Computers and Electrical Engineering 128 (2025) 110746

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

The blueprint of a new fact-checking system: A methodology to enrich RAG systems with new generated datasets☆

Salvador Lopez-Joya *, Jose A. Diaz-Garcia, M. Dolores Ruiz, Maria J. Martin-Bautista

Department of Computer Science and Artificial Intelligence, University of Granada, Daniel Saucedo Aranda, s/n, 18014 Granada, Spain

ARTICLE INFO

Keywords:
Fact-checking
RAG
Language models
Datasets
Natural language processing

ABSTRACT

In an era where digital misinformation spreads rapidly, Artificial Intelligence (AI) has become a crucial tool for fact-checking. However, the effectiveness of AI in this domain is often limited by the availability of high-quality and scalable datasets to train and guide algorithms. In this paper, we introduce VERIFAID (VERIfication FAISS-based framework for fake news Detection), a novel framework that improves fact-checking through a Retrieval-Augmented Generation (RAG) system based on automatically generated and dynamically growing datasets. Our approach improves evidence retrieval by building a scalable knowledge base, reducing the reliance on manually annotated data. The system consists of three key modules: two dedicated to dataset creation and one inference module that integrates advanced language models, such as LLaMA, within the RAG paradigm. To validate our methodology, we provide technical specifications for both the system and the dataset, together with comprehensive evaluations in zero-shot fact-checking scenarios. The results demonstrate the efficiency and adaptability of our approach and its potential to improve AI-driven fact verification at scale.

1. Introduction

Our society is profoundly impacted by the integration of Artificial Intelligence (AI) into our daily lives, leading many researchers and policymakers to describe this period as the dawn of a revolution. This transformation is driven by the emergence of Large Language Models (LLMs) that can improve productivity, facilitate self-learning [1], and improve customer service [2]. However, the same technology that brings these benefits can also be misused for harmful purposes, such as spreading misinformation to influence society in various negative ways [3]. This duality poses a significant challenge: AI has the potential to both combat and propagate harmful content.

A further challenge in leveraging AI to address misinformation, as well as other issues such as engineering applications, is the lack of high-quality data for training and testing algorithms [4]. Without sufficient data, AI models cannot achieve their full potential in accurately identifying and countering misinformation. Recent advances with the introduction of Language Models (LMs) and LLMs aim to address these challenges through zero-shot [5,6] or few-shot [7] scenarios—that is, classifying unseen instances without prior specific training. These approaches are showing promising results and appear to be well suited to real-world misinformation detection scenarios, where training and retraining models may be impractical. Consequently, new techniques, datasets, and methods tailored to these emerging scenarios are essential.

☆ This article is part of a Special issue entitled: 'idagi' published in Computers and Electrical Engineering.

* Corresponding author.
E-mail addresses: slopezjoya@ugr.es (S. Lopez-Joya), joseangeldiazg@ugr.es (J.A. Diaz-Garcia), mdruiz@decsai.ugr.es (M.D. Ruiz), mbautis@decsai.ugr.es (M.J. Martin-Bautista).

https://doi.org/10.1016/j.compeleceng.2025.110746
Received 17 March 2025; Received in revised form 1 August 2025; Accepted 1 October 2025
Available online 9 October 2025

Given that LLMs have been trained on vast amounts of data, a new research direction to solve zero-shot problems has emerged. This research line leverages this extensive internal knowledge for inference and classification. This approach can work in two ways: directly querying the LLM to classify or infer information, or enhancing its capabilities by integrating a retrieval module that searches for relevant contextual information, improving the accuracy and reliability of the model's responses. This leads to the concept of Retrieval-Augmented Generation (RAG) [8] system—a framework designed to enhance LLM performance by combining two key components:

- Retrieval Module: Searches for relevant external information, retrieving supporting documents or factual evidence.
- Generation Module: Uses the retrieved information to generate more accurate, contextually relevant, and informed responses.

By integrating these two elements, RAG systems significantly improve the quality of inference and text generation, making them particularly valuable in tasks like question answering [9]. Our research has three primary objectives: (1) to propose a comprehensive retrieval module capable of generating high-quality datasets, (2) to develop a robust fact-checking framework under Zero-Shot scenario integrating this data into a RAG system, (3) to create and make available a dataset to enhance other fact-checking systems and methodologies. Injecting the information of this dataset into the fact-checking system, we demonstrate that our RAG-based approach is a powerful tool for cross-topic fact-checking in few-shot scenarios, achieving competitive results over benchmark datasets without requiring prior training. Furthermore, our final dataset is designed to combat misinformation by providing a rich and diverse collection of textual data that can be used for both training and validation of AI models. In doing so, our goal is to improve the ability of AI to detect and mitigate false information, ultimately contributing to a more informed society.

The retrieval module and dataset have been developed using cutting-edge technologies, specifically Facebook AI Similarity Search (FAISS). This innovative technology utilizes embedding databases to perform similarity searches in a fast and efficient way, effectively overcoming the challenges posed by traditional embedding systems, particularly in terms of computational time required for calculating similarities and distances between embeddings [10]. FAISS constructs an embedding space by building neighbourhoods or clusters of the most similar elements, while also maintaining an index that allows for the rapid dismissal of vast amounts of comparisons. We leverage this capability in conjunction with web scrapers and language models to generate a substantial amount of content related to misinformation detection. This vast amount of knowledge provides one of the key strengths of our model, the ability to not only classify unseen examples with competitive accuracy but also provide claims and interpretative clues to support analysis and assist in the fact-checking process.

To summarize, our research makes the following key contributions to the state of the art:

- We propose the blueprint of VERIFAID (VERIfication FAISS-based framework for fake news Detection), an interpretable fact-checking system based on the RAG paradigm.
- We introduce a novel framework for a retrieval module in RAG systems, integrating language models with FAISS architecture to construct vast and efficient textual datasets, particularly suited for zero-shot scenarios.
- We present VERIFAID-dataset a comprehensive fact-checking dataset, derived from the most influential fact-checking sources, which can enhance the knowledge used in the RAG system's generation module while also providing interpretative insights to support fact-checking.

To validate our framework and dataset, we have evaluated the model over widespread benchmark datasets and we have also applied our system to a use case focused on zero-shot fact-checking. By exploring our results, we aim to demonstrate the versatility and effectiveness of our dataset for cross-domain applications. Additionally, we provide a comprehensive overview of the entire workflow, including the engineering details involved in creating the retrieval and generation modules. This detailed documentation serves as a blueprint for the community, enabling them to develop their own high-quality datasets, thereby facilitating further advancements in AI-driven fact-checking and other applications.

The paper is organized as follows. In the following section, we study related work and different approaches for creating vast datasets as well as existing datasets for fact-checking. In Section 3 we go into detail in the framework describing each of the steps and technical details, as well as the details of the created dataset. In Section 4, we explain the evaluation and validation of the framework providing a use case that shows the potential of our system. Finally, in Section 5, we examine the conclusions remarks and the future work of the research carried out.

2. Related work

To facilitate a thorough comparison with existing state-of-the-art systems, we have organized our related work into three categories according our main objectives: methodologies for dataset creation, existing fact-checking datasets and finally as our dataset tends to exhibit its potential uses in fact-checking under a zero-shot configuration, we also provide a review of the latest techniques in this area.

Table 1
Overview of popular Fact-Checking datasets.

Dataset	Claims	Labels	Topic	Evidence	
				Type	Source
FEVER [15]	185,445	Supported, Refuted, Not Enough Info	Diverse	Text	Wikipedia
FEVEROUS [16]	87,026	Supported, Refuted, Not Enough Info	Diverse	Text/Table	Wikipedia
CLIMATE-FEVER [17]	1535	Supports, Refutes, Not Enough Info	Climate	Text	Wikipedia
AVeriTeC [18]	4568	Supported, Refuted, Not Enough Info, Conflicting Evidence	Diverse	Text	Fact-Checking Websites
WatClaimCheck [19]	33,697	True, False, Partially True/False (source labels)	Diverse	Text	Fact-Checking Websites
PUBHEALTH [20]	11,832	True, False, Mixed, Unproven	Health	Text	Fact-Checking Websites
SciFact [21]	1409	Supports, Refutes	Science	Text	Research Papers
Fakeddit [22]	1,063,106	True, Satire, Misleading, Manipulated, False Connection, Imposter	Diverse	Text/Image	Reddit
PHEME [23]	4842	Supporting, Denying, Underspecified	Rumours Diverse	Text	Twitter
X-Fact [24]	31,189	True, Mostly-True, Partly-True, Mostly-False, False	Diverse	Text/Metadata	Fact-Checking Websites
LIAR [25]	12,836	Pants-fire, False, Barely-true, Half-true, Mostly-true, True	Politics	Text/Metadata	Politifact
VERIFAID dataset (Ours)	33,337	24 Labels (Politifact + Snopes System)	Diverse	Text/Metadata	Fact-Checking Websites

2.1. Vast dataset creation methodologies

Regarding the use of external information sources for dataset creation, which can lead to improvements in systems that leverage this data, significant advances have been made in the literature. For instance, [11] reviewed 248 papers addressing the problem of web crawling—developing automated systems capable of navigating the web to retrieve relevant information about a specific topic. Essentially, web crawlers represent the first step in creating datasets tailored to specific topics. Efforts to develop methodologies for dataset creation span diverse domains that require data-driven applications and analysis, including energy consumption [12] and document content analysis solutions [13]. Authors in [13] introduced a prototype system that, a decade ago, highlighted the necessity of automated systems for generating ground truth and datasets to support various types of evaluations. More recently, this kind of automatic systems has been demonstrated as the most useful, necessary and widespread. In [12], the authors made significant contributions to data-driven power system applications by introducing a scalable and efficient methodology for dataset creation. Their approach addresses critical challenges such as computational efficiency, dataset balance, and the handling of high-dimensional input spaces. The methodology integrates infeasibility certificates, strategic sampling, statistical fitting, and a strong emphasis on computational optimization to generate balanced datasets that accurately represent the security characteristics of power systems.

In the area of fact-checking, innovative approaches have emerged [14], introducing novel methodologies for claim identification and multilingual crawling to improve fact-checking capabilities. Building upon this, a multilingual fact-checking system has been proposed that leverages Wikipedia as its primary source of evidence. This system utilizes comprehensive Wikipedia snapshots to provide a rich and diverse information pool for assessing claim veracity. To accommodate different languages, machine translation is employed. The system incorporates the Question Answering for Claim Generation (QACG) method to generate claims using the retrieved Wikipedia documents. This approach enables the creation of adaptable training datasets for various languages by grounding claim generation and evidence within Wikipedia. Finally, an inference layer is employed to evaluate the retrieved claims and determine whether a given claim is true or false.

In our paper, we introduce a novel framework for a retrieval module in RAG systems, designed to create a vast and accurate dataset for use in fact-checking systems. One of the key distinctions from existing literature is that our approach provides a robust methodology that processes and leverages highly reliable sources in an automated manner, ensuring the creation of a reliable dataset for fact-checking across diverse topics. Additionally, our method offers a highly scalable solution by utilizing vector databases, an aspect that remains largely unexplored in current research.

