Fake News Detection with Retrieval Augmented Generative Artificial Intelligence

Anonymous Author(s)

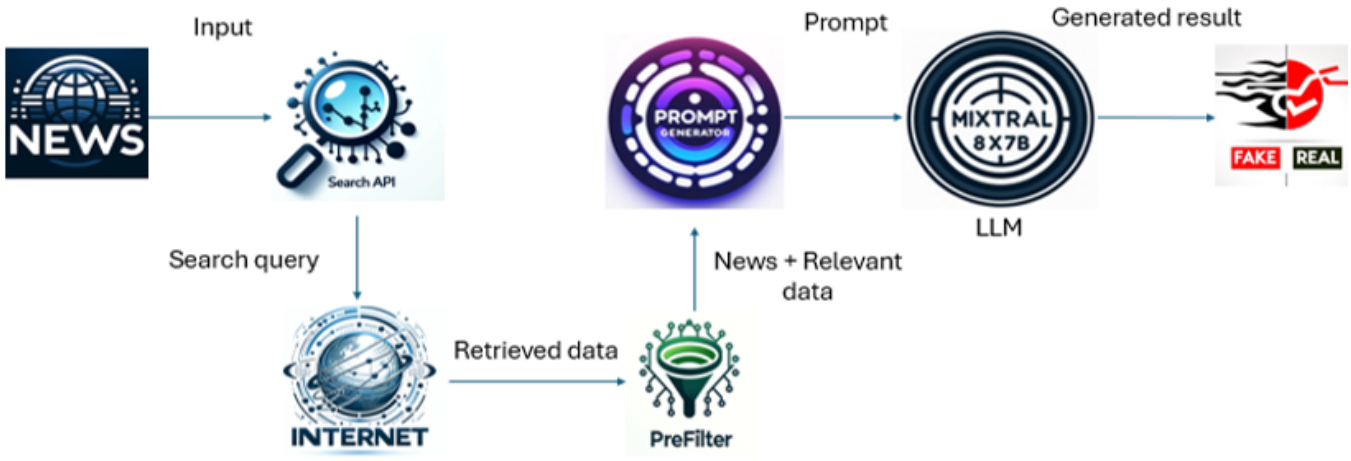

Figure 1: Fake News Detector Framework

ABSTRACT

The rapid spread of false information on social media has grown to be a serious problem that influences public opinion and decision-making. Fake news spreads rapidly and extensively, often outpacing efforts to debunk or mitigate its effects. Traditional methods for detecting fake news face numerous challenges, including the necessity for extensive model training and the potential for inherent biases. Although Large Language Models (LLMs) have seen substantial improvements recently, their use in fake news detection poses the risk of producing false or misleading information due to their possible hallucinations. This study presents a new strategy to combat fake news by integrating Mixtral-8x7B, a Sparse Mixture of Experts (SMoE) Large Language Model, with a Retrieval-Augmented Generation (RAG) framework. Our framework employs Google's search API to retrieve relevant articles in real time, harnessing Mixtral's sophisticated language processing capabilities and RAG's ability to access current information dynamically. Initial results are promising, indicating that our approach performs comparably to established fake news detection techniques. Our method operates without the need for extensive model training, offering significant cost savings and contributing to developing more efficient tools for detecting misinformation in the digital era, which will help stop the spread of misleading data more efficiently.

Workshop'4, August 2024, Barcelona, Spain

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

CCS CONCEPTS

• **Information systems** → *Retrieval effectiveness*; *Language models*; Similarity measures; • **Computer systems organization** → *Real-time operating systems.*

KEYWORDS

Fake News Detection, Sparse Mixture of Experts, Retrieval-Augmented Generation, Large Language Model, Google Search API

ACM Reference Format:
Anonymous Author(s). 2024. Fake News Detection with Retrieval Augmented Generative Artificial Intelligence. In . ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

False information poses a significant challenge in our digitally connected world. It spreads rapidly, impacting millions daily through enticing headlines and misleading content [36]. Hence, identifying fake news emerges as a critical issue garnering considerable research attention. Detecting fake news on social media consistently presents a fresh challenge, as it's often crafted to deceive readers. During the 2016 US presidential election, fake news proliferated more extensively on Facebook than genuine news [29]. False information, also known as fake news or misinformation, is longstanding in societies [9].

Over the past few years, social networks have become fertile breeding grounds for disseminating fake news, leading to widespread confusion and misinterpretation of critical social and political issues, particularly among individuals with limited access to reliable information. For instance, Wineburg et al. [35] indicate that numerous high school students struggle to discern fake news outlets on platforms like Facebook.

Furthermore, unlike some traditional sources of suspicion [11], fake news is difficult to recognize and carries significant societal ramifications, exemplified by incidents such as the Pizzagate shooting in 2016 [20]. Consequently, substantial efforts have been directed toward detecting fake news [26] and mitigating its adverse effects on society.

