Beyond Fact-Checking: A Scalable, Domain-Agnostic, and Explainable System for Automated Fake News Detection

Marco Aspromonte[1], Giuseppe Contissa[1], Federico Galli[1] and Andrea Loreggia[2]

[1]University of Bologna - Italy
[2]University of Brescia - Via Branze 38 - 25123 - Brescia -Italy

Abstract

The rapid spread of fake news through social media and online platforms poses significant challenges to public discourse and democratic processes. Traditional fact-checking methods, while effective, cannot keep pace with the vast influx of digital content. To address this issue, we propose a novel, domain-agnostic framework for automated fake news detection (AFND) using large language models (LLMs) and search engine integration. Unlike domain-specific solutions, our approach verifies claims against evidence collected from a dynamically generated set of evidence, leveraging LLMs to assess the truthfulness of the input by comparing it with authoritative sources. Our framework emphasizes explainability, providing users with clear, evidence-based reasoning for its classifications. This transparency is crucial in building trust and complying with regulations, such as the EU Digital Services Act, which demands both content monitoring and justification of decisions. The system supports multilingual and multimodal capabilities, enhancing its versatility across various contexts. Through empirical evaluation of datasets such as Politifact and Liar, we demonstrate significant improvements in accuracy, precision, recall, and F1 scores when knowledge augmentation is applied. Our results highlight the potential of LLM-driven solutions in the ongoing fight against disinformation, offering a scalable, explainable tool for automated fake news detection.

Keywords

Fake News Detection, Disinformation, Large Language Models (LLMs), Explanation

1. Introduction

Fake news, defined as intentionally misleading or false information presented as factual [1], has become a significant issue in the digital era [2]. Social media platforms facilitate the rapid spread of misinformation, amplifying its reach and impact on public discourse, democratic processes, and societal trust [3]. The algorithmic nature of online platforms further exacerbates the problem, often prioritizing engagement over accuracy, making it increasingly difficult for individuals to distinguish truth from falsehood. Given the vast volume of online content, manual fact-checking alone is insufficient to address the scale of the issue [4, 5]. Automated Fake News Detection (AFND) has emerged as a crucial solution for countering misinformation [6]. By leveraging machine learning (ML) and natural language processing (NLP), these systems can analyze large volumes of data in real-time, comparing claims against credible sources to assess their accuracy [2]. However, for AFND to be effective and widely adopted, explainability is essential [7]. Users must understand why content is flagged as misinformation, particularly in high-stakes areas such as elections or public health. Explainable AI (XAI) methods, including those powered by large language models (LLMs), enhance transparency by providing interpretable justifications for decisions, increasing user trust and system reliability.

Recent regulatory developments, such as the European Union's Digital Services Act (DSA) [8] and the 2022 Code of Practice on Disinformation, emphasize both the detection and transparency of

ROMCIR 2025: The 5th Workshop on Reducing Online Misinformation through Credible Information Retrieval (held as part of ECIR 2025: the 47th European Conference on Information Retrieval), April 10, 2025, Lucca, Italy
*Corresponding author.
†These authors contributed equally.
marco.aspromonte2@unibo.it (M. Aspromonte); giuseppe.contissa@unibo.it (G. Contissa); federico.galli7@unibo.it (F. Galli); andrea.loreggia@unibs.it (A. Loreggia)
0000-0002-9846-0157 (M. Aspromonte); 0000-0002-9846-0157 (G. Contissa); 0000-0002-9846-0157 (F. Galli); 0000-0002-9846-0157 (A. Loreggia)

misinformation. These regulations impose obligations on large platforms to combat disinformation while ensuring clear explanations for content moderation decisions. In this context, we propose a novel, domain-agnostic AFND approach that bypasses traditional training processes. Our method gathers relevant articles from the web as evidence for claim verification, leveraging LLMs to interpret and validate information against the collected evidences. This framework enables scalable, explainable misinformation detection, which we detail further in the following sections.

2. Background

In this section, we introduce basic notions that are useful for understanding the proposed framework.

2.1. Collected Evidence

Fact-checking against collected evidence involves comparing a claim with retrieved information to assess its accuracy. Our approach dynamically gathers evidence by querying search engines in real-time, ensuring up-to-date verification as facts evolve. This method offers two key advantages: (1) it avoids reliance on outdated or static data, and (2) it leverages search engine rankings to focus on the top k most relevant results, enhancing efficiency and credibility.