2.2. Fact-checking datasets adaptable to RAG systems

RAG systems rely on high-quality sources of information that provide both factual claims and their corresponding evidence. Several datasets have been developed over the years to support fact-checking tasks, ranging from manually curated claims with Wikipedia evidence to large-scale, noisy social media datasets (see a summary in Table 1). In this section, we review popular datasets that align with RAG-based fact-checking approaches.

One of the most widely used datasets in the fact-checking domain is FEVER [15], which provides over 185,000 claims extracted from Wikipedia. Each claim is annotated with a label (Supported, Refuted, or Not Enough Info) based on evidence retrieved from Wikipedia articles. This dataset has been essential in advancing research in claim verification, particularly in retrieval-based fact-checking models. Building upon FEVER, FEVEROUS [16] extends the dataset to include tabular data from Wikipedia, improving the ability of fact-checking models to handle structured information.

In the climate change domain, CLIMATE-FEVER [17] adapts the FEVER methodology to assess climate-related claims, providing a focused dataset with claims labelled as Supports, Refutes, or Not Enough Info. This dataset is particularly valuable for verifying misinformation in environmental discourse. Similarly, SciFact [21] introduces a science-focused dataset containing claims derived from biomedical research papers. This dataset supports retrieval-based fact-checking by linking claims to relevant scientific publications.

Several datasets have been created using claims from professional fact-checking organizations. AVeriTeC [18] is a recent dataset that compiles claims from multiple fact-checking websites and pairs them with supporting or refuting evidence. Unlike FEVER, which uses Wikipedia as its knowledge base, AVeriTeC incorporates real-world fact-checking sources, making it more aligned with practical fact-verification tasks. Similarly, WatClaimCheck [19] and PUBHEALTH [20] provide claims with veracity labels and supporting evidence, with a focus on diverse and health-related misinformation, respectively.

Social media platforms serve as major sources of misinformation, prompting the creation of datasets like Fakeddit [22] and PHEME [23]. Fakeddit is a large-scale dataset containing over one million Reddit posts categorized into multiple misinformation classes, including True, Satire, Misleading, and False Connection. It incorporates both textual and visual information, making it suitable for multimodal fact-checking. PHEME, on the other hand, focuses on rumour detection in Twitter conversations, labelling tweets as Supporting, Denying, or Underspecified in relation to a claim.

A notable multilingual dataset is X-Fact [24], which contains claims in multiple languages with veracity labels from fact-checking websites. Similarly, LIAR [25] provides short political statements classified into six levels of truthfulness, using Politifact as its primary source.

Unlike other works that present static datasets, our contribution focuses on a scalable and continuously growing database designed to serve as a reliable knowledge source for fact-checking systems based on the RAG paradigm. Instead of relying on manually annotated data, our approach enables the automatic and dynamic incorporation of new information without human intervention, ensuring that the knowledge base continuously evolves with the flow of information from fact-checking websites. Our dataset in the VERIFAID framework consists of over 33,000 verified news articles and hoaxes, placing it among the top five largest datasets for fact-checking. Additionally, we offer the most comprehensive compendium of real information, enabling other systems or fact-checkers to effectively verify and contrast news pieces.

2.3. Zero-shot and few-shot configurations for fact-checking

The emergence of LLMs has revolutionized how systems and humans interact with AI. The immense amount of data used during their training has enabled these models to unlock remarkable potential for performing classification tasks on previously unseen examples. Two key paradigms in this context are zero-shot learning, where models rely solely on pre-trained knowledge, and few-shot learning, where they leverage a small set of examples from the target dataset to guide its structure, context, and predictions. These approaches are particularly relevant for fact-checking, as misinformation spreads rapidly, making it impractical to rely solely on traditional machine learning models trained on static datasets.

Authors of [26] proposed TELLER, a systematic framework designed for trustworthy fake news detection that prioritizes both accuracy and reliability. It consists of two key components: the cognition system, which integrates human expertise with LLMs to generate structured Yes/No questions aligned with logical predicates, and the decision system, which aggregates responses to classify news. The cognition system leverages external tools like search engines to enhance information retrieval. While this framework performs well on datasets like LIAR and Politifact using supervised training, it differs from our zero-shot approach, which does not rely on explicit training sets. Similarly, [27] introduced MAPLE, an approach that takes advantage of microlanguage evolution paths to analyse the relationship between claims and evidence. Additionally, the authors proposed *SemSim*, a new evaluation metric designed to measure semantic similarity.

In [28], the authors proposed a RAG system that uses information retrieved from the Internet. To classify a claim, the system performs an online search, filtering and classifying the retrieved data to extract the most relevant evidence. Finally, GPT-4 is employed to classify the claim based on this evidence. While the system achieved impressive results, it relies on a final classification layer trained on datasets like BuzzFeedNews, PolitiFact-Snopes-2024, and FakeNewsNet, making it non-zero-shot. In contrast, our approach is designed specifically for scenarios with minimal labelled data, making it more applicable to real-world where the lack of labelled data is a significant challenge.

In a similar way of previous work, Singhal et al. [29] proposed an automatic fact-checking system using a RAG configuration, built on the AVeriTeC dataset, which pairs claims with a knowledge store containing articles. Each claim is annotated with question–answer pairs, a veracity label, and a justification. The system, employing a hybrid LLM architecture, classifies claims into categories such as Supported, Refuted, Conflicting Evidence/Cherrypicking, or Not Enough Evidence, aiming to provide transparent, evidence-based veracity predictions and foster public trust. While the system outperforms the baseline on AVeriTeC, its evaluation lacks comparisons with other models. A key strength of the model is its ability to demonstrate that leveraging external information — specifically, retrieving the most relevant articles for each claim — significantly improves performance.

The concept of leveraging external information shown in [29] and [28] aligns with our approach, where we extend this idea by incorporating a much larger volume of information sourced from social news newspapers, unlocking a vast potential for knowledge injection into our system.

In [30], the authors introduce KAPALM, a novel model that leverages knowledge graphs for fake news classification. The system combines BERT with an adapter and integrates a knowledge graph through a graph neural network for classification. They employed a few-shot configuration, using 2 to 100 random news samples from the dataset to adjust the knowledge representation, focusing on the PolitiFact and GossipCop datasets. Their model achieved impressive results when trained on the full dataset, with F1 scores of the fake news of 0.9134 on PolitiFact and 0.7168 on GossipCop. However, in the few-shot setting, the F1 scores dropped to 0.5278 and 0.4226, respectively, when only two labelled examples were used for training.

An interesting perspective on the few-shot paradigm is presented by [31], where the authors propose a prompt-based system that also incorporates user interactions to align and guide classification. They fine-tuned a BERT model using various types of prompts, such as '*Prompt 3: Article with [MASK]: {news article input}*', replacing the [MASK] token with labels like real or fake. The model was trained on datasets like GossipCop, PolitiFact, and FANG, using 16 to 128 examples. A key innovation is the social alignment matrix, which estimates users' likelihood of engaging with true information based on past interactions, improving classification. This approach achieved 0.8530 accuracy on PolitiFact, demonstrating the value of user behaviour integration.

Also in the field of prompt learning, Ouyang et al. [32] introduced COOL, a novel framework for domain-adaptive fake news detection using a few-shot training paradigm. COOL incorporates a comprehensive knowledge extraction module that retrieves and integrates both structured and unstructured knowledge relevant to news stories related to COVID-19. The framework employs a hybrid prompt learning strategy, blending traditional hard prompts with learned soft prompts. Additionally, it leverages an adversarial contrastive training approach to model domain-invariant interaction patterns between news and knowledge.

In the context of zero-shot configurations, [33] proposed a zero-shot fact-checking system for the FEVEROUS challenge, indexing Wikipedia sentences and phrases related to claims. It used BERT with a Question-Answering configuration to classify claims as refuted or supported. While its results were not the best in the challenge, it offers a solid foundation for future RAG-based systems. Our framework builds on this concept but enhances indexing using FAISS and improves QA capabilities through LLMs instead of BERT-based retrieval.

Using LLMs, [34] proposed a system that achieved impressive results on benchmark datasets such as LIAR and PolitiFact. The system employs GPT-3.5 as the LLM in conjunction with an information retrieval model to enhance fake news detection. Starting with a claim, the system retrieves the most relevant news articles from the web, generates a set of evidence from these documents, and infers whether the claim is true or false. For claims lacking sufficient evidence, the system iteratively refines its search process, improving both the retrieval mechanism and the prompt engineering for the LLM. The system achieved remarkable results, with F1 scores exceeding 0.8. While this surpasses the state of the art for a zero-shot configuration, it relies on an exhaustive evidence retrieval process that mirrors searching for truth on the Internet, making it less predictive. The authors also evaluated the system without the retrieval component, reaching an F1 score of 0.56 on the LIAR dataset, which is closer to typical state-of-the-art results.

Compared to previous research, our system operates in a pure zero-shot scenario, making it more representative of real-world conditions where training data is unavailable. Despite this constraint, it achieves competitive results even when compared to trained and few-shot models. One of the key differences between our system and the existing literature is that it does not rely on retrieving new information or external clues during the inference process. This is particularly important in tasks like fake news detection, where searching the internet may yield claims about a given statement.

3. VERIFAID - VERIfication FAISS-based framework for fake news Detection

As we have discussed, there is a need for reliable and efficient fact-checking systems, and the RAG paradigm presents a promising approach to develop these systems. This section outlines the methodology for creating a fact-checking system that integrates different data sources, uses advanced text processing techniques, and uses a vector database for efficient retrieval.

As can be seen in Fig. 1 we break down the process into four subsections: Fact-checking sources; Metadata, claim, and label extraction; FAISS vector database; and Answer generation process.

3.1. Fact-checking sources

Ensuring the reliability of a fact-checking system is based on the integrity and objectivity of its data sources. To construct a robust and transparent verification pipeline, we have prioritized well-established, reputable fact-checking organizations. These sources form the basis of our system, providing claims, evaluations and supporting evidence.

A major challenge in compiling a comprehensive dataset is the heterogeneity of data formats. Fact-checking sources provide information in both structured and unstructured forms. While structured data facilitates the extraction of claims and verdicts, unstructured data often requires advanced processing techniques, including natural language understanding models, to derive meaningful insights.

For this study, we selected four primary fact-checking platforms: PolitiFact, Snopes, FullFact, and FactCheck.org. The selection criteria were based on credibility, data availability, and diversity in evaluation methodologies. PolitiFact and Snopes provide explicit claims and fact-check ratings, facilitating direct label extraction. FullFact, which offers detailed textual assessments without explicit labels, required additional processing via LLMs to infer verdicts. FactCheck.org presented the greatest challenge due to the lack of pre-extracted claims and evaluations, requiring the use of LLM-based techniques to produce structured outputs.

Data spans the period from 2014 to early 2025. Fig. 3 illustrates the temporal distribution of fact-checking articles published across these sources, while Figs. 4(a), 4(b), 4(c), and 4(d) show the class distribution of the fact-checking articles. Fig. 2 gives a distribution of the number of articles per source included in the dataset.