2 PROBLEMS AND MOTIVATIONS

Detecting fake news using machine learning has been approached with various algorithms, each presenting unique strengths and limitations. Previous studies have utilized methods such as Naive Bayes, passive-aggressive classifiers, and Deep Neural Networks, demonstrating the potential for automating fake news detection. However, these models often suffer from biases due to inconsistent dataset quality. Furthermore, as explored by Altheneyan et al., big data technologies and ensemble models have shown promise but require substantial computational resources, limiting their practical deployment [24][5].

Notable advancements have been witnessed in LLMs [12][13] and pre-training methodologies [28][22]. The exceptional capabilities demonstrated by LLMs such as GPT-4 [1][25] underscore their superior performance in various tasks.

The explosive growth of fake news and its erosion of democracy, justice, and public trust have heightened the demand for effective fake news detection and intervention strategies. Researchers are increasingly focusing on developing innovative methods to combat the spread of misinformation [4].

Recent advancements in Large Language Models (LLMs) and the rise of fake news have motivated us to explore their potential for improving detection methods. Studies, such as those by Hu et al. [14], have shown that while LLMs like GPT-3.5 and ChatGPT provide valuable multi-perspective rationales, they often struggle with consistency and specificity in tasks like fake news detection compared to fine-tuned smaller models. When targeted and fine-tuned, these smaller models can outperform LLMs due to their efficiency and specificity in handling niche tasks.

Furthermore, LLMs like GPT-4 and Llama, despite their ability to generate coherent and contextually relevant text, are prone to "hallucination," where they produce information that is not factually accurate. This occurs because LLMs predict text based on statistical likelihoods from their training data, leading to plausible yet potentially misleading or fabricated outputs, especially in complex or niche topics [32].

Retrieval Augmented Generation (RAG) frameworks have emerged as a solution to address these issues [23]. By incorporating external data during the generation process, RAG frameworks help LLMs produce more accurate and reliable outputs. Research indicates that RAG enhances the quality of structured outputs and improves the generalization of LLMs across various domains, significantly reducing the occurrence of misleading information [16].

By integrating real-time data retrieval, RAG frameworks ensure that the information used by LLMs is up-to-date and contextually appropriate, thereby enhancing the credibility and accuracy of the generated content [18]. This understanding has guided our approach to leveraging LLMs and real-time data retrieval to improve the generalizability and effectiveness of fake news detection.

The financial and computational costs of model training are substantial and well-documented [33]. When there is a shift in the domain, it necessitates retraining the model from scratch or employing transfer learning. While transfer learning can reduce the need for large volumes of new data, it still requires significant training data and may be ineffective when the domain changes drastically.

However, the advent of LLMs has revolutionized this landscape. These models are pre-trained on vast amounts of data across diverse topics, allowing them to generalize well across different domains. Although the initial training of these models is resource-intensive, once trained, they exhibit versatility across various applications without the need for extensive retraining. This ability to generalize significantly reduces the overhead associated with domain adaptation.

The emergence of open-source LLMs has further lowered the barrier to entry for creating new AI applications. Researchers can leverage these pre-trained models and build upon them without incurring the prohibitive costs of data collection and computational resources for training [30]. This accessibility of advanced AI technologies fosters innovation and accelerates the development of specialized applications across diverse fields.

While significant costs and logistical challenges burden traditional model training approaches[9], the advent of LLMs offers a promising alternative. These models, supported by open-source initiatives, enable researchers to overcome traditional barriers, streamline workflows, reduce costs, and accelerate innovation across various domains.

3 RELATED WORKS

Detecting fake news using machine learning has seen various approaches with significant achievements and notable limitations. Mandical et al. [24] applied algorithms like Naive Bayes, Passive Aggressive Classifier, and Deep Neural Networks to multiple datasets, highlighting the feasibility of automating the classification process. These algorithms each have unique strengths; for instance, Naive Bayes is known for its simplicity and efficiency, Passive Aggressive Classifier excels in handling large-scale data streams, and Deep Neural Networks can capture complex patterns in data. However, their study did not address the biases that might arise from varied dataset quality, potentially limiting the generalizability of the models. Inconsistent data quality can lead to skewed results, as models trained on high-quality datasets may perform poorly when exposed to real-world data with noise and inconsistencies.

The integration of big data technologies in fake news detection was examined by Altheneyan et al. [5], where a stacked ensemble model employed on a decentralized Spark cluster showed improved

performance metrics. This approach leverages the power of ensemble learning, where multiple models are combined to improve prediction accuracy and robustness.

By distributing the computational load across a Spark cluster, they could efficiently process vast amounts of data. However, the necessity of substantial computational resources could restrict the deployment of such systems in less resource-intensive settings. The high cost of maintaining and operating a Spark cluster and the need for technical expertise make it less feasible for small organizations or individual researchers with limited budgets.