We primarily use Google due to its robust ranking algorithm, which evaluates page authority based on link quality and quantity. This ensures that retrieved evidence is both relevant and authoritative.

No data is stored or retained; all retrieved information is used solely for validation and discarded immediately, ensuring privacy and compliance with data protection laws.

2.2. Generative AI and Large Language Models (LLMs)

Generative Artificial Intelligence is a subset of traditional machine learning, where models extract patterns from vast datasets to generate new content. Large language models (LLMs) are trained on trillions of words over weeks or months, developing billions of parameters and emergent capabilities beyond basic language processing.

LLMs are primarily based on the transformer architecture, leveraging attention mechanisms [9] like self-attention and multi-head attention to enhance text understanding. These mechanisms improve context awareness, long-range dependency capture, and nuanced interpretation. In this work, we compare different LLM architectures within our framework.

3. Related Work

Automated Fake News Detection (AFND) has gained attention due to the spread of misinformation on social media [10]. Initially, it was tackled as a binary classification problem using machine learning models, later improving with deep learning approaches such as Bi-GRUs [**?**]. Pre-trained language models have further enhanced performance, incorporating techniques like Knowledgeable Prompt Learning [11] and knowledge graph integration [12]. Few-Shot AFND (FSAFND) remains a challenge, relying on large language models (LLMs) for in-context learning [12, 13]. However, LLMs face issues such as ambiguity and hallucinations [14]. Recent efforts mitigate these limitations by incorporating external evidence [15] or generating justifications for classifiers [16]. The Dual-Perspective Augmented Fake News Detection model [17] exemplifies this approach by integrating internal and external knowledge. Beyond accuracy, explainability is crucial for fostering trust in AFND systems [7, 18]. Explainable AI (XAI) methods enhance transparency by generating interpretable justifications.

4. Proposed Approach

The proposed approach leverages the integration of search engine tools and incorporates LLM technology for fake news detection. LLM can extract information from local files as well as single and multiple web

pages. These data are then processed by LLMs, which have the ability to understand and analyse natural language in an advanced manner. Thanks to these models, every statement contained in potentially fake content can be assessed for truthfulness. In this context, claim verification is performed by comparing the collected information with authoritative and reliable sources.

In our approach, the proposed system simulates human fact-checking behaviour by gathering relevant information from a wide range of websites and online sources. It then evaluates the truthfulness of a given statement by cross-referencing the collected data with the claim, much like a human would consult multiple sources to verify a fact. While this process involves retrieving and assessing potential evidence, inconsistencies among the retrieved information may arise due to variations in content across sources. Although we acknowledge that such inconsistencies can affect the final judgment, our current approach relies on the LLM's capacity to resolve them. We plan to extensively analyse this issue in future work to refine our method and further improve the reliability of fact-checking.

The news/fact to be verified is given as input into the system. Search engine APIs are used to retrieve a list of websites that are related to the input. These websites will be used as a evidence for validating the input.

The validation process takes place through a comparative analysis between the claims extracted from potentially misleading texts and the official documents listed above. This rigorous approach ensures that every claim is verified with reliable sources, minimising the risk of spreading false or misleading information. In this way, the methodology provides a detailed description of the contents, highlighting both the accurate and potentially deceptive components. This detailed analysis facilitates the classification of information, helping users discern between reliable news and fake news.

Below are the macro-elements that compose the framework: 1. **Search Engine:** The first step of the process consists of collecting various website pages that are related to the input. Each URL is saved in a temporary dataset. The user can specify URLs that should be blacklisted and thus removed and not considered in the collected evidence. 2. **Collected Evidence:** The data collected from the web undergoes a verification and validation process to ensure the reliability of the sources (blacklist control). The LLM will use this data to compare and evaluate the claim, ensuring that the analysis is based on accurate and up-to-date information related to the topic addressed. The information collection algorithm is designed to "vote" for the most semantically relevant source. 3. **LLM textual inference:** The input is compared with information from the collected evidence. These sources are scalable based on the context so that the appropriate collected evidence can be selected as needed. The system is able to determine whether each claim is true, false, or unverifiable, providing a complete and precise response. This is done through an in-depth contextual analysis that takes multiple variables and linguistic nuances into account. 4. **Prediction:** The framework generates an outcome reporting whether the framework does not have enough information for classifying the claim or whether the claim is real or fake given the collected evidence. In the latter case, the system also provides an explanation about the validation of the claim, explaining why the claim is classified in a given class and what parts of the collected evidence refute or support the claim.

