending real and fake content or mimicking journalistic writing style adds moderate complexity to the detection tasks. Overall, the model shows strong performance across all techniques but also highlights the nuanced difficulty of detecting more subtly manipulated text, as it is evident from the results in content-based generation in the table.

Lastly, in table 14 I report the model's test accuracy, grouped by the specific LLM responsible for generating each set of fake news articles (in the "LLM News Dataset").

Model	Fake-vs-Real Accuracy (%)	Human-vs-LLM Accuracy (%)
GPT3.5	99.7	99.6
Llama2 13b	98.1	98.1
Llama2 7b	98.3	98.2
Mistral 7b	99.2	99.1

Table 14: Classification Accuracy by LLM Model for Fake-vs-Real and Human-vs-LLM Targets (Tested on "LLM News dataset")

The results show consistently high accuracy across all LLMs for both classification tasks. Notably, the model achieved near-perfect performance on text generated by GPT-3.5, with 99.7% accuracy for Fake-vs-Real and 99.6% for Human-vs-LLM. Similarly, Mistral 7B also yielded strong results, closely following GPT-3.5. Slightly lower, but still excellent, performance was observed for both versions of Llama2 (7B and 13B), with accuracies above 98% in all cases. These findings suggest that despite differences in model architecture and scale, the generated text retains detectable features that the classifier can reliably detect, further demonstrating the robustness and generalizability of the dual-head BERT model. A natural extension of this research would involve evaluating the model against newer generations of LLMs. Since the introduction of GPT-3.5, OpenAI has released several subsequent models, each bringing notable advancements in both scale and capability. Testing the model's robustness against these more sophisticated models would provide valuable insights into its adaptability and future-proofing.

In summary, a range of deep learning architectures was implemented and evaluated, including BiLSTM, DistilBERT, and BERT-base. Among them, BERT-base achieved the highest accuracy and lowest loss, showing superior performance across all key evaluation metrics, and thus meeting the objective of identifying the most effective model for binary fake news classification. After selecting BERT-base as the best-performing model, the architecture was extended to include a second classification head, enabling it to simultaneously detect whether a piece of content was fake or real and whether it was written by a human or generated by a language model. The model achieved 85% accuracy in Fake-vs-Real detection and 96% accuracy in Human-vs-LLM classification, demonstrating strong capability in both tasks. Importantly, it also showed resilience across different data sources, with high performance sustained even on previously unseen datasets like MegaFake. In addition, the model performed exceptionally well on identifying LLM-generated fake news generated by different language models and different generation techniques. Also, while the model performed best on story-based fake news, it showed slightly lower performance on more subtly manipulated content such as content-based alterations, suggesting an area for potential future research. Overall, by testing the model across various generation techniques and LLM architectures, the study confirmed that while LLMs can produce convincing fake news, their outputs still exhibit detectable patterns that allow for reliable classification. These results collectively affirm the effectiveness of the proposed system in both identifying misinformation and distinguishing between human and AI-generated content.

5 Conclusion

This study set out to investigate a highly relevant and timely research question: Can a classification model accurately detect LLM-generated fake news? To answer this, the research was structured around four key objectives. These included: (1) constructing a robust, high-quality dataset sourced from a diverse range of sources, (2) designing and evaluating multiple deep learning architectures for fake news detection, (3) extending the best-performing model to simultaneously classify whether text was human or LLM-generated, and (4) testing the model's resilience against fake news generated by different language models as well as different prompting strategies. Each of these objectives was systematically addressed, and the results provide a clear and affirmative response to the central research question.

The first objective was centered on building a comprehensive dataset to support robust and generalizable fake news detection. This study made a substantial contribution by integrating data from six human-written news datasets and one LLM-generated news dataset, resulting in a large corpus with both "Fake-vs-Real" and "Human-vs-LLM" labels. Crucially, the study went beyond the typical practice of relying on a single data source by incorporating diverse formats (e.g., Reddit posts, full-length articles, political statements), linguistic styles, and topic domains. The results in Tables 2 and 1 demonstrate that several popular pre-trained fake news models, when evaluated on this combined dataset, performed significantly worse than on their original datasets, highlighting the issue of low generalizability that this study aimed to address. These findings validate the underlying hypothesis that many existing models are overfitted to narrow datasets, and they underscore the importance of training on a heterogeneous corpus. By overcoming this limitation, the present study ensured that its results reflect realistic deployment conditions and broader applicability.

The second objective involved experimenting with a variety of deep learning model architectures. Three major models were developed and assessed: a bidirectional Long Short-Term Memory (BiLSTM) network, a DistilBERT fine-tuned transformer model, and a fine-tuned BERT-base model. The comparative analysis (Tables 6–8) clearly demonstrated the superiority of transformer-based approaches, particularly BERT-base, which consistently outperformed the other models in terms of accuracy, loss, precision, recall, and F1-score. These results reaffirm previous findings in the literature about the effectiveness of transformer architectures for text classification tasks. In addition, the dual evaluation across training and testing phases revealed that BERT-base was not prone to overfitting, further reinforcing its robustness. Thus, the study fulfilled its second objective by identifying and selecting the most effective model for fake news detection based on comprehensive evaluation metrics.