In a more focused study on large language models, Hu et al. [14] investigated the effectiveness of LLMs like GPT-3.5 in detecting fake news. The results suggested that while LLMs provide valuable multi-perspective rationales, they do not perform well in isolation compared to fine-tuned smaller models. This is because LLMs, although powerful in generating coherent and contextually relevant text, may lack the specificity and tuning required for niche tasks like fake news detection unless they are specifically fine-tuned for such purposes. Smaller models, when fine-tuned, can outperform LLMs by being more targeted and efficient in their scope.

Also, Huang et al. [15] assessed ChatGPT's capabilities in generating, explaining, and detecting fake news, revealing issues with consistency and a need for innovative methods to enhance performance in ambiguous content situations. Their study highlighted that while ChatGPT can generate plausible explanations, it sometimes struggles with consistency, particularly in scenarios where the context is vague or contradictory.

LLMs like GPT-4 and Llama have revolutionized the field of natural language processing with their ability to generate coherent and contextually relevant text. However, these models are also prone to a phenomenon known as "hallucination," where they generate information that is not factually accurate or even entirely fictional [32]. This occurs because LLMs, proficient in language patterns and associations, do not inherently understand the truth or verify facts; they instead predict text based on statistical likelihoods derived from their training data. This can lead to outputs that, while plausible, may be completely made up or misleading, particularly when handling complex or niche topics where verifiable data may have been sparse in their training sets. For example, in generating news articles, an LLM might fabricate details that sound credible but are entirely unfounded, leading to the dissemination of misinformation.

Following this understanding, RAG frameworks have proven to be instrumental in mitigating the issue of hallucinations in LLMs. By dynamically incorporating external data during the generation process, RAG helps LLMs produce more accurate and reliable outputs. This technique leverages the depth and breadth of external knowledge bases, ensuring the generated content is relevant and verifiable.

Studies indicate that RAG enhances the quality of structured outputs and improves the overall generalization of LLMs across various domains, significantly reducing the occurrence of misleading or fabricated information generated by these models. By integrating real-time data retrieval, RAG frameworks ensure that the information used by LLMs is up-to-date and contextually appropriate, thereby enhancing the credibility and accuracy of the generated content [23][21].

LLMs are available in various forms, some being open-source and others proprietary. Open-source LLMs, such as the Mixtral-8x7B, are freely accessible and can be modified to suit specific research needs. This contrasts with proprietary models, often with usage restrictions and licensing fees.

4 METHODOLOGY

By utilizing Google's Search API for real-time retrieval of news articles, our model accesses the most current and relevant information for analysis. The RAG framework reduces hallucinations commonly associated with large language models, improving the reliability of the results. We developed and applied optimized prompt engineering techniques to guide the LLM in producing more accurate and contextually relevant outputs then We conducted experiments to fine-tune parameters such as top-k sampling, top-p sampling, temperature, max tokens, stop tokens, and temperature decay to achieve optimal performance of the Mixtral-8x7B model, further enhancing the overall system performance. These contributions collectively ensure a robust, trustworthy classification system and represent an advancement in combating fake news with cutting-edge technology.

4.1 Real time article retrieval

We used the Google Search API provided by VALUESERP to gather relevant information. This API allows us to retrieve the top 10 search results related to the title of each news article in our dataset. The API returns results in JSON format, with each JSON file containing several fields, including the web page's title, the domain from which the content originates, and a snippet that provides a summary or excerpt of the page content.

4.1.1 Prefiltering. Our system includes an important step for filtering out noise and irrelevant data before analysis. We use DistilBERT, a pre-trained transformer model, to turn the titles of the news articles in our dataset and the titles of the Google Search results into embeddings. These embeddings are dense vectors that capture the text's meaning, representing the context and meaning of words and phrases. With DistilBERT, we get the advantages of efficiency and smaller size while keeping high accuracy and performance.

After creating the embeddings, we compute the cosine similarity [31] between them to assess how relevant the search results are to the original news article. Cosine similarity measures the cosine of the angle between two vectors, giving us a metric for their similarity. High cosine similarity means the search result is closely related to the news title, while low cosine similarity indicates less relevance.

The cosine similarity between two vectors $\mathbf{A}$ and $\mathbf{B}$ is given by the formula:

$$\text{cosine similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} \tag{1}$$

We use a dynamic threshold for cosine similarity to ensure we gather the most relevant search results. We start with a threshold of 0.85. If we don't find relevant news at this threshold, we lower it to 0.80, and finally to 0.70. This approach ensures we collect the most relevant search results possible. If no relevant data with a cosine similarity above 0.70 is found, we pass all the search results to the LLM for final determination. This additional step improves the accuracy and reliability of our fake news detection system by ensuring a thorough evaluation of the news articles.

In this step, we also add a credibility annotation to each retrieved result. We have a list of credible news sources, and after calling the search API, it returns a list of JSON files, each containing information about the results. We check the domain of each result and add a source credibility field, marking it as either true or false based on our list. This annotation helps the LLM classify the authencity of news.