4.1. Multilingualism and Multimodality

The number of languages supported by the tool depends closely on the selected LLM model. For instance, the GPT-4o model that could be adopted by the system supports 50 languages, covering 97% of the languages spoken worldwide[1]. Furthermore, many LLMs are multimodal, accepting input not only in the form of text but also in video, image, and sound. This multimodal capability makes the tool extremely versatile and powerful, enhancing user interaction and expanding the possibilities for use in various contexts and applications.

In this work, we only test input in the form of text, leaving other content types as future work.

[1] https://openai.com/index/gpt-4o-and-more-tools-to-chatgpt-free/ - Last visited 9 October 2024

5. Empirical Evaluation

In this section, we detail the LLM models as well as the datasets adopted to assess the performance. In the end, we describe the empirical evaluation of our proposed approach. First, we check whether each baseline model is able to predict if an input statement is real or fake, and then we provide LLMs with some evidence about the input. This allows us to compare the performance and test if the evidence is useful to classify the input and improves the performance of the model. Moreover, we compare the ability to provide an explanation with or without evidence.

5.1. Large Language Models (LLMs)

In the experiments, we adopted several different Large Language Models (LLMs) to compare their performance and assess the extent to which the choice of model influences the final outcomes. Below is a list of the models used in our study, accompanied by a brief description of their key characteristics and features:

1. **mistral-7b-instruct-v0.3-bnb-4bit**[2]: This model employs sliding window attention, where each layer attends to the previous 4,096 hidden states. This architecture achieves linear computational complexity and a 2x speedup for sequences of up to 16k tokens with a 4k window while also reducing cache memory usage by 50% for sequences of 8,192 tokens without degrading model performance. It is designed for highly efficient processing of long-context documents, making it ideal for fact-checking and other tasks requiring extensive input sequences. Despite being quantised to 4 bits, it retains strong performance, particularly in instruction-following tasks. 2. **Phi-3-Mini-128K-Instruct**[3]: This model has 3.8 billion parameters, is part of the Phi-3 family, and is designed for lightweight yet high-performance tasks. It was trained on the Phi-3 dataset, which consists of both synthetic and high-quality public data with an emphasis on reasoning abilities. The model supports both 4K and 128K token context lengths, making it highly versatile. Its fine-tuning included preference optimisation to enhance instruction-following behaviour and adherence to safety protocols. Phi-3-Mini excels in areas like common sense reasoning, language understanding, math, coding, long-term context handling, and logical reasoning, performing competitively against models with up to 13 billion parameters. 3. **GPT-4o mini**: GPT-4o mini is a smaller variant of the GPT-4 architecture, optimised for tasks requiring general-purpose language understanding. Despite its smaller parameter count compared to full-scale GPT models, it retains much of the versatility of GPT-4, performing well across diverse tasks, including language comprehension, reasoning, and fact-checking. Its performance is particularly notable in contexts where computational resources are limited but where multi-turn reasoning and instruction-following are crucial. 4. **gemma2-9b-it**[4]: This 9-billion-parameter model is specifically optimised for high-level reasoning across multilingual datasets. The gemma2-9b-it model balances large-scale language understanding with efficiency, making it suitable for scenarios requiring rich semantic analysis in multiple languages. Its training includes a focus on factual accuracy, with the model excelling in both general-purpose NLP tasks and domain-specific knowledge applications. It shows robust performance in classification tasks, such as fact-checking, where detailed contextual understanding is required. 5. **LLaMA3-8b-8192**[5]: it is an 8-billion-parameter model from Meta's latest LLaMA series, optimised for longer sequences with an 8,192-token context window. This expanded context window makes it particularly well-suited for handling long-range dependencies in tasks like document-level fact-checking, legal document processing, and scientific text analysis. The model balances large-scale language understanding with efficiency, and its architecture is optimised to maintain performance while handling long contexts. It excels at instruction-following and has demonstrated strong results in tasks involving factual reasoning, general knowledge, and commonsense reasoning. Its 8k token capacity allows it to process extensive inputs while preserving high inference speed and

[2]https://huggingface.co/unsloth/mistral-7b-instruct-v0.3-bnb-4bit - Last visited 30 September 2024
[3]https://huggingface.co/microsoft/Phi-3-mini-128k-instruct - Last visited 9 October 2024
[4]https://huggingface.co/google/gemma-2-9b-it - Last visited 12 October 2024
[5]https://huggingface.co/unsloth/llama3-8b-8192 - Last visited 12 October 2024

accuracy.