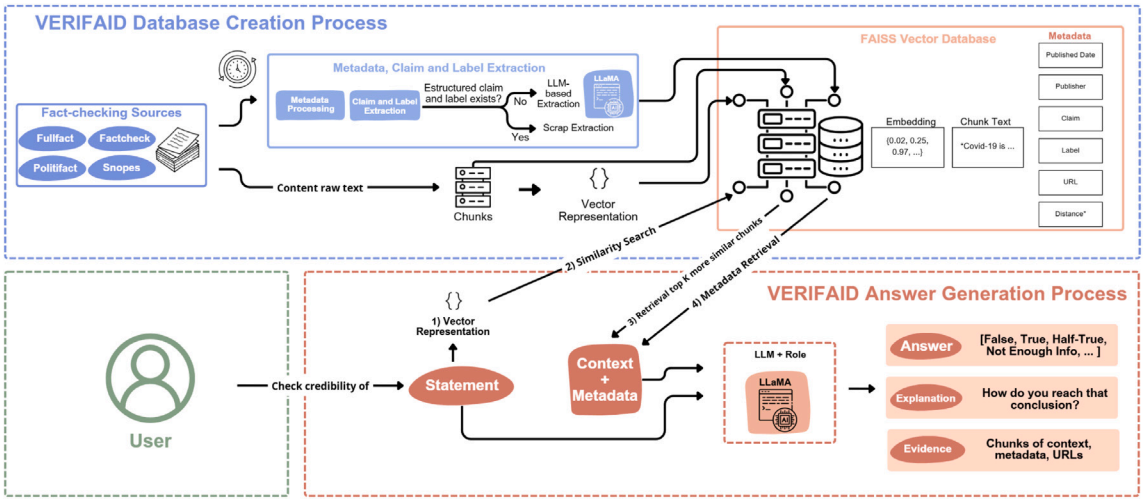


Fig. 1. The blueprint of a new Fact-Checking system using RAG. Distance is marked because it is dynamically calculated when the context is retrieved.

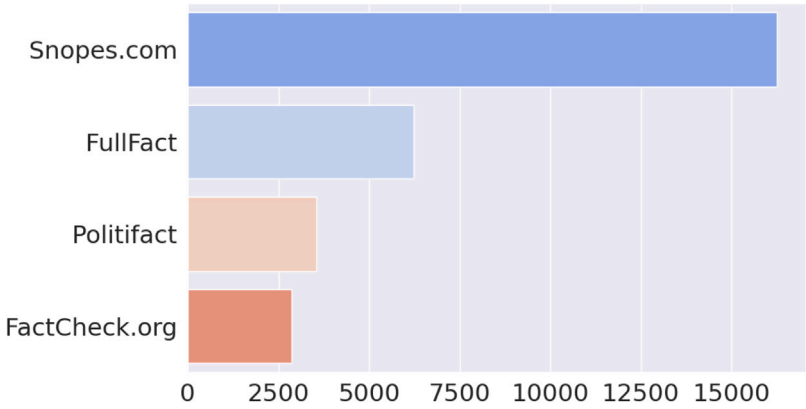


Fig. 2. Distribution of fact-checked news by publisher.

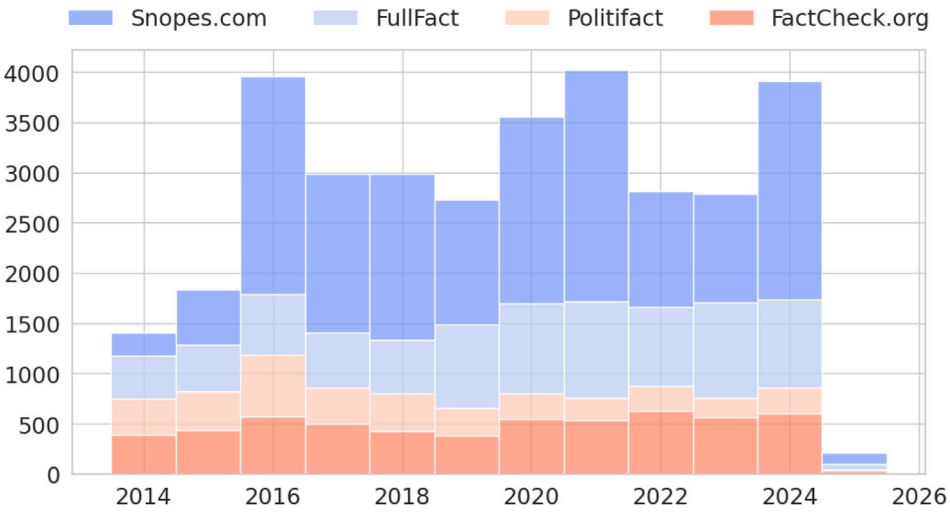
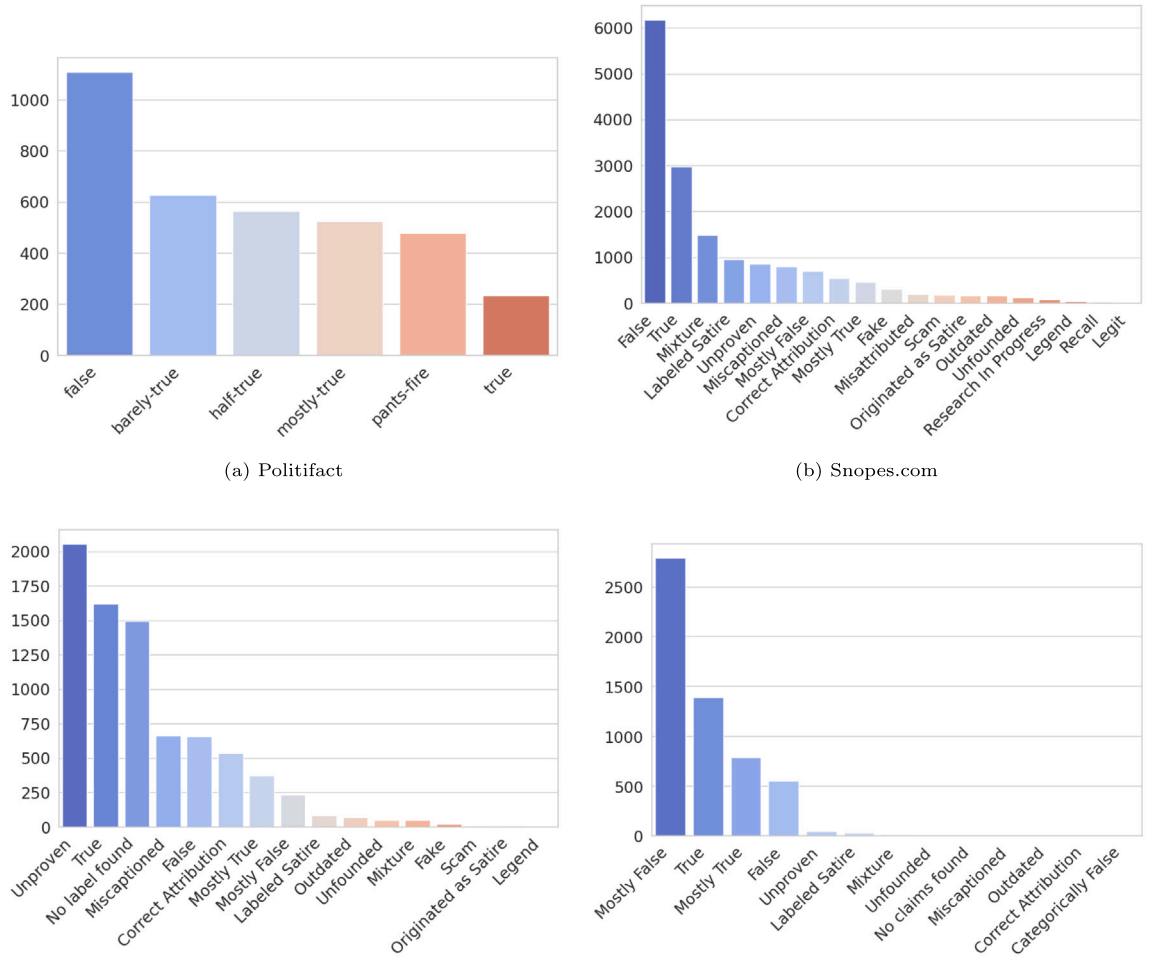


Fig. 3. Distribution of fact-checked news by publisher over time.



(a) Politifact (b) Snopes.com

(c) FullFact (d) FactCheck.org

Fig. 4. Label distribution of fact-checked news across the different sources.

3.2. Metadata, claim and label extraction

To enhance the utility of the dataset, we have included extensive metadata along with each fact-checked statement. Each entry in our database includes:

- *Publication metadata:* Publisher, publication date, and source URL.
- *Claim-related information:* The extracted claim and its associated evaluation (if present).
- *Semantic similarity metric:* The squared Euclidean (L2) distance between the fact-checked statement embedding and the input query embedding. This measure is not stored in the database as such and is calculated dynamically in the generation and inference process.

The full content of a fact-check entry is constructed by concatenating the claim, its evaluation, and the supporting textual content. This structure ensures that the system retrieves comprehensive context when assessing a statement's veracity.

To ensure the reliability and consistency of fact-checking labels, we have implemented different extraction strategies depending on the data source:

- *PolitiFact.* This source provides fact-checking articles where claims are clearly defined, and evaluations follow a predefined rating scale, including labels such as `True`, `Mostly-True`, `Half-True`, `Mostly-False`, `False`, and `Pants-on-Fire`. Since both claims and labels are presented in a consistent and identifiable way, direct extraction techniques were applied, ensuring minimal processing while preserving the integrity of the original assessments.

LLM Fact-Checking Role and Structure

```
You are a fact-checking expert trained to evaluate the truthfulness of
statements. Your task is to assess whether a statement is true or false.

Also in each chunk could be information about the original article like
published date, publisher, the original claim, credibility evaluation of
the original article by the fact-checking web and the squared Euclidean
(L2) distance from statement provided.

You must respond with a single word:
- "TRUE" if the statement is true
- "FALSE" if the statement is false.
```

Fig. 5. Prompt of the LLM fact-checker role, structure in which the information will be provided and expected output.

- *Snopes*. Similar to PolitiFact, Snopes maintains an evaluation system with explicitly stated verdicts. Snopes ratings include categories such as `False`, `True`, `Mixture`, `Labelled-Satire`, `Unproven`, `Miscaptioned`, `Mostly-False`, `Correct-Attribution`, `Mostly-True`, `Fake`, `Misattributed`, `Scam`, `Originated-as-Satire`, `Outdated`, `Unfounded`, `Research-In-Progress`, `Recall`, `Legend` and `Legit`. Given this consistency, direct label extraction was possible. However, minor preprocessing was required to normalize variations in wording across different articles, ensuring uniformity in our dataset.
- *FullFact*. Unlike PolitiFact and Snopes, FullFact does not assign explicit categorical labels to fact-checks. Instead, it provides a detailed natural language evaluation for each claim. To standardize this, we employed an LLM to process the textual evaluations and infer a label using the Politifact and Snopes rating systems as a reference. Information from both the claim and the content of the full article has been used to ensure the highest accuracy in the labels extracted. Additionally, since FullFact articles often analyse multiple claims within a single report, we stored separate entries for each identified claim and its corresponding evaluation.
- *FactCheck.org*. This source presents a challenge, as it does not explicitly highlight individual claims within its articles. Instead, fact-checking assessments are embedded in longer discussions. To address this, we utilized an LLM to extract up to two claims from each article, ensuring that each claim was logically distinct and verifiable. After claim extraction, the LLM was further employed to determine the corresponding label based on the context in which the claim was evaluated using the Politifact and Snopes rating systems as a reference. To maintain transparency, we stored a full copy of the original article alongside each extracted claim, allowing independent verification.