4.2 Prompt Engineering

We conducted experiments to compare the results of zero-shot [27] and few-shot prompting [7]. In zero-shot prompting, the model is given a task without any prior examples, relying solely on its pre-trained knowledge to generate responses. In contrast, few-shot prompting involves providing the model with a few examples related to the task, enabling it to learn from these examples and improve its performance. Our findings indicated that few-shot prompting significantly enhanced the quality of the answers generated by the LLM.

The improvement with few-shot prompting can be attributed to the model's ability to recognize patterns and apply the learned examples to new, unseen prompts. This method reduces ambiguity and provides a clearer framework within which the LLM can operate, resulting in more accurate and reliable outputs. By leveraging the few-shot prompting technique, we ensure the LLM is better equipped to handle the nuances and complexities of fake news detection.

We also employed chain of thought prompting [34]. This method helps divide the LLM task into manageable stages rather than providing an immediate answer. By following this structured process, the steps outlined in the instructions are executed sequentially, ensuring a thorough and organized completion of the task by the LLM.

These techniques collectively enabled more effective prompt engineering, ensuring the LLM produced high-quality, contextually appropriate outputs that met my objectives.

4.3 LLM

After preparing the data through the above steps, we use the LLM to classify the news article as fake or real. The LLM analyzes the filtered search results, credibility annotations, and structured prompts to make an informed decision.

We utilize the Mixtral 8x7B model, a Sparse Mixture of Experts (SMoE) language model, which enhances efficiency and performance by using multiple experts within its layers. Mixtral 8x7B, a Sparse Mixture of Experts (SMoE) model with open weights licensed under Apache 2.0. Mixtral 8x7B outperforms both Llama 2 70B and GPT-3.5 on most benchmarks. The model's architecture allows it to use only a subset of its parameters for each token. It enables faster inference speeds at low batch sizes and higher throughput at large batch sizes. Mixtral is a decoder-only model where each feed-forward block selects from 8 distinct parameter groups (experts). At every layer, a router network chooses two experts to process each token, combining their outputs additively. This approach increases model parameters while controlling computational cost and latency. Pretrained with multilingual data using a 32k token context size, Mixtral matches or exceeds the performance of Llama 2 70B and GPT-3.5 across various benchmarks, excelling in mathematics, code generation, and multilingual tasks. It effectively retrieves information from its extensive context window, irrespective of sequence length or information position. Additionally, Mixtral 8x7B – Instruct, a fine-tuned chat model, surpasses GPT-3.5 Turbo, Claude-2.1, Gemini Pro, and Llama 2 70B in human evaluations, demonstrating reduced biases and a balanced sentiment profile. Instructions are available under the Apache 2.0 license, allowing for free academic and commercial use [17].

5 EXPERIMENTS

5.1 Experimental Setup and Configuration

All experiments were conducted using Amazon Web Services (AWS) cloud services. We utilized an AWS SageMaker notebook instance configured as ml.t3.2xlarge, offer- ing 8 vCPUs and 32 GB of memory to meet our computational needs. This setup provided a managed environment for developing, training, and deploying our mod- els. To deploy the MixTral LLM, we used AWS Bedrock, which supports efficient, scalable model deployment. Additionally, we accessed the Google Search API via VALUESERP to gather relevant news articles. This setup enabled seamless integra- tion, robust execution, and reliable experimentation, ensuring our system had the necessary computational power and access to up-to-date data.

5.2 ISOT Fake News Data Set

The dataset we used features two categories of articles: fake and real news, sourced from the real world. Authentic articles were sourced from Reuters.com through web crawling, while fake news was compiled from various unreliable sites identified by PolitiFact and Wikipedia, mainly focusing on political and global news topics. Each article includes details like the title, text, type, and publication date. The dataset has been cleaned and processed, punctuation and errors in the fake news articles were preserved [3][2].

Table 1: Dataset Distribution Summary

Label	#Article	Details
Real	21,417	World News: 10,147 Politics news: 11,272 Gov. news: 1,570 Mid. east: 778
Fake	23,481	Left news: 4,459 Polities: 6,841 News: 9,050

5.3 Performance metrics

This section defines the key performance metrics used to evaluate our fake news detection model, such as precision, recall, F1 score, and accuracy. These metrics provide a comprehensive view of the model's effectiveness. Accuracy measures the proportion of true results (both true positives and true negatives) among the total number of cases examined. Precision indicates the proportion of true positives among the total predicted positives, showing the accuracy of the positive predictions. Recall represents the proportion of true positives that were correctly identified by the model, highlighting its ability to capture all relevant instances. The F1 score combines precision and recall into a single metric by taking their harmonic mean, providing a balanced measure of the model's performance.

5.4 Semantic Similarity Check

For the pre-filtering part, we explored various semantic similarity measures alongside cosine similarity to check the relevance of retrieved articles. One alternative similarity measure we tested was the Euclidean distance between the embeddings of the titles. Another measure we evaluated was the Manhattan distance, which also aims to quantify the semantic closeness between the retrieved articles and our dataset. To determine the most effective similarity check approach, we ran an experiment on a sample of 200 data points from our dataset. These methods, like cosine similarity, ensure the relevance of the filtered data.