The above models were selected for their diverse strengths in areas such as context-length handling, instruction-following, multilingual capabilities, and reasoning. By using this range of models, we aim to determine the impact of model choice on the accuracy and robustness of the results across different datasets and tasks.

5.2. Datasets

In order to evaluate our proposed approach, we tested our approach on different datasets that are commonly used in the literature. In particular, we adopted the following datasets: 1. **PolitiFact**[19]: A widely-used fact-checking dataset that contains real-world claims and their corresponding truthfulness ratings, derived from the PolitiFact website. It focuses on political statements made by public figures, covering a variety of topics. 2. **Liar**[20]: A dataset composed of short statements labelled for their truthfulness, sourced from fact-checking websites. The dataset contains 12,836 short statements in English collected in a grounded, more natural context, such as political debate, TV ads, Facebook posts, tweets, interviews, news releases, and many others. 3. **Weibo21**[21]: this dataset focuses on fact-checking claims across multiple domains, including Science, Military, Education, Accidents, Politics, Health, Finance, Entertainment, and Society. It is structured to assess a system's ability to handle both fact verification and claim validation tasks, making it a comprehensive benchmark for multi-domain fact-checking systems. The dataset is in Chinese.

We utilized balanced subsets of each dataset: for PolitiFact and Liar datasets, we randomly sampled 150 true and 150 false claims, resulting in a total of 300 samples per dataset. For the Weibo21 dataset, given its multi-domain nature, we adopted a stratified sampling strategy to maintain class balance within each category. Specifically, we selected around 18-20 samples true claims per category and 18-20 false claims per category. This resulted in a balanced distribution of claims for each category in the dataset.

When performing a query, the results returned by a search engine are not always consistent. This variability can be attributed to changes in the indexed content and inherent randomness in the search engine's algorithms. To ensure consistency in the evidence used by large language models during experiments, particularly when comparing different models, we took a controlled approach. For each sample in our dataset, we retrieved and stored the search results from the chosen search engine at the time of collection. This guarantees that all models are evaluated using the same evidence, regardless of when or how often the experiment is repeated.

5.3. Metrics

Each model was evaluated in two configurations: **Baseline**, i.e., a vanilla version without additional knowledge and **Knowledge Injection (KI)**, where external knowledge from articles is previously injected.

We employed standard accepted metrics, namely precision, recall, F1 score, and accuracy, as performance indicators.

In our evaluation, we compute various performance metrics based on the number of articles k used to support the inference process. However, in some cases, the total number of available articles N may be less than the desired k. To handle this, we calculate the metrics using the actual number of articles available, denoted by n, where $n = \min(k, N)$. This ensures that when k exceeds the number of available articles, the metric is computed using all available data.

5.4. Performance on Different Datasets

Table 1 presents the performance values for the different models on the three datasets considered. Each model's performance is evaluated under baseline conditions and various Knowledge Injection (KI) configurations using $K = 3$, $K = 5$, and $K = 7$.

Table 1
Performance values for the different models of the three datasets. For each model, metric, and dataset, the best value is in bold.