After selecting BERT-base as the best-performing model, the third objective was addressed by extending it into a dual-headed classification model. This adaptation allowed the model to simultaneously classify news content along two dimensions: (1) whether it was real or fake, and (2) whether it was written by a human or generated by an LLM. The performance of the dual-head model was exceptional across both tasks, achieving 85% accuracy in Fake-vs-Real detection and 96% accuracy in Human-vs-LLM classification (Table 10). Importantly, the model demonstrated high F1 scores on both tasks, and it was particularly successful in identifying LLM-generated text, with a precision of 99%. These results indicate that the dual-task architecture did not compromise the model's performance, in fact, it enhanced its utility by addressing two critical dimensions of misinformation simultaneously. This directly supports the research question by showing that LLM-generated fake news is not only detectable but also distinguishable from both real news and human-written fake news.

The fourth and final objective examined the model's ability to generalize to LLM-generated misinformation produced by different architectures on one hand, and different prompting strategies on the other. The first was evaluated using the "LLM News Dataset" (test set)- which was created using various LLMs, including GPT-3.5, Llama2-7B, Llama2-13B, and Mistral-7Band- and the second on the the completely unseen "MegaFake" dataset, which was created using multiple fake news generation strategies. The model achieved nearly perfect classification results on the "LLM News Dataset," with Fake-vs-Real and Human-vs-LLM accuracies exceeding 98% across all tested LLMs (Table 14). It also maintained strong generalization on the MegaFake dataset, despite not being exposed to it during training, achieving 89% accuracy for Fake-vs-Real and 84% for Human-vs-LLM classification. This finding is particularly notable given the dataset's complexity, which involved multiple misinformation strategies including content-based manipulation, integration of true and false information, style-based mimicry, and fictional story generation. While detection performance was highest on story-based content (likely due to its semantic distance from real news), content-based manipulations were more difficult to detect, illustrating that subtle factual distortions remain the most challenging for current detection models. These results reinforce the robustness of the model and demonstrate its capacity to identify fake content across various LLM architectures and prompting strategies, as well as its generalizability to unseen data.

Beyond these key contributions, the model's performance also points to broader implications for the future of misinformation detection. The fact that the model was able to detect machine-generated misinformation across multiple generation techniques and model architectures demonstrates that LLMs, while capable of producing highly convincing content, still exhibit systematic linguistic patterns that can be algorithmically identified. These patterns may arise from probabilistic sampling methods used during generation (such as nucleus sampling or temperature scaling), which, while designed to promote fluency, also result in detectable regularities in syntax, phrasing, or cohesion. Understanding and leveraging these cues could offer a meaningful advantage in future detection systems.

6 Discussion

The current misinformation literature faces two main gaps: (1) most fake news detectors are trained on narrow, single-source datasets like LIAR or Fakeddit, which limits their ability to generalize to real-world data, and (2) the rise of highly convincing fake news generated by Large Language Models (LLMs) presents a new challenge, as existing detection methods often fail to reliably identify or generalize across different LLM outputs and generation strategies.

This thesis addressed these gaps in two key ways. First, it introduced a large, diverse dataset that combines six different sources, including both human-written and LLM-generated content, to improve generalizability and avoid overfitting to a specific format or domain. Second, it proposed a dual-head BERT-based model that simultaneously classifies whether a news article is fake or real and whether it was written by a human or an LLM. This joint-learning approach not only improves overall detection accuracy but also tackles the unique challenges posed by AI-generated misinformation, demonstrating strong performance even on unseen content and across different LLM architectures and generation strategies.

Nonetheless, the study also faced a number of limitations. Due to constraints in GPU memory and computational time, the study employed fixed-length input sequences capped at 60 tokens. Future work could explore alternative segmentation strategies, such as dynamic chunking based on sentence boundaries, paragraph length, or semantic coherence. Moreover, interpretability remains a significant challenge. While the model achieved high accuracy, it operates as a black box, offering little insight into which features it relies on to make decisions. Incorporating interpretability tools, such as attention heatmaps, SHAP values, or gradient-based saliency maps, could provide valuable transparency and help identify the specific linguistic or structural patterns that distinguish fake from real or human from machine-generated content.

Another limitation is that the classification approach in this study was binary for both tasks. While this provided a practical framework and allowed the model to perform strongly across datasets, future studies could expand this to include more fine-grained classification schemes. For instance, instead of a simple Fake-vs-Real distinction, models could be trained to identify varying degrees of misinformation (e.g., satire, partially true, manipulated headlines), as is done in the original six-category LIAR dataset. Similarly, the Human-vs-LLM classification could be extended to identify which specific LLM was used to generate a piece of content.

In a nutshell, this study provided evidence that can affirmatively answer the question of whether we can accurately detect LLM-generated fake news. Although there were some limitations, this study's contributions aim to help address the growing challenges posed by increasingly sophisticated text generation technologies, and lay the ground for future work.

References

Abaghyan, G. (2025). Fake News Detection: NLP Machine Learning Dataset. `https://www.kaggle.com/datasets/abaghyangor/fake-news-dataset/data`.