Table 2: Result of different similarity check techniques

	Threshhold	Precision	Recall	Accuracy
Manhattan	100	0.88	0.68	0.77
Euclidean	4	0.86	0.70	0.78
Cosine	0.85	0.93	0.82	0.88

Cosine similarity, Euclidean distance, and Manhattan distance are techniques we used to measure the semantic closeness between vectors representing text embeddings. Cosine similarity measures the cosine of the angle between two vectors, focusing on their orientation rather than magnitude, making it effective for comparing the semantic similarity of texts [31][10].

Euclidean distance, on the other hand, calculates the straight-line distance between two points in multi-dimensional space, quantifying how far apart two vectors are in terms of absolute distance. Manhattan distance, also known as L1 distance, measures the sum of the absolute differences between the coordinates of two points, assessing the distance one would travel along the axes to move from one point to another.

In the context of our experiment, cosine similarity outperformed both Euclidean and Manhattan distances, demonstrating higher precision, recall, and accuracy in ensuring the relevance of retrieved articles during the pre-filtering process according to Table 2. This indicates that cosine similarity is the most suitable method for our similarity check tasks.

5.5 Hyperparameter Tuning

We used the same 500 samples for hyperparameter tuning for the next step. When calling the Mixtral-8x7B Sparse Mixture of Experts LLM, you typically specify parameters to guide the inference process. These include top-k sampling (top k) to limit the sampling pool to the top-k predictions and top-p (nucleus) sampling (top p) to consider only the most likely predictions. The temperature parameter controls the randomness of forecasts, with lower values making the model more deterministic. The max tokens parameter sets the maximum number of tokens to generate, while stop tokens are used to specify when the model should stop generating further text.

Additionally, temperature can be used to control the decay of the temperature parameter over time, stabilizing the generation process.We eliminated these samples from the data to avoid any bias toward these samples. The following table summarizes the results, including accuracy, precision, recall, and F1 score for five different sets of parameters:

Table 3: LLM Parameter Tuning Results

	Set 1	Set 2	Set 3	Set 4	Set 5
Top-k	100	50	30	20	60
Top-p	0.2	0.9	0.7	0.8	0.85
Temperature	0.0	0.7	0.3	0.5	0.6
Accuracy	**0.91**	**0.88**	**0.90**	**0.89**	**0.90**

After conducting experiments with various parameter sets, we chose the first set because it provided the most accurate and reliable results for our fake news detection system. As shown in Table 3, the first set uses a low temperature (0.0), a conservative top-p value (0.2), and a higher top-k value (100). These parameters help ensure the model generates highly deterministic and precise outputs by focusing on the most probable tokens.

5.6 Without RAG

In this experiment, we only used the LLM to detect fake news without any extra information about the data we had.

Table 4: Classification Results without RAG

	Precision	Recall	F1-Score	Support
Class 0	0.58	1	0.73	21167
Class 1	1	0.33	0.49	23231
Accuracy	0.65			
Macro Avg	0.79	0.66	0.61	44398
Weighted Avg	0.80	0.65	0.61	44398

Table 5: Confusion Matrix Without RAg

	Predicted Positive	Predicted Negative
Actual Positive	21141	26
Actual Negative	15595	7636

Table 4 shows the classification results for the experiment using the RAG framework, including precision, recall, F1-score, and support for each class. Table 5 illustrates the model's performance by displaying the Confusion Matrix for the experiment using the RAG framework.

5.7 With RAG

This experiment integrated the RAG framework with our LLM, MixTral 8x7B Sparse Mixture of Experts, to detect fake news. The RAG framework enhanced the LLM's performance by providing relevant articles retrieved in real-time using Google's Search API.

Table 6: Classification Results with RAG

	Precision	Recall	F1-Score	Support
Class 0	0.82	0.95	0.88	21167
Class 1	0.94	0.81	0.87	23231
Accuracy	0.88			
Macro Avg	0.88	0.88	0.88	44398
Weighted Avg	0.88	0.88	0.88	44398

Table 7: Confusion Matrix With RAG

	Predicted Positive	Predicted Negative
Actual Positive	20019	1148
Actual Negative	4346	18885

Table 6 shows the classification results for the experiment using the RAG framework, including precision, recall, F1-score, and support for each class. Table 7 illustrates the model's performance by displaying the Confusion Matrix for the experiment using the RAG framework.

Here are successful examples of the LLM's answers detecting fake news:
Table 8: First Example of LLM's Response

Title of the news	**HANNITY AND TUCKER CARLSON: THE 5 MAJOR FORCES Aligned To Take Down President Trump**
With RAG	**Fake News Determination: True** Reason: While there is a credible source (foxnews.com) that supports the claim, the majority of the sources are not credible and contradict the claim. Additionally, the snippet from the credible source does not directly support the claim."
Without RAG	**Fake News Determination: False** Reason: This news is from an opinion-based news network, Fox News. The article is expressing the opinions of the two hosts, Sean Hannity and Tucker Carlson. It is not fake news.