		Politifact				Liar				Weibo21			
Model	**Config.**	**Acc.**	**Prec.**	**Rec.**	**F1**	**Acc.**	**Prec.**	**Rec.**	**F1**	**Acc.**	**Prec.**	**Rec.**	**F1**
GPT4o mini	Baseline	0.55	0.51	0.77	0.62	0.53	0.49	**0.74**	0.59	**0.67**	**0.78**	0.48	0.60
	K = 3	**0.88**	0.83	**0.92**	0.87	**0.73**	**0.78**	0.50	**0.62**	0.59	0.57	0.82	0.67
	K = 5	0.87	**0.84**	**0.92**	**0.88**	**0.73**	**0.78**	0.50	**0.62**	0.59	0.56	**0.85**	**0.68**
	K = 7	0.87	0.83	0.91	0.87	0.72	**0.78**	0.50	0.61	-	-	-	-
Phi3 mini	Baseline	0.69	0.70	0.58	0.63	0.55	0.53	0.75	0.62	0.62	0.65	0.55	0.60
	K = 3	**0.86**	**0.90**	**0.85**	**0.88**	0.61	**0.58**	0.80	0.67	**0.68**	**0.70**	**0.65**	**0.67**
	K = 5	0.84	**0.90**	0.84	0.87	**0.63**	**0.58**	**0.82**	**0.68**	0.67	0.68	0.64	0.66
	K = 7	**0.86**	**0.90**	**0.85**	0.87	-	-	-	-	-	-	-	-
Mistral 7b	Baseline	0.53	0.50	0.67	0.57	0.56	**0.55**	0.67	**0.60**	0.55	0.50	0.60	0.55
	K = 3	**0.63**	**0.72**	**0.64**	**0.67**	**0.58**	0.25	**0.96**	0.39	**0.62**	**0.60**	**0.65**	**0.62**
	K = 5	0.61	0.69	0.62	0.65	**0.58**	0.23	0.87	0.36	0.60	0.58	0.63	0.60
	K = 7	0.60	0.68	0.61	0.64	-	-	-	-	-	-	-	-
LLaMA 8B	Baseline	0.53	0.52	**0.87**	0.65	0.64	0.65	0.91	0.75	**0.64**	**0.70**	0.54	0.61
	K = 3	**0.79**	**0.80**	0.62	**0.70**	0.88	0.83	0.92	0.87	0.59	0.56	**0.73**	**0.64**
	K = 5	0.73	0.78	0.51	0.61	0.88	**0.84**	0.92	**0.88**	0.58	0.55	0.72	0.63
	K = 7	0.74	0.79	0.51	0.62	**0.89**	**0.84**	**0.93**	**0.88**	-	-	-	-
Gemma2 9b	Baseline	0.69	**0.75**	0.76	0.75	**0.62**	**0.63**	0.55	0.59	**0.70**	**0.77**	0.60	0.67
	K = 3	**0.73**	0.74	0.81	**0.77**	0.57	0.56	**0.67**	**0.61**	**0.70**	0.70	**0.73**	**0.72**
	K = 5	0.67	0.65	**0.88**	0.75	-	-	-	-	-	-	-	-
	K = 7	-	-	-	-	-	-	-	-	-	-	-	-

As seen in Table 1, on the Politifact dataset, all models exhibit performance improvements when external knowledge is injected. In particular, GPT-4o mini and Phi3-mini show significant improvements in accuracy, precision, recall, and F1 score with 3 articles, with diminishing returns for more articles. Mistral 7b and LLaMA 8B also demonstrate enhanced performance with 3 articles, though LLaMA 8B's recall drops significantly in the $K = 5$ and $K = 7$ configurations. The gemma2-9b-it model demonstrates competitive performance, particularly in recall with 5 articles. For the final model, it was not possible to run the experiment with $K = 7$ due to token length limitations.

Mistral 7b performs poorly on the Liar dataset, especially in terms of precision and F1 score for K=3 and K=5, which could be attributed to the model's inability to focus on relevant information across multiple retrieved documents (as seen in the high recall but very low precision). This result might suggest that Mistral struggles with effectively distributing its attention, leading to a large number of false positives. Additionally, the fact that no results were retrieved for K=7 hints at potential limitations in handling increased complexity when more documents are involved, further supporting the idea that its attention mechanism might be less effective for this task.

Overall, Liar seems a much harder dataset as its statements have been collected with the intention of being difficult to classify.

As presented in Table 1, the estimated performance for Mistral 7b and Phi3-mini follows a similar trend to their results on the Politifact and Liar datasets, with improvements in accuracy, precision, and F1 score when additional knowledge is injected.

For the Weibo21 dataset, we evaluated the models' performance across various categories beyond political claims. We do not report the results due to space limits. In general, they reveal that Knowledge Injection improve the accuracy on some categories.

5.5. Discussion

The results clearly demonstrate that external Knowledge Injection (KI) significantly enhances the performance of all models.

For GPT-4o mini, these gains are most pronounced with the first few articles introduced, but as more articles are added, the performance plateaus or even declines slightly. This suggests that GPT-4o mini extracts the most relevant information from the initial articles, while additional sources offer diminishing returns and may introduce irrelevant or conflicting data, reducing overall effectiveness.

The gemma2-9b-it model's performance on the Liar dataset shows a competitive baseline but limited improvement with Knowledge Injection configurations, highlighting the need for further exploration in optimizing this model for different configurations.