The process of automatic extraction of claims and labels requires special attention to maintain the quality of our database. In Section 4.4 we develop this point in depth and include several filters to ensure the integrity of the stored information.

3.3. FAISS vector database

The collected and structured fact-checking data is stored in a high-dimensional vector space to enable efficient similarity search and retrieval. The vector database chosen is FAISS (Facebook AI Similarity Search) because it provides an optimal trade-off between accuracy and retrieval speed, making it well-suited for large-scale fact-checking datasets. Also the indexing performance of the FAISS database is optimized for real-time inference, enabling fact-checking responses to be generated almost instantaneously.

For embedding creation, we employed the `jina-embeddings-v3` model. This model was chosen for two key reasons: it supports Rotary Position Embeddings, enabling it to handle long input sequences of up to 8192 tokens. This feature is especially relevant given the potentially extensive length of fact-checking articles. It offers query retrieval capabilities, making it well-suited for RAG applications. The `jina-embeddings-v3` model has proven its performance as shown in [35]. The articles were chunked into segments of 1000 characters with an overlap of 150 characters to ensure consistency during embedding generation.

3.4. Answer generation process

During the answer generation process, the goal is to verify whether a given claim is real or fake by leveraging the structured and enriched dataset. The claim to be verified is first codified into an embedding using the same embedding model utilized during

dataset creation. This embedding represents the semantic meaning of the claim, making it comparable to the embeddings of text chunks stored in the vector database.

A similarity-based retrieval is then performed, using the squared Euclidean (L2) distance in VERIFAID database to identify the most relevant chunks of information. Among the stored embeddings, the top four chunks with the highest similarity scores are selected. These chunks represent the most contextually relevant evidence in the database, offering diverse yet targeted perspectives related to the claim. The inclusion of metadata such as publication date, source of the chunk and its relevance to the claim ensures that the retrieved context is not only semantically similar but also temporally and contextually appropriate.

With these top four chunks serving as the context, the RAG framework generates a response. This response is expected to provide both a verdict indicating whether the claim is real or fake and an explanation that outlines the reasoning behind the verdict. The explanation is based on the retrieved context, bringing together the most relevant details from the evidence to form a coherent narrative.

`Llama-3.1-8B-Instruct` is used as the core fact-checking model, guided by a predefined prompt that ensures its responses rely strictly on retrieved evidence. The prompt explicitly instructs the model to generate a verdict and justification based only on this information, preventing hallucinations. For the same purpose, the temperature of the model is set to 0.01 and the generation of new tokens is limited to 256. A low temperature value ensures consistent and deterministic predictions by reducing randomness. Similar studies have used low values with the same goal [36,37]. A preliminary study to determine those values is included in Section 4.

The retrieval-augmented setup enhances the performance of the model by supporting its predictions with concrete evidence, rather than relying solely on the pre-trained knowledge of the model. This not only improves the factual accuracy of the outputs but also makes the inference process more robust to the complexities of real-world claims.

4. System validation and evaluation

In this section, we provide both the validation and evaluation of our proposed system. Additionally, we conduct a sanity check on the label extraction process to ensure the accuracy and quality of the generated labels.

The validation is performed through a case study, illustrating a real-world example that demonstrates the ability of the system to process and verify claims effectively. This case study provides a step-by-step walkthrough of how the system retrieves evidence and gives a verdict.

The evaluation, on the other hand, is conducted through a series of experiments designed to assess the robustness and effectiveness of our proposed system. Our evaluation focuses on assessing the ability of the model to accurately verify claims, the impact of different retrieval and augmentation strategies, and the overall efficiency of the retrieval-augmented framework.

4.1. Validation: A real fact-checking use case

In this section, we present a real example of how our system processes fact-checking using Retrieval-Augmented Generation. We will evaluate a statement, retrieve relevant documents, and assess its truthfulness.

Once the embedding database is loaded, we initialize the model and configure its fact-checking role. We provide it with simple yet precise instructions regarding the type of input it will receive and the expected output format. Fig. 5 displays the system prompt that yielded the best experimental results, as discussed in Section 4.3.1.

Next, we select the statement to be evaluated. For this use case, we use an example from the CoAID dataset:

> *"A Pentagon study found that people who get the flu vaccine are 36% more likely to get COVID-19".*

To process this statement, we use the same embedding model that was employed to build the database (`jina-embeddings-v3`). The statement is first converted into an embedding representation, and its squared Euclidean (L2) distance is computed against all embeddings in the database. We then retrieve the four closest chunks based on the smallest L2 distance, ranking them from closest to farthest. Additionally, we extract the metadata associated with the original document for each of these chunks.

All this information, together with the distance and the original statement, is provided in a structured way to our generation model (`Llama-3.1-8B-Instruct`). Fig. 6 shows the final prompt containing the retrieved context relevant to the given statement. Since chunks can be up to 1000 characters long, only a part is shown in the example.

The vocabulary of the model is limited to only two tokens: `TRUE` and `FALSE`. This ensures consistency when evaluating benchmark datasets and prevents the generation of unexpected words. In this case, the response of the model is `FALSE`, being correct in its evaluation.

When not working with labelled datasets, we can leverage the ability of the model to provide reasoning behind its verdict. In this case, the number of possible tokens would remain unchanged and the role would be modified as shown in Fig. 7.

Repeating the experiment with this example and with the specific role to give a justification of the evaluation gives the verdict and explanation in Fig. 8.

LLM Prompt Example with Context Retrieval

Statement to evaluate: A Pentagon study found that people who get the flu vaccine are 36% more likely to get COVID-19.

Answer using the following context:

{ Chunk 1 : { content : "In addition, while having the flu may boost the immune system, it can significantly weaken someone's overall health status, and make them more susceptible to complications should they be unfortunate enough to have both influenza and COVID-19 in the same year." ...}
{ published date : 2020-04-27 00:00:00}
{ publisher : FactCheck.org}
{ original claim : Claim 1: The flu vaccine increases someone's risk of contracting COVID-19 by 36%, Claim 2: The flu shot predisposes individuals to infection with the more severe forms of coronavirus, such as COVID-19}
{ credibility evaluation of the original article : Label 1: False, Label 2: False}
{ squared Euclidean (L2) distance from statement : 0.5692708}}

{ Chunk 2 : { content : The overall results of the study showed little to no evidence supporting the association of virus interference and influenza vaccination. Individual respiratory virus results were mixed, and some rebutted virus interference. Additionally ...}
{ published date : 2020-05-14 09:38:26}
{ publisher : Snopes.com}
{ original claim : People vaccinated for seasonal influenza face higher chances of catching SARS-COV-2, which causes COVID-19, or testing positive for the virus when they do not actually have it.}
{ credibility evaluation of the original article : False}
{ squared Euclidean (L2) distance from statement : 0.6078619}}

{ Chunk 3 : { content : There have been some results suggestive of virus interference, which the Children's Health Defense cites in its post. But none of the studies referenced assessed risks of the flu shot when it comes to COVID-19. For example ...}
{ published date : 2020-04-27 00:00:00}
{ publisher : FactCheck.org}
{ original claim : Claim 1: The flu vaccine increases someone's risk of contracting COVID-19 by 36%, Claim 2: The flu shot predisposes individuals to infection with the more severe forms of coronavirus, such as COVID-19}
{ credibility evaluation of the original article : Label 1: False, Label 2: False}
{ squared Euclidean (L2) distance from statement : 0.62524533}}
{ Chunk 4 : { content : In terms of whether the shot makes people more susceptible to other respiratory viruses, which in 2020 could include COVID-19, no strong scientific or medical evidence shows that flu shots increase people's risk of catching other viruses as of this writing. One 2012 study in the peer-reviewed medical journal Clinical Infectious Diseases, which considered ...}
{ published date : 2020-04-27 00:00:00}
{ publisher : Snopes.com}
{ original claim : People vaccinated for seasonal influenza face higher chances of catching SARS-COV-2, which causes COVID-19, or testing positive for the virus when they do not actually have it.}
{ credibility evaluation of the original article : False}
{ squared Euclidean (L2) distance from statement : 0.64493686}}

Fig. 6. Example of the context retrieved for the claim. Four chunks with metadata are shown, but only a part of the content per chunk is included in the figure.

LLM Fact-Checking Role and Structure for Reasoning

```
...
You must respond with:
- "TRUE" if the statement is true
- "FALSE" if the statement is false.

You should also provide a brief justification of the process you have
followed to reach this verdict.
```

Fig. 7. Initial prompt modified to include justification for evaluation.

LLM Verdict and Justification

```
Evaluation: FALSE
Justification: Fact-checking sources, including FactCheck.org and Snopes,
have evaluated similar claims and found them to be false. There is no strong
scientific evidence supporting the idea that the flu vaccine increases the
risk of contracting COVID-19. While some older studies suggested a potential
effect of virus interference, later research, including peer-reviewed studies
and CDC evaluations, found no conclusive link between flu vaccination and
increased susceptibility to respiratory viruses, including COVID-19. The claim
misrepresents scientific findings, likely stemming from a misinterpretation
of a Pentagon study, which did not specifically assess the relationship
between flu vaccination and COVID-19 risk. Therefore, the claim is inaccurate
and misleading.
```

Fig. 8. Example of LLM output with verdict and justification.

Table 2
Datasets used for evaluation.

Dataset	Size	Labels	Topic	Content type	Time span
FakeNewsNet-GossipCop [38]	22,478	Fake, Real	Celebrity Rumours	Text	2016+
FakeNewsNet-PolitiFact [38]	753	Fake, Real	Political News	Text	2016+
CoAID [39]	5457	Fake, Real	COVID-19 Claims and News	Text	2020+

4.2. System evaluation

The evaluation process consists of multiple experimental setups designed to measure the performance of the system under different conditions. We explore how the inclusion of retrieved fact-checking evidence, metadata, and role-specific instructions affects accuracy and reliability. By analysing variations in several performance metrics, we aim to determine the most effective configuration for claim verification.

We evaluate our approach using the datasets shown in Table 2. These include GossipCop [40] dataset, a widely recognized benchmark for celebrity rumours detection; CoAID [39] dataset, which provides health-related misinformation in the context of the COVID-19 pandemic; and PolitiFact, which provides fact-checked political news headlines. By using these datasets, we ensure that our results are consistent with real-world fact-checking scenarios.