Ayoobi, N., Shahriar, S., and Mukherjee, A. (2024). Seeing through ai's lens: Enhancing human skepticism towards llm-generated fake news. In *Proceedings of the 35th ACM Conference on Hypertext and Social Media*, pages 1–11.

Chen, C. and Shu, K. (2024). Can llm-generated misinformation be detected?

Denniss, E. and Lindberg, R. (2025). Social media and the spread of misinformation: infectious and a threat to public health. *Health Promotion International*, 40(2):daaf023.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gehrmann, S., Strobelt, H., and Rush, A. (2019). GLTR: Statistical Detection and Visualization of Generated Text. In Costa-jussà, M. R. and Alfonseca, E., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116, Florence, Italy. Association for Computational Linguistics.

Hu, B., Mao, Z., and Zhang, Y. (2025a). An overview of fake news detection: From a new perspective. *Fundamental Research*, 5(1):332–346.

Hu, B., Sheng, Q., Cao, J., Li, Y., and Wang, D. (2025b). Llm-generated fake news induces truth decay in news ecosystem: A case study on neural news recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25)*. ACM.

Indurthi, V., Syed, B., Gupta, M., and Varma, V. (2020). Predicting clickbait strength in online social media. In *Proceedings of the 28th International Conference on Com-*

putational Linguistics, pages 4835–4846. International Committee on Computational Linguistics.

Jruvika (2018). Fake News Detection. `https://www.kaggle.com/datasets/jruvika/fake-news-detection/data`.

Kaliyar, R. K., Goswami, A., and Narang, P. (2021). Fakebert: Fake news detection in social media with a bert-based deep learning approach. *Multimedia Tools and Applications*, 80(8):11765–11788.

Kanchana, M., Kumar, V. M., Anish, T. P., and Gopirajan, P. (2023). Deep fake bert: Efficient online fake news detection system. *ResearchGate*.

Kokiantonis, A. (2022). Fake News Project. `https://www.kaggle.com/datasets/antonioskokiantonis/newscsv/data`.

Komatsuzaki, A. (2019). One epoch is all you need. *arXiv preprint arXiv:1906.06669*.

Li, J., Bin, Y., Zou, J., Zou, J., Wang, G., and Yang, Y. (2023). Cross-modal Consistency Learning with Fine-grained Fusion Network for Multimodal Fake News Detection. arXiv:2311.01807 [cs].

Monti, F., Frasca, F., Eynard, D., Mannion, D., and Bronstein, M. M. (2019). Fake news detection on social media using geometric deep learning. *arXiv preprint arXiv:1902.06673*.

Nakamura, K., Levy, S., and Wang, W. Y. (2019). r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. *CoRR*, abs/1911.03854.

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

Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., and Stein, B. (2018). A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 231–240, Melbourne, Australia. Association for Computational Linguistics.

Rubin, V., Conroy, N., Chen, Y., and Cornwell, S. (2016). Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, pages 7–17, San Diego, California. Association for Computational Linguistics.

Ruchansky, N., Seo, S., and Liu, Y. (2017). 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. ACM.

Sallami, D., Chang, Y.-C., and Aïmeur, E. (2024). From Deception to Detection: The Dual Roles of Large Language Models in Fake News. arXiv:2409.17416 [cs].

Su, J., Zhuo, T. Y., Mansurov, J., Wang, D., and Nakov, P. (2023). Fake news detectors are biased against texts generated by large language models. *arXiv preprint arXiv:2309.08674*.

Tacchini, E., Ballarin, G., Della Vedova, M. L., Moret, S., and de Alfaro, L. (2017). Some like it hoax: Automated fake news detection in social networks. *arXiv preprint arXiv:1704.07506*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, u., and Polosukhin, I. (2017). Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wang, L. Z., Ng, K. C., Ma, Y., and Fan, W. (2024). Megafake: A theory-driven dataset of fake news generated by large language models. Available at SSRN: `https://ssrn.com/abstract=5095309` or `http://dx.doi.org/10.2139/ssrn.5095309`.

Wang, W. Y. (2017). “liar, liar pants on fire”: A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Wu, J., Yang, S., Zhan, R., Yuan, Y., Wong, D. F., and Chao, L. S. (2024). A survey on llm-generated text detection: Necessity, methods, and future directions.

Xu, C. and Kechadi, T. (2024). An enhanced fake news detection system with fuzzy deep learning. *IEEE Access*, 12:88006 – 88021.

Yang, Y., Zheng, L., Zhang, J., Cui, Q., Li, Z., and Yu, P. S. (2018). Ti-cnn: Convolutional neural networks for fake news detection. *arXiv preprint arXiv:1806.00749*.

Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., and Choi, Y. (2019). Defending against neural fake news. In *Advances in Neural Information Processing Systems 32*, pages 9054–9065. Curran Associates, Inc.

Zhou, X. and Zafarani, R. (2018). A survey of fake news: Fundamental theories, detection methods, and opportunities. *arXiv preprint arXiv:1812.00315*.

APPENDIX

This is the link to the GitHub repository of this study, where all the necessary replication code can be found.