Table 9: Second Example of LLM's Response

Title of the news	**HOW SENATE DEMOCRATS PLAN TO FORCE GUN-CONTROL Amendment On GOP's Bill To Repeal Obamacare**
With RAG	**Fake News Determination: True** Reason: The search results do not support the claim made in the news title and description. The articles are about the Senate passing a bill to repeal Obamacare, and do not mention any plans by Senate Democrats to force gun-control amendments on the GOP's bill to repeal Obamacare.
Without RAG	**Fake News Determination: False** Reason: The news is from Los Angeles Times and it is about a real event that happened in 2015.

Here are examples of which our framework failed:
Table 10: First Example of LLM's Response

Title of the news	**'No visas, bad jobs: Venezuelan emigrants reluctantly return home'**
With RAG	**Fake News Determination: True** Reason: The search results do not support the claim made in the news title and description. The articles are about refugee job fairs and the definition of refugees, and do not mention any Venezuelan emigrants returning home due to lack of jobs or visas."
Label of The news	**Real News** Reason: The news article is from Reuters, a reputable news agency.

Table 11: Second Example of LLM's Response

Title of the news	**Donald Trump Shames Kids For Not Registering To Vote For Him: 'They Feel Very, Very Guilty**
With RAG	**Fake News Determination: True** Reason: The news is supported by multiple credible sources, including Politico and CBC, and the snippets match the news description.
Label of The news	**Fake News** Reason: Although the news had been published by several credible sources, its labeled as fake in the dataset

Tables 8 and 9 show examples of how our framework helped to identify the authenticity of the news. Tables 10 and 11 show examples of which our framework failed to detect the authenticity of the news .

5.8 Results and discussion

In this section we compare the results of a previous study with our framework.

Table 12: Model Performance Comparison

Model	Accuracy	Precision	Recall	F1 Score
Logistic R	0.77	0.81	0.75	0.75
LSTM	0.726	0.760	0.718	0.718
DistilBERT	0.71	0.74	0.73	0.71
BERT	0.67	0.63	0.70	0.64
Ours	**0.88**	**0.94**	**0.81**	**0.87**

Table 4.8.1 compares Blackledge et al. [6] results with our framework. They explored the use of transformers for classifying news articles based on textual content. They specifically focused on both in-distribution and out-of-distribution generalization. The results shown are their out-of-distribution generalization experiments in which they trained their model on another data set and tested it on an ISOT data set.

They proposed a two-step classification pipeline to address fake news's subjective and inconsistent nature. This process involved identifying and removing opinion-based news articles from the training data, thus filtering out the most subjective samples to prevent models from learning patterns that do not generalize well.

The study demonstrated that transformers outperformed traditional models in news classification tasks. DistilBERT achieved a peak accuracy of 77.5 in out-of-distribution generalization. The two-step classification process improved deBERTa's accuracy by 7.8 to a peak of 80.8 and its F1 score by 10.1.

Kaliyar et al.[19] demonstrated the performance of their proposed model, FakeBERTa, a BERT-based deep convolutional approach for fake news detection. Their model combines BERT with three parallel blocks of 1D-CNN, each having different kernel-sized convolutional layers and various filters for enhanced learning. Built on top of a bidirectional transformer encoder-based pre-trained word embedding model (BERT), FakeBERTa achieves an impressive accuracy of 98.90.

6 COCLUSION

As we can see from the results (Table 4, 6), the RAG framework significantly enhanced our fake news detection model results. The improvements in precision, recall, F1-score, and overall accuracy demonstrate that integrating real-time, relevant context from external sources allows the LLM to make more accurate and informed decisions. This enhanced performance underscores the efficacy of the RAG framework in providing the necessary context to the LLM, thereby reducing errors and increasing the reliability of the detection process. We can see a false positive number increase after applying our framework which needs furthure investigation to deal with.

One of the key strengths of our framework is that it is not biased toward a specific dataset, as no training step is involved. Traditional machine learning models often suffer from biases introduced during the training phase, where the model learns patterns particular to the training data, which may not generalize well to new, unseen data. In contrast, our approach utilizes a dual-stage process involving real-time data retrieval and analysis, which ensures that the information processed and generated by the LLM is grounded in current, verifiable data. This methodology avoids the pitfalls of overfitting and bias, providing a robust and adaptive solution for fake news detection.

From the above examples (Table 4.8.3, 4.8.4), we can see that how RAG elevated the results in these three cases. The RAG framework helps provide more context and verify the information against credible sources. It identifies contradictions and evaluates the credibility of the sources, leading to a correct determination of fake news. Without RAG, The LLM fails to recognize the fake news correctly in all three cases because the LLM relies solely on the news article's content without cross-referencing it with other sources. This leads to incorrect determinations, as it doesn't account for the credibility of the sources or the lack of supporting evidence.