In addition to assessing classification performance, we evaluate the efficiency of the FAISS index creation process, index sizes, and mean retrieval times of the VERIFAID database, and we compare it with Weaviate. Table 3 presents all this information for different fact-checking sources, including the combined dataset. The results were obtained using a server with the following specifications: Intel(R) Xeon(R) Silver 4210R CPU @ 2.40 GHz, 2 × NVIDIA RTX A5000 24 GB, and 4 × Samsung DDR4 32 GB.

4.3. Preliminary study

To select the hyperparameters for subsequent experiments and evaluate the generalization capabilities of our RAG framework for fake news detection, we conducted a series of preliminary experiments using a custom evaluation dataset. This dataset consists

Table 3
Creation times, mean retrieval times, and index sizes for FAISS and Weaviate across fact-checking sources.

Source	Creation time		Mean retrieval time		Index size
	FAISS	Weaviate	FAISS	Weaviate	
PolitiFact	69.90 s	94.23 s	0.19 ms	2.15 ms	24.479
Snopes	298.74 s	381.57 s	0.89 ms	4.82 ms	118.318
FullFact	78.23 s	103.97 s	0.23 ms	2.48 ms	29.861
FactCheck.org	105.23 s	133.41 s	0.27 ms	2.92 ms	35.517
Combined Sources	565.00 s	713.20 s	1.59 ms	7.22 ms	208.175

of 1000 samples, equally divided into four balanced and randomly sampled subsets (250 samples each) derived from four diverse publicly available datasets: CoAID [39], FakeVSSatire [41], TruthSeeker [42], and GossipCop [38].

Each subset was chosen to represent a different domain and linguistic style. CoAID focuses on health-related misinformation, particularly about COVID-19; FakeVSSatire includes both fake and satirical news, introducing a new challenge for the classifier; GossipCop contains celebrity-related rumour news; and TruthSeeker consists of tweets with both true and false labels, often containing informal or noisy language. The goal of combining such heterogeneous sources was to simulate real-world variability. All experiments were carried out using the full RAG framework as described in the following sections, which includes a fact-checking role, the retrieval of the top four relevant contexts, and the inclusion of metadata.

Our experiments varied multiple parameters to assess their impact on classification performance: we evaluated three instruction-tuned large language models (`Qwen2.5-7B-Instruct`, `Mistral-Small-24B-Instruct`, and `Meta-Llama-3.1-8B-Instruct`), two prompt strategies (generated vs. handcrafted), three temperature settings (0.9, 0.5, and 0.01), and two embedding models for retrieval (`all-mpnet-base-v2` and `jina-embeddings-v3`). It is worth noting that the `Mistral-Small-24B-Instruct` model was loaded using 4-bit quantization to reduce memory footprint.

As shown in Table 4, the `Meta-Llama-3.1-8B-Instruct` model consistently outperformed the other LLMs across most configurations. The best result was achieved with a handcrafted prompt, low temperature (0.01), and `jina-embeddings-v3` as the embedding model, reaching an F1 score of 0.618 and accuracy of 0.627. These results indicate that thoughtful prompt engineering, lower temperature (for more deterministic outputs), and improved embedding quality can significantly enhance RAG-based fake news classification. The relatively low performance of some configurations also highlights the sensitivity of such frameworks to prompt formulation and retrieval quality.

4.3.1. Benchmark performance

To evaluate the effectiveness of our retrieval-augmented fact-checking approach, we conducted a series of experiments on GossipCop, CoAID and Politifact datasets. Unlike traditional machine learning pipelines that require separate training, validation, and test sets, our evaluation is performed on the entire dataset, given that we employ a zero-shot classification paradigm. Our experiments explore various configurations, incorporating fact-checking roles, retrieved-context, and metadata augmentation to assess their impact on classification performance. The results, presented in Tables 5–7, provide a detailed breakdown of accuracy (ACC), precision, recall, and F1 scores under different configurations.

We begin by evaluating our approach in GossipCop, a dataset focused on rumour detection in celebrity articles (see Table 5). The baseline model, `Llama-3.1-8B-Instruct`, achieved an accuracy of 0.551 and a F1-Macro score of 0.522. Introducing a fact-checking role within the prompt led to a significant improvement, increasing accuracy to 0.639 and F1-Macro to 0.578. This enhancement suggests that role-specific prompting can effectively guide the decision-making process of the model in a fact-checking context.

However, the inclusion of retrieved fact-checking context from external sources resulted in a decrease in performance when used in isolation. Specifically, using only the top 1 retrieved-context decreased accuracy to 0.287, while expanding to the top 4 retrieved contexts resulted in a small improvement (ACC: 0.292, F1: 0.280). This finding suggests that while external evidence is useful, unstructured retrieval alone may introduce noise that confounds the model.

The most effective configuration was achieved by integrating the fact-checking role, retrieved-context (top-4) and metadata augmentation. This configuration achieved the best performance, with an accuracy of 0.676 and an F1-Macro score of 0.581. These results highlight the role of metadata in reducing the noise of retrieved content and improving the ability of the model to distinguish between real and fake news.

We also evaluated our method on CoAID, a dataset of fake and real news and claims in COVID-19 (see Table 6). The baseline model achieved an accuracy of 0.672 and an F1-Macro score of 0.620. Introducing a fact-checking role significantly improves performance, increasing accuracy to 0.746 and F1-Macro to 0.687, confirming the effectiveness of role-based prompting.

Similar to GossipCop, the inclusion of retrieved fact-checking context alone led to a decrease in performance (Top-1 Context: ACC: 0.567, F1: 0.535). However, the introduction of metadata in addition to retrieved evidence reversed this trend, with top-4 context plus metadata achieving an accuracy of 0.647 and F1-Macro score of 0.599.

The best configuration on CoAID was again the fact-checking role, top-4 context and metadata, achieving an accuracy of 0.778 and an F1-Macro score of 0.714, demonstrating the robustness of our system across different misinformation domains.

Finally we evaluated our method on Politifact, which includes a wide range of political claims and fact-checks. The baseline performance of the model was relatively strong, with an accuracy of 0.694 and F1-Macro of 0.693. As with previous datasets,

Table 4
Results of several hyperparameter combinations on evaluation dataset.

LLM	Prompt	Temp.	Embedding M.	F1	ACC
Qwen2.5-7B-Instruct	generated_prompt	0.9	all-mpnet-base-v2	0.425	0.529
Qwen2.5-7B-Instruct	generated_prompt	0.9	jina-embeddings-v3	0.458	0.546
Qwen2.5-7B-Instruct	generated_prompt	0.5	all-mpnet-base-v2	0.428	0.531
Qwen2.5-7B-Instruct	generated_prompt	0.5	jina-embeddings-v3	0.459	0.547
Qwen2.5-7B-Instruct	generated_prompt	0.01	all-mpnet-base-v2	0.431	0.534
Qwen2.5-7B-Instruct	generated_prompt	0.01	jina-embeddings-v3	0.462	0.549
Qwen2.5-7B-Instruct	handcrafted_prompt	0.9	all-mpnet-base-v2	0.482	0.560
Qwen2.5-7B-Instruct	handcrafted_prompt	0.9	jina-embeddings-v3	0.517	0.582
Qwen2.5-7B-Instruct	handcrafted_prompt	0.5	all-mpnet-base-v2	0.483	0.560
Qwen2.5-7B-Instruct	handcrafted_prompt	0.5	jina-embeddings-v3	0.516	0.582
Qwen2.5-7B-Instruct	handcrafted_prompt	0.01	all-mpnet-base-v2	0.484	0.562
Qwen2.5-7B-Instruct	handcrafted_prompt	0.01	jina-embeddings-v3	0.518	0.585
Mistral-Small-24B-Instruct	generated_prompt	0.9	all-mpnet-base-v2	0.372	0.509
Mistral-Small-24B-Instruct	generated_prompt	0.9	jina-embeddings-v3	0.380	0.513
Mistral-Small-24B-Instruct	generated_prompt	0.5	all-mpnet-base-v2	0.356	0.501
Mistral-Small-24B-Instruct	generated_prompt	0.5	jina-embeddings-v3	0.367	0.509
Mistral-Small-24B-Instruct	generated_prompt	0.01	all-mpnet-base-v2	0.352	0.501
Mistral-Small-24B-Instruct	generated_prompt	0.01	jina-embeddings-v3	0.356	0.504
Mistral-Small-24B-Instruct	handcrafted_prompt	0.9	all-mpnet-base-v2	0.435	0.523
Mistral-Small-24B-Instruct	handcrafted_prompt	0.9	jina-embeddings-v3	0.452	0.535
Mistral-Small-24B-Instruct	handcrafted_prompt	0.5	all-mpnet-base-v2	0.420	0.519
Mistral-Small-24B-Instruct	handcrafted_prompt	0.5	jina-embeddings-v3	0.418	0.521
Mistral-Small-24B-Instruct	handcrafted_prompt	0.01	all-mpnet-base-v2	0.414	0.521
Mistral-Small-24B-Instruct	handcrafted_prompt	0.01	jina-embeddings-v3	0.411	0.520
Meta-Llama-3.1-8B-Instruct	generated_prompt	0.9	all-mpnet-base-v2	0.538	0.572
Meta-Llama-3.1-8B-Instruct	generated_prompt	0.9	jina-embeddings-v3	0.559	0.585
Meta-Llama-3.1-8B-Instruct	generated_prompt	0.5	all-mpnet-base-v2	0.536	0.569
Meta-Llama-3.1-8B-Instruct	generated_prompt	0.5	jina-embeddings-v3	0.563	0.593
Meta-Llama-3.1-8B-Instruct	generated_prompt	0.01	all-mpnet-base-v2	0.536	0.571
Meta-Llama-3.1-8B-Instruct	generated_prompt	0.01	jina-embeddings-v3	0.569	0.596
Meta-Llama-3.1-8B-Instruct	handcrafted_prompt	0.9	all-mpnet-base-v2	0.596	0.606
Meta-Llama-3.1-8B-Instruct	handcrafted_prompt	0.9	jina-embeddings-v3	0.605	0.616
Meta-Llama-3.1-8B-Instruct	handcrafted_prompt	0.5	all-mpnet-base-v2	0.613	0.624
Meta-Llama-3.1-8B-Instruct	handcrafted_prompt	0.5	jina-embeddings-v3	0.609	0.619
Meta-Llama-3.1-8B-Instruct	handcrafted_prompt	0.01	all-mpnet-base-v2	0.612	0.621
Meta-Llama-3.1-8B-Instruct	handcrafted_prompt	0.01	jina-embeddings-v3	**0.618**	**0.627**

Table 5
Evaluation metrics for different experiments on GossipCop dataset.