LLaMA 8B demonstrates the most significant improvement on the Liar dataset, particularly in the cleaned configurations, where it achieves an accuracy of 88.97% with $K = 5$ articles. GPT-4o mini, Phi3-mini, and Mistral 7b also show considerable gains, though Mistral 7b struggles with precision in $K = 5$ and $K = 7$ configurations, suggesting potential issues with noise in additional articles.

Phi3-mini, Mistral 7b, and LLaMA 8B exhibit robust and steady improvements with the addition of external articles. This is particularly evident in their recall metrics, where all three models show a consistent ability to retrieve and utilize relevant information from the growing pool of external sources. LLaMA 8B, in particular, demonstrates exceptional performance on the Liar dataset, achieving the highest overall accuracy and F1 score across all configurations. This result suggests that LLaMA 8B is more adept at processing and integrating a larger volume of external knowledge, making better use of the supplementary data to enhance its performance without being overwhelmed by noise.

6. Explanation

Beyond classification, the system generates explanations by outlining the evidence supporting its predictions, ensuring transparency and interpretability. Example 6.1 illustrates how leveraging relevant evidence enhances both accuracy and justification, reinforcing confidence in the system's decisions.

Example 6.1. *"The past president (Barack Obama) brought Al Sharpton into the White House something like 80 times."*

- **Ground Truth Class**: REAL
- **Predicted Class (with Baseline approach)**: FAKE
- **Explanation (with Baseline approach)**: The claim that Barack Obama brought Al Sharpton into the White House "something like 80 times" is exaggerated and misleading. While Al Sharpton did visit the White House during Obama's presidency, official records indicate that he visited approximately 30 times. This includes meetings related to civil rights and social justice issues. The figure of 80 visits lacks substantiation and seems to overstate Sharpton's presence at the White House.
- **Predicted Class (with KI approach)**: REAL
- **Explanation (with KI approach)**: The evidence states that "as of December 2014, Sharpton had made 72 official recorded visits to the White House." This portion of the evidence supports the claim by providing a concrete number of visits. Furthermore, the statement indicates that not all visits were recorded, implying the actual number could be higher. Congressman Grothman estimated "around 80", which aligns closely with the claim that Al Sharpton was brought into the White House approximately 80 times. Thus, the claim is validated by the evidence provided.
- **Evidence**: The following evidence (refined by the model inference) was presented:

  > "Biden Is Even Worse than Obama (unemployment, Iraq, generation, poll) - Politics and Other Controversies - Democrats, Republicans, Libertarians, Conservatives, Liberals, Third Parties, Left-Wing, Right-Wing, Congress, President - Page 4 - City-Data Forum ... Bernadette Green, Grothmanś director of communications, said that as of December

2014, Sharpton had made 72 official recorded visits to the White House. 'Keep in mind that not all visits were recorded, so that number could be higher, and we have no way of knowing,' Green said in an email. 'Congressman Grothman is correct in saying 'around 80' as he was estimating, and was not too far off the actual number (and again, these are only the visits recorded up until Dec. 2014).'"

The knowledge-injected (KI) approach shows improvement over the baseline in both examples by providing more accurate explanations aligned with the evidence, even when the claim classification is correct. Looking at the ground truth class, in the first case, the KI approach correctly supports the claim about Al Sharpton's White House visits with detailed evidence, while the baseline mistakenly refutes the claim.

In the second example, both approaches correctly classify the R. Kelly claim as FAKE, but the KI explanation more directly reflects the comprehensive evidence, making it more reliable and detailed. This result highlights KI's ability to better leverage evidence for both classification and explanation.

In general, the two examples suggest that LLMs can distil and refine complex information, presenting it in a concise manner that directly addresses the claim. This characteristics leads to improved readability, as the evidence is organised and explicitly linked to the claim's verification or debunking. By synthesising information in a structured format, LLMs facilitate a better understanding of the relationship between claims and supporting evidence.

7. Conclusions and Future Work

In this paper, we presented a scalable, domain-agnostic, explainable framework for automated fake news detection that integrates large language models (LLMs) with search engine tools to verify the truthfulness of claims. By dynamically generating some evidence from relevant web sources and comparing claims against these data, the proposed approach addresses the growing challenges posed by the rapid spread of disinformation in the digital age. The framework also emphasises explainability, ensuring that users can understand and trust the system's decisions, which is crucial in sensitive areas such as political discourse or regulatory compliance.

The empirical evaluation of our approach on the Politifact, Weibo21, and Liar datasets demonstrated significant performance improvements, especially when leveraging knowledge augmentation. This result highlights the effectiveness of combining LLMs with external data sources for scalable and transparent fake news detection.