Our methodology holds significant value due to its unique approach of eliminating the training step, which helps avoid biases typically associated with model training phases and prevents the cost of training complex models. By testing our framework on the entire ISOT dataset, we ensure a comprehensive and robust evaluation of its performance. This thorough testing allows us to accurately gauge the model's effectiveness in real-world scenarios, demonstrating its capability to classify news articles with high precision and recall. From Table 4.8.1, The high precision of our framework minimizes false positives, which is crucial for maintaining trust in the system, and we can see that we outperformed Blackedge et al.'s experiments, which were done on a large portion of the data set. Our approach offers a reliable and efficient solution for fake news detection, showing superior performance metrics compared to traditional models. Although Kaliyar et al. achieved a much better performance in accuracy compared to our framework, they used 2080 sample of the data set for testing compared to 44398 records we used and they tested their model on the data set they trained in contrast with out source generalization approach of Blackedge et al.'s.

6.1 Future Works

There are several avenues for future research to enhance the effectiveness of our framework. One potential direction is to fine-tune different language models for fake news detection. By employing various models as classifiers and comparing their performance with Mixtral-8x7B, we can identify the most effective ones for this purpose. In recent months, Llama3 70B and Mixtral-8x22B have been released both are open source, which are more powerful than the LLM we used. Additionally, applying our RAG framework to different datasets could provide further insights into its generalizability and robustness. These efforts could lead to more accurate and reliable methods for combating fake news across diverse sources and topics.

A significant challenge in our study is the model's difficulty in accurately detecting fake news published by credible sources, as we can see in Table 11. These instances of misinformation are particularly problematic because they often contain elements of truth, making it harder for the model to distinguish between genuine

and false information. The current framework, which relies heavily on contextual retrieval and analysis using Mixtral-8x7B, is not equipped to handle this level of nuance. This issue underscores the need for additional mechanisms or features that can better assess the sources' credibility rather than just the content.

To enhance fake news detection, we propose a two-step search mechanism. The first step involves our current method, which searches for relevant news articles based on the initial query. The second step employs an advanced search strategy where the news article's title is used to find related news from a later date. This step aims to identify updates or additional reports about the news. Additionally, this search includes checking the news title and tags related to fake news to determine if any subsequent publications have identified the news as fake after its original publication date.

Moreover, we can build an efficient web scraping to get the full text of the search results, which could be challenging but would definitely provide us with more valuable data for detecting fake news compared to a summarized snippet of the news

Another approach to improve our framework involves integrating Named Entity Recognition (NER) [8]to enhance the prefiltering stage. While our current method utilizes the cosine similarity of embeddings to retrieve related articles, NER can provide a more precise filtering mechanism by extracting and focusing on key entities such as names, organizations, and dates. By combining the contextual richness of embeddings with the precision of entity-based filtering, we can significantly improve the relevance of the articles retrieved for comparison. This hybrid approach will allow our framework to more effectively cross-reference and verify the authenticity of news articles, leveraging the strengths of both semantic and entity-based analysis without requiring a traditional training phase.

Additionally, exploring the use of various embedding models can enhance the prefiltering stage of our fake news detection framework. While our current approach utilizes DistilBERT, other models such as BERT, RoBERTa, ALBERT, GPT-3, or T5 might provide improved contextual understanding and relevance. These models can capture deeper nuances and domain-specific contexts, potentially leading to more accurate and relevant article retrieval. By evaluating and potentially combining embeddings from multiple models, we aim to develop a more comprehensive and robust prefiltering mechanism. This approach seeks to further refine our framework's ability to identify and analyze pertinent news articles, thereby increasing its overall effectiveness in detecting fake news. Also we can experiment different types of similarity checks to measure their effect in the performance of our framework.

By incorporating these enhancements into our framework, we aim to develop a more reliable and efficient system for detecting fake news. Utilizing open-source large language models notably reduces the costs associated with data collection and computational training in machine learning approaches. These models offer versatility across various domains, enhancing their practicality and cost-effectiveness for diverse applications.

REFERENCES

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

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

[3] Hadeer Ahmed, Issa Traore, and Sherif Saad. 2018. Detecting opinion spams and fake news using text classification. *Security and Privacy* 1, 1 (2018), e9.

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

[5] Alaa Altheneyan and Aseel Alhadlaq. 2023. Big data ML-based fake news detection using distributed learning. *IEEE Access* 11 (2023), 29447–29463.

[6] Ciara Blackledge and Amir Atapour-Abarghouei. 2021. Transforming fake news: Robust generalisable news classification using transformers. In *2021 IEEE International Conference on Big Data (Big Data)*. IEEE, 3960–3968.

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[8] Indra Budi and Ryan Randy Suryono. 2023. Application of named entity recognition method for Indonesian datasets: a review. *Bulletin of Electrical Engineering and Informatics* 12, 2 (2023), 969–978.