GossipCop experimentation	ACC	Prec.	Recall	F1-Macro	F1-Micro	F1-Real	F1-Fake
Llama-3.1-8B-Instruct Base	0.551	0.566	0.594	0.522	0.551	0.638	0.406
+ Fact-checking Role	0.639	**0.584**	**0.614**	0.578	0.639	0.738	**0.418**
+ Top 1 Context	0.287	0.517	0.508	0.274	0.287	0.178	0.370
+ Top 4 Context	0.292	0.520	0.511	0.280	0.292	0.190	0.371
+ Top 1 Context + Metadata	0.481	0.558	0.576	0.471	0.481	0.543	0.398
+ Top 4 Context + Metadata	0.508	0.561	0.584	0.492	0.508	0.582	0.402
+ Fact-checking Role + Top 1 Context	0.359	0.516	0.516	0.358	0.359	0.352	0.365
+ Fact-checking Role + Top 4 Context	0.420	0.535	0.542	0.417	0.420	0.458	0.376
+ Fact-checking Role + Top 1 Context + Metadata	0.385	0.528	0.530	0.385	0.385	0.397	0.372
+ Fact-checking Role + Top 4 Context + Metadata	**0.676**	0.578	0.591	**0.581**	**0.676**	**0.781**	0.381

introducing a fact-checking role improved all metrics, reaching an accuracy of 0.716 and F1-Macro of 0.716. The addition of external context alone reduced performance, but when combined with metadata and the fact-checking role, the results improved again. The best configuration overall used only the fact-checking role, but the combination of role, top-4 context, and metadata was a close second, achieving an accuracy of 0.709 and F1-Macro of 0.709.

After manually evaluating 50 randomly selected error cases, we observed that the performance degradation when using retrieved contexts alone arises because the chunks are often noisy or incomplete, which can mislead the model if no additional guidance is provided. The base model or the role-prompted version performs better in these cases because they rely on internally consistent reasoning patterns, while the raw retrieved contexts may introduce conflicting signals.

The combination of role prompting, top-4 retrieved contexts, and metadata achieves more stable and reliable results, even if the performance metrics are close to those of the role-only setup. This configuration improves interpretability by providing traceable evidence, increases robustness by leveraging up-to-date external information, and reduces the risk of hallucinations. In real-world

Table 6
Evaluation metrics for different experiments on CoAID dataset.

CoAID experimentation	ACC	Prec.	Recall	F1-Macro	F1-Micro	F1-Real	F1-fake
Llama-3.1-8B-Instruct Base	0.672	0.646	0.758	0.620	0.672	0.760	0.479
+ Fact-checking Role	0.746	0.684	0.812	0.687	0.746	0.824	0.549
+ Top 1 Context	0.567	0.614	0.699	0.535	0.567	0.657	0.413
+ Top 4 Context	0.547	0.616	0.699	0.521	0.547	0.632	0.410
+ Top 1 Context + Metadata	0.594	0.619	0.710	0.555	0.594	0.686	0.425
+ Top 4 Context + Metadata	0.647	0.637	0.743	0.599	0.647	0.737	0.460
+ Fact-checking Role + Top 1 Context	0.681	0.653	0.769	0.629	0.681	0.768	0.490
+ Fact-checking Role + Top 4 Context	0.679	0.656	0.775	0.629	0.679	0.765	0.493
+ Fact-checking Role + Top 1 Context + Metadata	0.426	0.596	0.635	0.420	0.426	0.480	0.360
+ Fact-checking Role + Top 4 Context + Metadata	**0.778**	**0.700**	**0.826**	**0.714**	**0.778**	**0.850**	**0.579**

Table 7
Evaluation metrics for different experiments on the Politifact dataset.

Politifact experimentation	ACC	Prec.	Recall	F1-Macro	F1-Micro	F1-Real	F1-Fake
Llama-3.1-8B-Instruct Base	0.694	0.706	0.700	0.693	0.694	0.674	0.712
+ Fact-checking Role	**0.716**	**0.716**	**0.717**	**0.716**	**0.716**	**0.725**	0.706
+ Top 1 Context	0.564	0.653	0.586	0.521	0.564	0.376	0.665
+ Top 4 Context	0.592	0.681	0.612	0.558	0.592	0.435	0.681
+ Top 1 Context + Metadata	0.602	0.637	0.615	0.590	0.602	0.519	0.660
+ Top 4 Context + Metadata	0.632	0.655	0.641	0.626	0.632	0.581	0.671
+ Fact-checking Role + Top 1 Context	0.620	0.656	0.633	0.610	0.620	0.546	0.674
+ Fact-checking Role + Top 4 Context	0.688	0.714	0.697	0.683	0.688	0.648	**0.719**
+ Fact-checking Role + Top 1 Context + Metadata	0.608	0.667	0.625	0.588	0.608	0.496	0.680
+ Fact-checking Role + Top 4 Context + Metadata	0.709	0.708	0.709	0.709	0.709	0.722	0.695

Table 8
Performance comparison of models across datasets. Accuracy (ACC) and F1-Macro score have been taken directly from the original paper.

Dataset	Model	Zero-shot	ACC	F1-Macro
GossipCop	DetectYSF. [43] (few-16)	✗	0.663	–
	P&A [31] (few-16)	✗	0.612	–
	DAFND [44] (few-8)	✗	**0.828**	0.535
	Cross-Modal A. [43] (few-2)	✗	0.719	–
	LLaMa-ZS [45] (zero-shot)	✓	–	0.550
	Our Approach	✓	0.676	**0.581**
CoAID	COOL [32] (few-16)	✗	**0.793**	0.590
	KPL [32] (few-2)	✗	0.549	0.338
	RPL [32] (few-2)	✗	0.543	0.417
	COOL [32] (few-2)	✗	0.602	0.447
	Our Approach	✓	0.778	**0.714**
Politifact	KPL [46] (few-2)	✗	–	0.537
	LLaMa-ZS [45] (zero-shot)	✓	–	0.610
	Our Approach	✓	**0.709**	**0.709**

applications, these qualities are more valuable than marginal improvements in accuracy, making this combination the most reliable choice despite the similar numbers.

To contextualize our results, we compare our approach with state-of-the-art models from the literature. Table 8 shows the performance of our retrieval-augmented fact-checking model in a zero-shot setting, alongside several few-shot and supervised baselines. On GossipCop, our approach can compete with many of the methods in the literature as it has the highest F1-Macro score and is comparable to other metrics. By far the highest accuracy is achieved by DAFND, but in fake news detection problems, it can be seen in the literature that a higher degree of importance is given to metrics such as the F1 score, where both classes are taken into account. For this metric, our approach even outperforms this method.

In addition to the results shown in Table 8 we can also highlight the studies [46,47] where they give the F1 score of the `Fake` class for the GossipCop dataset, these are 0.422 for two-shot in [47] and 0.378 for two-shot in [46]. As can be seen in Table 5, our F1 score in the `Fake` class (0.381) is competitive, even surpassing the 0.378 of [46].

In CoAID, our approach achieves the state of the art in F1 score, competing with few-shot models with 2 and 16 train samples. Furthermore, in the original study where the dataset is presented, several baselines used as benchmarks are presented, and the highest F1 score obtained in those methods is 0.581, which is far behind the one provided by our approach.

On the Politifact dataset, although the configuration including all variables (fact-checking role, top-4 retrieved context, and metadata) was only the second best in our internal experiments, it still surpasses state-of-the-art models of a similar nature, such as KPL and LLaMa-ZS. Our approach achieves an accuracy of 0.709 and an F1-Macro score of 0.709, outperforming KPL (F1: 0.537) and LLaMa-ZS (F1: 0.610). From our experimentation, we derive the following key insights:

Sanity Check Prompt

```
You are an expert evaluator for LLM-based label extraction. Given a claim,
its extracted label and the original text, you will judge the quality of
the label extraction. Evaluate the quality of the label extraction. Choose
a number from 1 to 5 where:

1 = Very poor, 2 = Poor, 3 = Fair, 4 = Good, 5 = Excellent

Respond only with the number.

Original text: {...} Claim: {...} Extracted Label: {...}  Judgment:
```

Fig. 9. Sanity check prompt used for evaluating label extraction quality.

1. Introducing role-specific prompting consistently improved classification accuracy and F1 scores across both datasets.
2. Including more chunks of context alone can introduce noise, but when combined with metadata, it enhances the ability of the model to prioritize relevant information.
3. The best results were obtained when retrieved evidence was augmented with metadata.
4. Our method outperformed multiple few-shot and some fully supervised models without requiring additional labelled training data, making it highly scalable for real-world misinformation detection.

We recognize that contemporary generative AI systems often operate at the intersection of ethical ambiguity and practical risk, frequently lacking adequate oversight or interpretability. To address this, our system is explicitly created as a decision-support mechanism, rather than an autonomous decision-maker.

While the LLM produces a verdict, the interface clearly communicates that this may not always be reliable. In order to promote transparency and user verification, all retrieved context chunks associated with each claim are provided, along with the URLs of their sources, allowing the user to review the evidence. Additionally, any labels extracted automatically by an LLM from fact-checking websites are clearly marked with a tag so users can critically assess their trustworthiness.

It is important to note that the system is designed to work alongside a human fact-checker. The fact-checker uses the system to accelerate the evaluation process while maintaining authoritative control over the final judgment. The human operator remains entirely responsible for deeper contextual exploration and final verification. This hybrid model upholds core ethical principles, particularly in important areas such as political discourse and public health.

4.4. Sanity check for label extraction

In this section, we perform a sanity check to evaluate the quality of label extraction from fact-checking sources such as FactCheck.org and FullFact. Label extraction evaluation was performed using the `Mistral-Small-24B-Instruct` model, as it was necessary to use a distinct model for generation and evaluation. This ensures that the model responsible for generating the labels is different from the one assessing their quality. The quality of these labels was assessed on a scale from 1 to 5. The evaluation focuses on determining how accurately the extracted labels reflect the truthfulness of the claims in the context of the original articles. The evaluation was carried out based on the prompt shown in Fig. 9.

The quality ratings for the extracted labels are summarized in Fig. 10 (left). These results show that the majority of the labels (84.1%) were evaluated as either `Good` or `Excellent`. This indicates that the label extraction process performed well overall, with the model successfully identifying the truthfulness of most claims. However, a small proportion of labels (11%) received a lower score of `Poor`, or `Very Poor`. This subset of labels suggests areas where the extraction process might be refined, particularly in handling complex or ambiguous claims.

In Fig. 10 (right), the distributions of ratings by publisher are shown. It can be observed that for FactCheck.org, where claims were also extracted with an LLM, there are very few labels rated as `Very Poor`, `Poor`, or `Fair`, with almost all labels being rated as `Good` or `Excellent`. In contrast, FullFact, where the claims and ratings were written in natural language by the authors, shows a higher proportion of `Very Poor`, `Poor`, and `Fair` ratings.

To further address the subset of low-quality labels, we add an additional refinement step. These labels were reprocessed using a different instruction-tuned LLM, `Qwen2.5-7B-Instruct`, with an updated prompt explicitly stating that the claims in question were particularly ambiguous or challenging.

After re-extraction, we re-evaluated the quality of the new labels using the same Mistral-based evaluation framework described earlier. As shown in Fig. 11 (updated version), this refinement step substantially reduced the proportion of Poor and Very Poor ratings from 11% to just 1.3%. This suggests that much of the initially low-quality output could be attributed to model uncertainty in borderline cases, which can be mitigated by targeted reprocessing with explicit instructions and model diversity.