In future work, we plan to extend our framework to handle more complex, multi-faceted statements. The current system is optimised for verifying relatively short claims, but many real-world instances of disinformation involve nuanced or multi-layered assertions that require a deeper level of contextual understanding and multi-step reasoning. By enhancing the framework to process these more intricate claims, incorporating advanced LLM-based techniques, and expanding the scope to include multi-modal content such as images and videos, we aim to further improve the system's accuracy and robustness. This expansion will enable the framework to tackle a broader range of fake content, making it an even more powerful tool in the fight against disinformation.

References

[1] M. D. Molina, S. S. Sundar, T. Le, D. Lee, "fake news" is not simply false information: A concept explication and taxonomy of online content, American behavioral scientist 65 (2021) 180–212.

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

[3] T. Duile, S. Tamma, Political language and fake news: Some considerations from the 2019 election in indonesia, Indonesia and the Malay World 49 (2021) 82–105.

[4] F. Galli, A. Loreggia, G. Sartor, The regulation of content moderation, in: International Conference on the Legal Challenges of the Fourth Industrial Revolution, Springer, 2022, pp. 63–87.
[5] A. Loreggia, G. Sartor, et al., Artificial intelligence and the moderation of digital platforms, Sistemi Intelligenti 34 (2022) 53–73.
[6] A. Alaphilippe, A. Gizikis, C. Hanot, K. Bontcheva, Automated tackling of disinformation, Technical Report, European Parliamentary Research Service, 2019.
[7] A. Athira, S. M. Kumar, A. M. Chacko, A systematic survey on explainable ai applied to fake news detection, Engineering Applications of Artificial Intelligence 122 (2023) 106087.
[8] M. Leiser, Analysing the european union's digital services act provisions for the curtailment of fake news, disinformation, & online manipulation, 2023. URL: osf.io/preprints/socarxiv/rkhx4. doi:10.31235/osf.io/rkhx4.
[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems 30, Curran Associates, Inc., 2017, p. 5998–6008.
[10] K. Shu, A. Sliva, S. Wang, J. Tang, H. Liu, Fake news detection on social media: A data mining perspective, ACM SIGKDD explorations newsletter 19 (2017) 22–36.
[11] G. Jiang, S. Liu, Y. Zhao, Y. Sun, M. Zhang, Fake news detection via knowledgeable prompt learning, Information Processing & Management 59 (2022) 103029.
[12] J. Ma, C. Chen, C. Hou, X. Yuan, Kapalm: Knowledge graph enhanced language models for fake news detection, in: Findings of the Association for Computational Linguistics: EMNLP 2023, 2023, pp. 3999–4009.
[13] B. Hu, Q. Sheng, J. Cao, Y. Shi, Y. Li, D. Wang, P. Qi, Bad actor, good advisor: Exploring the role of large language models in fake news detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, 2024, pp. 22105–22113.
[14] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, ACM Computing Surveys 55 (2023) 1–38.
[15] T. W. Teo, H. N. Chua, M. B. Jasser, R. T. Wong, Integrating large language models and machine learning for fake news detection, in: 2024 20th IEEE International Colloquium on Signal Processing & Its Applications (CSPA), IEEE, 2024, pp. 102–107.
[16] B. Wang, J. Ma, H. Lin, Z. Yang, R. Yang, Y. Tian, Y. Chang, Explainable fake news detection with large language model via defense among competing wisdom, in: Proceedings of the ACM on Web Conference 2024, 2024, pp. 2452–2463.
[17] Y. Liu, J. Zhu, K. Zhang, H. Tang, Y. Zhang, X. Liu, Q. Liu, E. Chen, Detect, investigate, judge and determine: A novel llm-based framework for few-shot fake news detection, arXiv preprint arXiv:2407.08952 (2024).
[18] V. U. Gongane, M. V. Munot, A. D. Anuse, A survey of explainable ai techniques for detection of fake news and hate speech on social media platforms, Journal of Computational Social Science (2024) 1–37.
[19] N. Vo, K. Lee, Where are the facts? searching for fact-checked information to alleviate the spread of fake news, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 7717–7731.
[20] W. Y. Wang, " 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), Association for Computational Linguistics, 2017.
[21] Q. Nan, J. Cao, Y. Zhu, Y. Wang, J. Li, Mdfend: Multi-domain fake news detection, in: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, 2021, pp. 3343–3347.