[9] Nadia Conroy, Victoria Rubin, and Yimin Chen. 2015. Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology* 52, 3 (01 2015), 1–4. https://doi.org/10.1002/pra2.2015.145052010082

[10] Gabriele Corso, Zhitao Ying, Michal Pándy, Petar Veličković, Jure Leskovec, and Pietro Liò. 2021. Neural distance embeddings for biological sequences. *Advances in Neural Information Processing Systems* 34 (2021), 18539–18551.

[11] Emmanuel Gbenga Dada, Joseph Stephen Bassi, Haruna Chiroma, Shafi'i Muhammad Abdulhamid, Adebayo Olusola Adetunmbi, and Opeyemi Emmanuel Ajibuwa. 2019. Machine learning for email spam filtering: review, approaches and open research problems. *Heliyon* 5, 6 (2019), e01802. https://doi.org/10.1016/j.heliyon.2019.e01802

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[13] Luciano Floridi and Massimo Chiriatti. 2020. GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds and Machines* 30 (12 2020), 1–14. https://doi.org/10.1007/s11023-020-09548-1

[14] Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2024. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 22105–22113.

[15] Yue Huang and Lichao Sun. 2023. Harnessing the power of chatgpt in fake news: An in-depth exploration in generation, detection and explanation. *arXiv preprint arXiv:2310.05046* (2023).

[16] Gautier Izacard and Edouard Grave. 2021. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (Eds.). Association for Computational Linguistics, Online, 874–880. https://doi.org/10.18653/v1/2021.eacl-main.74

[17] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088* (2024).

[18] Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active Retrieval Augmented Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 7969–7992. https://doi.org/10.18653/v1/2023.emnlp-main.495

[19] Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. 2021. FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimedia tools and applications* 80, 8 (2021), 11765–11788.

[20] Cecilia Kang and Adam Goldman. 2016. In Washington Pizzeria Attack, Fake News Brought Real Guns. *New York Times* (December 2016), A1. https://www.nytimes.com/2016/12/05/us/politics/fake-news-guns.html Published online December 5, 2016.

[21] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.

[22] Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. Pre-trained language models for text generation: A survey. *Comput. Surveys* 56, 9 (2024), 1–39.

[23] Jiarui Li, Ye Yuan, and Zehua Zhang. 2024. Enhancing LLM Factual Accuracy with RAG to Counter Hallucinations: A Case Study on Domain-Specific Queries in Private Knowledge-Bases. *arXiv preprint arXiv:2403.10446* (2024).

[24] Rahul R Mandical, N Mamatha, N Shivakumar, R Monica, and AN Krishna. 2020. Identification of fake news using machine learning. In *2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*. IEEE, 1–6.

[25] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.

[26] Peng Qi, Yuyan Bu, Juan Cao, Wei Ji, Ruihao Shui, Junbin Xiao, Danding Wang, and Tat-Seng Chua. 2023. FakeSV: A Multimodal Benchmark with Rich Social Context for Fake News Detection on Short Video Platforms. *Proceedings of the AAAI Conference on Artificial Intelligence* 37 (06 2023), 14444–14452. https://doi.org/10.1609/aaai.v37i12.26689

[27] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[28] Nils Rethmeier and Isabelle Augenstein. 2023. A primer on contrastive pretraining in language processing: Methods, lessons learned, and perspectives. *Comput. Surveys* 55, 10 (2023), 1–17.

[29] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter* 19 (08 2017). https://doi.org/10.1145/3137597.3137600

[30] Vivienne Sze, Yu-Hsin Chen, Joel Emer, Amr Suleiman, and Zhengdong Zhang. 2017. Hardware for machine learning: Challenges and opportunities. In *2017 IEEE Custom Integrated Circuits Conference (CICC)*. 1–8. https://doi.org/10.1109/CICC.2017.7993626

[31] Tan Thongtan and Tanasanee Phienthrakul. 2019. Sentiment classification using document embeddings trained with cosine similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. 407–414.

[32] SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313* (2024).

[33] Meng Wang, Weijie Fu, Xiangnan He, Shijie Hao, and Xindong Wu. 2022. A Survey on Large-Scale Machine Learning. *IEEE Transactions on Knowledge and Data Engineering* 34, 6 (2022), 2574–2594. https://doi.org/10.1109/TKDE.2020.3015777

[34] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[35] Sam Wineburg, Sarah McGrew, Joel Breakstone, and Teresa Ortega. 2016. *Evaluating Information: The Cornerstone of Civic Online Reasoning*. Technical Report. Stanford History Education Group. https://purl.stanford.edu/fv751yt5934

[36] Shuo Yang, Kai Shu, Suhang Wang, Renjie Gu, Fan Wu, and Huan Liu. 2019. Unsupervised Fake News Detection on Social Media: A Generative Approach. *Proceedings of the AAAI Conference on Artificial Intelligence* 33, 1 (07 2019), 5644–5651. https://doi.org/10.1609/aaai.v33i01.33015644

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009