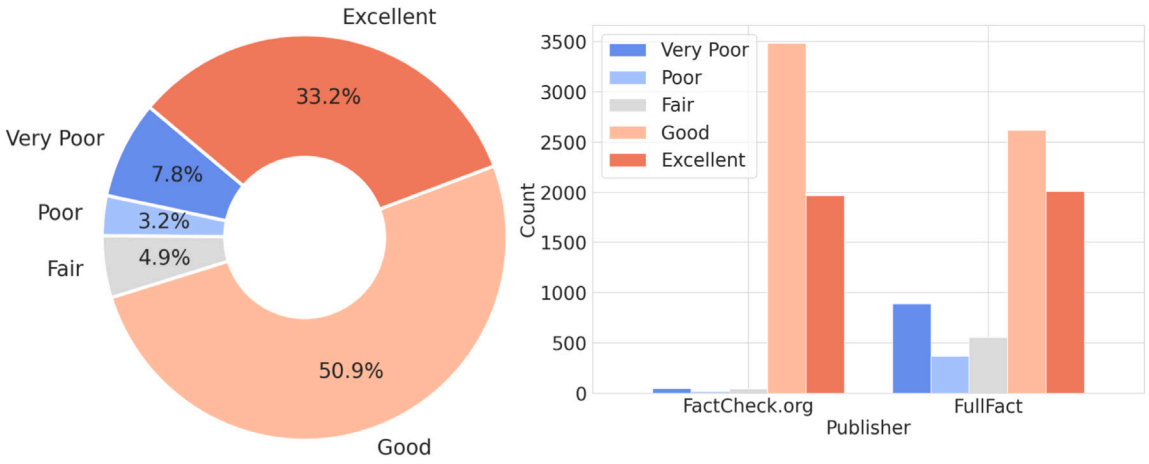


Fig. 10. Quality evaluation of label extraction using Mistral from FactCheck.org and FullFact (left) and evaluation by publisher (right).

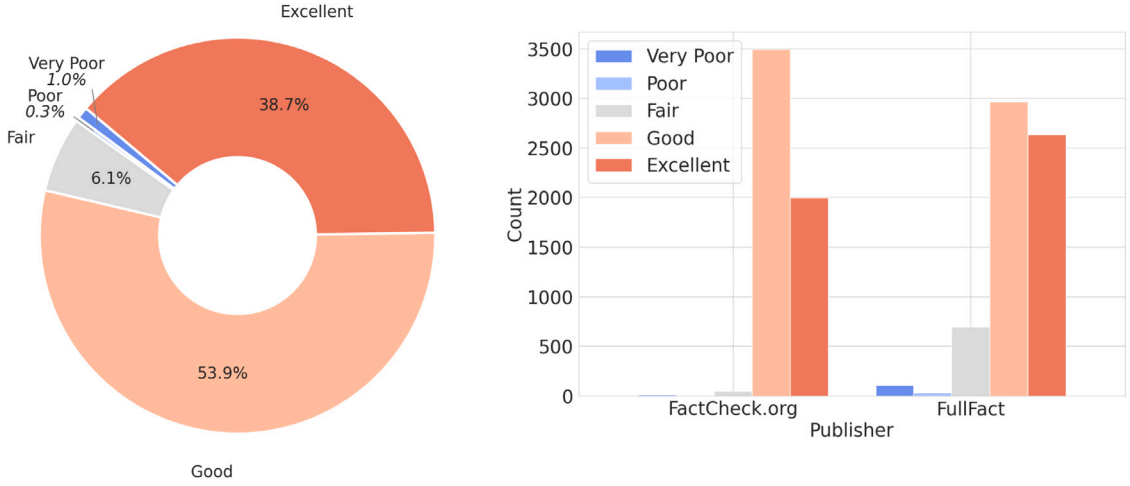


Fig. 11. Quality evaluation of label extraction using Mistral from FactCheck.org and FullFact (left) and evaluation by publisher (right). Low-quality labels were reprocessed with Qwen and re-evaluated.

The remaining 1.3% of labels that continue to receive low ratings after the second pass can either be manually verified by human annotators or excluded from the final dataset, in order to maintain high-quality standards. In addition, as we said in the previous section, all labels automatically generated by an LLM are explicitly flagged with a tag that is provided with the chunks of context. Fig. 12 shows the complete flow of this process, which is integrated into our system. This diagram summarizes all the stages involved in claim and label extraction. The modular design also facilitates future extensions or changes to the pipeline, such as the use of different LLMs for multilingual processing, the addition of more relevant metadata, or extra steps for quality extraction checks.

5. Conclusion and future work

In this paper, we proposed and validated a novel framework for fact-checking, integrating a RAG system enriched with a scalable dataset construction methodology. Our approach not only serves as a blueprint for building interpretable fact-checking systems but also demonstrates how retrieved knowledge can enhance classification and analysis in zero-shot scenarios. By leveraging FAISS-based retrieval and language models, our system efficiently gathers and processes external knowledge, enabling accurate classification of unseen examples while also providing interpretative clues and contextual insights to support fact-checking efforts. Through extensive evaluation on widely-used datasets, we have shown that our approach is competitive compared with traditional models trained on these datasets, proving its effectiveness in tackling misinformation.

As part of our future work, we plan to integrate the proposed system into an online platform that will be continuously updated with the latest fact-checking data. This will enable real-time misinformation detection and ensure the system remains adaptable to

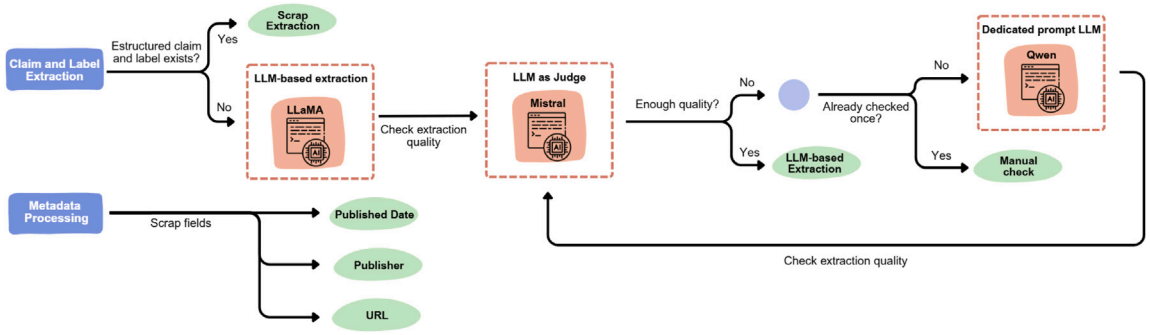


Fig. 12. Claim and label extraction process flowchart.

new narratives. Additionally, we aim to explore advanced techniques for improving the relevance of retrieval, search efficiency, and search accuracy. Furthermore, we will explore techniques to improve the contextualization of embeddings, ensuring that retrieved information is not only relevant but also accurately framed within the broader context of the claim being evaluated.

Regarding evaluation, we consider it essential to continue incorporating new datasets, particularly those that contain challenging content such as satire or informal language. Expanding the diversity of evaluated material allows for a more comprehensive assessment of system performance under varied linguistic and contextual conditions. In line with this, an important direction for future work is the integration of multilingual datasets, which would enable the development of a more robust fact-checking system capable of operating effectively across multiple languages.

Finally, we plan to explore how content-level features such as sentiment, toxicity, and public engagement dynamics can influence claim interpretation, with the goal of leveraging them to improve our system.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The research reported in this paper was supported by the DesinfoScan project: Grant TED2021-129402B-C21 funded by MICIU/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR, and FederaMed project: Grant PID2021-123960OB-I00 funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU. Finally, the research reported in this paper is also funded by the European Union (BAG-INTEL project, grant agreement no. 101121309).

Appendix. Supplementary material

The code and dataset associated with the experiments in this article will be available in a GitHub repository upon request to the corresponding author after acceptance of the article.

Data availability

Data will be made available on request.

References

[1] Li B, Bonk CJ, Wang C, Kou X. Reconceptualizing self-directed learning in the era of generative AI: An exploratory analysis of language learning. IEEE Trans Learn Technol 2024. http://dx.doi.org/10.1109/TLT.2024.3386098.

[2] Ferraro C, Demsar V, Sands S, Restrepo M, Campbell C. The paradoxes of generative AI-enabled customer service: A guide for managers. Bus Horiz 2024. http://dx.doi.org/10.1016/j.bushor.2024.04.013, URL https://www.sciencedirect.com/science/article/pii/S0007681324000582.

[3] Chu-Ke C, Dong Y. Misinformation and literacies in the era of generative artificial intelligence: A brief overview and a call for future research. Emerg Media 2024;2(1):70–85. http://dx.doi.org/10.1177/27523543241240285.

[4] Ponzio F, Urgese G, Ficarra E, Di Cataldo S. Dealing with lack of training data for convolutional neural networks: The case of digital pathology. Electronics 2019;8(3):256. http://dx.doi.org/10.3390/electronics8030256, URL https://www.mdpi.com/2079-9292/8/3/256.

[5] Baashirah R. Zero-shot automated detection of fake news: An innovative approach (ZS-FND). IEEE Access 2024;12. http://dx.doi.org/10.1109/ACCESS.2024.3462151.

[6] Lin H, Yi P, Ma J, Jiang H, Luo Z, Shi S, Liu R. Zero-shot rumor detection with propagation structure via prompt learning. In: Proceedings of the 37th AAAI conference on artificial intelligence, vol. 37, (no. 4):2023, p. 5213–21. http://dx.doi.org/10.1609/aaai.v37i4.2565.

[7] Gao W, Ni M, Deng H, Zhu X, Zeng P, Hu X. Few-shot fake news detection via prompt-based tuning. J Intell Fuzzy Systems 2023;44(6):9933–42. http://dx.doi.org/10.3233/JIFS-221647.

[8] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih W-t, Rocktäschel T, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. Adv Neural Inf Process Syst 2020;33:9459–74.

[9] Xu Z, Cruz MJ, Guevara M, Wang T, Deshpande M, Wang X, Li Z. Retrieval-augmented generation with knowledge graphs for customer service question answering. In: Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval. 2024, p. 2905–9.

[10] Novotnỳ V, Ayetiran EF, Štefánik M, Sojka P. Text classification with word embedding regularization and soft similarity measure. 2020, arXiv preprint arXiv:2003.05019.

[11] Kumar M, Bhatia R, Rattan D. A survey of web crawlers for information retrieval. Wiley Interdiscip Rev: Data Min Knowl. Discov 2017;7(6). http://dx.doi.org/10.1002/widm.1218.

[12] Venzke A, Molzahn DK, Chatzivasileiadis S. Efficient creation of datasets for data-driven power system applications. Electr Power Syst Res 2021;190:106614. http://dx.doi.org/10.1016/j.epsr.2020.106614, URL https://www.sciencedirect.com/science/article/pii/S0378779620304181.

[13] Tkaczyk D, Szostek P, Bolikowski L. GROTOAP2-the methodology of creating a large ground truth dataset of scientific articles. D-Lib Mag 2014;20(11/12).

[14] Drchal J, Ullrich H, Mlynář T, Moravec V. Pipeline and dataset generation for automated fact-checking in almost any language. Neural Comput Appl 2024;36(30):19023–54. http://dx.doi.org/10.1007/s00521-024-10113-5.

[15] Thorne J, Vlachos A, Christodoulopoulos C, Mittal A. FEVER: a large-scale dataset for fact extraction and VERification. In: Walker M, Ji H, Stent A, editors. Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long papers). New Orleans, Louisiana: Association for Computational Linguistics; 2018, p. 809–19. http://dx.doi.org/10.18653/v1/N18-1074, URL https://aclanthology.org/N18-1074/.

[16] Aly R, Guo Z, Schlichtkrull MS, Thorne J, Vlachos A, Christodoulopoulos C, Cocarascu O, Mittal A. The fact extraction and VERification over unstructured and structured information (FEVEROUS) shared task. In: Aly R, Christodoulopoulos C, Cocarascu O, Guo Z, Mittal A, Schlichtkrull M, Thorne J, Vlachos A, editors. Proceedings of the fourth workshop on fact extraction and verification. Dominican Republic: Association for Computational Linguistics; 2021, p. 1–13. http://dx.doi.org/10.18653/v1/2021.fever-1.1, URL https://aclanthology.org/2021.fever-1.1/.

[17] Diggelmann T, Boyd-Graber J, Bulian J, Ciaramita M, Leippold M. Climate-fever: A dataset for verification of real-world climate claims. 2020, arXiv preprint arXiv:2012.00614.

[18] Schlichtkrull MS, Guo Z, Vlachos A. AVeriTeC: A dataset for real-world claim verification with evidence from the web. In: Thirty-seventh conference on neural information processing systems datasets and benchmarks track. 2023, URL https://openreview.net/forum?id=fKzSz0oyaI.

[19] Khan K, Wang R, Poupart P. WatClaimCheck: A new dataset for claim entailment and inference. In: Muresan S, Nakov P, Villavicencio A, editors. Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers). Dublin, Ireland: Association for Computational Linguistics; 2022, p. 1293–304. http://dx.doi.org/10.18653/v1/2022.acl-long.92, URL https://aclanthology.org/2022.acl-long.92/.

[20] Kotonya N, Toni F. Explainable automated fact-checking for public health claims. In: Webber B, Cohn T, He Y, Liu Y, editors. Proceedings of the 2020 conference on empirical methods in natural language processing. Online: Association for Computational Linguistics; 2020, p. 7740–54. http://dx.doi.org/10.18653/v1/2020.emnlp-main.623, URL https://aclanthology.org/2020.emnlp-main.623/.

[21] Wadden D, Lin S, Lo K, Wang LL, van Zuylen M, Cohan A, Hajishirzi H. Fact or fiction: Verifying scientific claims. In: Webber B, Cohn T, He Y, Liu Y, editors. Proceedings of the 2020 conference on empirical methods in natural language processing. Online: Association for Computational Linguistics; 2020, p. 7534–50. http://dx.doi.org/10.18653/v1/2020.emnlp-main.609, URL https://aclanthology.org/2020.emnlp-main.609/.

[22] Nakamura K, Levy S, Wang WY. Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In: Calzolari N, Béchet F, Blache P, Choukri K, Cieri C, Declerck T, Goggi S, Isahara H, Maegaard B, Mariani J, Mazo H, Moreno A, Odijk J, Piperidis S, editors. Proceedings of the twelfth language resources and evaluation conference. Marseille, France: European Language Resources Association; 2020, p. 6149–57, URL https://aclanthology.org/2020.lrec-1.755/.

[23] Zubiaga A, Liakata M, Procter R, Wong Sak Hoi G, Tolmie P. Analysing how people orient to and spread rumours in social media by looking at conversational threads. PLOS ONE 2016;11(3):1–29. http://dx.doi.org/10.1371/journal.pone.0150989.

[24] Gupta A, Srikumar V. X-Fact: A new benchmark dataset for multilingual fact checking. In: Zong C, Xia F, Li W, Navigli R, editors. Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 2: short papers). Online: Association for Computational Linguistics; 2021, p. 675–82. http://dx.doi.org/10.18653/v1/2021.acl-short.86, URL https://aclanthology.org/2021.acl-short.86/.

[25] Wang WY. "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). Vancouver, Canada: Association for Computational Linguistics; 2017, p. 422–6. http://dx.doi.org/10.18653/v1/P17-2067, URL https://aclanthology.org/P17-2067.

[26] Liu H, Wang W, Li H, Li H. TELLER: A trustworthy framework for explainable, generalizable and controllable fake news detection. In: Ku L-W, Martins A, Srikumar V, editors. Findings of the association for computational linguistics: ACL 2024. Bangkok, Thailand: Association for Computational Linguistics; 2024, p. 15556–83. http://dx.doi.org/10.18653/v1/2024.findings-acl.919, URL https://aclanthology.org/2024.findings-acl.919.

[27] Zeng X, Zubiaga A. MAPLE: Micro analysis of pairwise language evolution for few-shot claim verification. In: Graham Y, Purver M, editors. Findings of the association for computational linguistics: EACL 2024. St. Julian's, Malta: Association for Computational Linguistics; 2024, p. 1177–96, URL https://aclanthology.org/2024.findings-eacl.79/.

[28] Niu C, Guan Y, Wu Y, Zhu J, Song J, Zhong R, Zhu K, Xu S, Diao S, Zhang T. VeraCT scan: Retrieval-augmented fake news detection with justifiable reasoning. In: Cao Y, Feng Y, Xiong D, editors. Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 3: system demonstrations). Bangkok, Thailand: Association for Computational Linguistics; 2024, p. 266–77. http://dx.doi.org/10.18653/v1/2024.acl-demos.25, URL https://aclanthology.org/2024.acl-demos.25.

[29] Singal R, Patwa P, Patwa P, Chadha A, Das A. Evidence-backed fact checking using RAG and few-shot in-context learning with LLMs. In: Schlichtkrull M, Chen Y, Whitehouse C, Deng Z, Akhtar M, Aly R, Guo Z, Christodoulopoulos C, Cocarascu O, Mittal A, Thorne J, Vlachos A, editors. Proceedings of the seventh fact extraction and vERification workshop. Miami, Florida, USA: Association for Computational Linguistics; 2024, p. 91–8. http://dx.doi.org/10.18653/v1/2024.fever-1.10, URL https://aclanthology.org/2024.fever-1.10/.

[30] Ma J, Chen C, Hou C, Yuan X. KAPALM: Knowledge grAPh enhanced language models for fake news detection. In: Bouamor H, Pino J, Bali K, editors. Findings of the association for computational linguistics: EMNLP 2023. Singapore: Association for Computational Linguistics; 2023, p. 3999–4009. http://dx.doi.org/10.18653/v1/2023.findings-emnlp.263, URL https://aclanthology.org/2023.findings-emnlp.263.

[31] Wu J, Li S, Deng A, Xiong M, Hooi B. Prompt-and-Align: prompt-based social alignment for few-shot fake news detection. In: Proceedings of the 32nd ACM international conference on information and knowledge management. 2023, p. 2726–36. http://dx.doi.org/10.1145/3583780.3615015.

[32] Ouyang Y, Wu P, Pan L. COOL: Comprehensive knowledge enhanced prompt learning for domain adaptive few-shot fake news detection. 2024, arXiv preprint arXiv:2406.10870.

[33] Temiz O, Kılıç ÖO, Kızıldağ AO, Temizel TT. A fact checking and verification system for FEVEROUS using a zero-shot learning approach. In: Proceedings of the fourth workshop on fact extraction and vERification. 2021, p. 113–20. http://dx.doi.org/10.18653/v1/2021.fever-1.13, URL https://aclanthology.org/2021.fever-1.13/.

[34] Li G, Lu W, Zhang W, Lian D, Lu K, Mao R, Shu K, Liao H. Re-search for the truth: Multi-round retrieval-augmented large language models are strong fake news detectors. 2024, arXiv preprint arXiv:2403.09747.

[35] Sturua S, Mohr I, Akram MK, Günther M, Wang B, Krimmel M, Wang F, Mastrapas G, Koukounas A, Wang N, et al. Jina-embeddings-v3: Multilingual embeddings with task Lora. 2024, arXiv preprint arXiv:2409.10173.
[36] Bangerter ML, Fenza G, Furno D, Gallo M, Loia V, Stanzione C, You I. A hybrid framework integrating LLM and ANFIS for explainable fact-checking. IEEE Trans Fuzzy Syst 2024;PP:1–11. http://dx.doi.org/10.1109/TFUZZ.2024.3431710.
[37] Choi EC, Ferrara E. FACT-GPT: Fact-checking augmentation via claim matching with LLMs. In: Companion proceedings of the ACM web conference 2024. 2024, p. 883–6. http://dx.doi.org/10.1145/3589335.3651504.
[38] Shu K, Zheng G, Li Y, Mukherjee S, Awadallah AH, Ruston S, Liu H. Leveraging multi-source weak social supervision for early detection of fake news. 2020, arXiv preprint arXiv:2004.01732.
[39] Cui L, Lee D. Coaid: Covid-19 healthcare misinformation dataset. 2020, arXiv preprint arXiv:2006.00885.
[40] Shu K, Mahudeswaran D, Wang S, Lee D, Liu H. FakeNewsNet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. Big Data 2020;8(3):171–88.
[41] Golbeck J, Mauriello M, Auxier B, Bhanushali KH, Bonk C, Bouzaghrane MA, Buntain C, Chanduka R, Cheakalos P, Everett JB, et al. Fake news vs satire: A dataset and analysis. In: Proceedings of the 10th ACM conference on web science. 2018, p. 17–21.
[42] Dadkhah S, Zhang X, Weismann AG, Firouzi A, Ghorbani AA. The largest social media ground-truth dataset for real/fake content: Truthseeker. IEEE Trans Comput Soc Syst 2023;11(3):3376–90.
[43] Jiang Y, Wang T, Xu X, Wang Y, Song X, Maynard D. Cross-modal augmentation for few-shot multimodal fake news detection. Eng Appl Artif Intell 2025;142:109931.
[44] Liu Y, Zhu J, Zhang K, Tang H, Zhang Y, Liu X, Liu Q, Chen E. Detect, investigate, judge and determine: A novel LLM-based framework for few-shot fake news detection. 2024, arXiv preprint arXiv:2407.08952.
[45] Leite JA, Razuvayevskaya O, Bontcheva K, Scarton C. Weakly supervised veracity classification with LLM-predicted credibility signals. EPJ Data Sci 2025;14(1):16.
[46] Jiang G, Liu S, Zhao Y, Sun Y, Zhang M. Fake news detection via knowledgeable prompt learning. Inf Process Manage 2022;59(5):103029.
[47] Ma J, Chen C, Hou C, Yuan X. Kapalm: Knowledge graph enhanced language models for fake news detection. In: Findings of the association for computational linguistics: EMNLP 2023. 2023, p. 3999–4009.