**IAES International Journal of Artificial Intelligence (IJ-AI)**
Vol. 15, No. 4, August 2026, pp. 3144~3153
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i4.pp3144-3153 

# Social news factuality verification using large language models

**Tran Duc Duong[1], Hai Hoan Do[2]**
[1]Faculty of Information Technology, Posts and Telecommunications Institute of Technology, Hanoi, Vietnam
[2]Economic Research Institute of Postal Technology, Posts and Telecommunications Institute of Technology, Hanoi, Vietnam

**Article Info**

***Article history:***

Received Jul 5, 2025
Revised May 13, 2026
Accepted Jun 19, 2026

***Keywords:***

Claim extraction
Claim verification
Large language models
News clustering
Social news verification

**ABSTRACT**

Social media platforms have greatly accelerated the spread of news, but this rapid information flow also amplifies the risk of misinformation. Traditional automatic detection methods that rely solely on textual features often struggle with nuanced, emerging content. This paper present a novel pipeline that verifies the factuality of social news by clustering related articles into events and using large language models (LLMs) to extract and verify claims against trusted news sources. The approach groups social-media posts and mainstream reports on the same event, extracts atomic claims with a model like GPT-4, and checks each claim's truthfulness by comparing it to the cluster's reliable news. This pipeline was evaluated on a newly constructed Vietnamese news dataset of 1,765 articles (including 723 social-media items), manually annotating claims as true or false. The LLM-based method achieved high accuracy (≈88.9% F1-score on claim verification and 92.1% F1-score on overall news verification). These results demonstrate that carefully prompted LLMs, combined with event-level clustering of evidence, can outperform traditional methods (e.g., bidirectional encoder representations from transformers (BERT)-based classifiers) in verifying news. The paper discusses advantages of clustering over simple retrieval, scalability considerations for LLMs, and prospects for multilingual and knowledge-enhanced verification.



***Corresponding Author:***

Tran Duc Duong
Faculty of Information Technology, Posts and Telecommunications Institute of Technology
Km. 10 Nguyen Trai, Ha Dong, Ha Noi, Viet Nam
Email: ducdt@ptit.edu.vn

## 1. INTRODUCTION

The proliferation of social-media news has democratized information but also facilitated the rapid spread of false news. Misinformation online can sway public opinion and even incite real-world harm, as past events have shown. Traditional fact-checking sites (e.g., PolitiFact, Snopes) rely on labor-intensive manual review, which is too slow for large-scale social feeds. Consequently, automated methods are needed. Early fake news detectors used content-based cues (linguistic style, sentiment, and stance) to classify news [1]–[4]. However, such methods often lack external grounding and can be easily fooled by adversarial or context-shifted content. In response, knowledge-based approaches have been developed: they extract check-worthy statements and compare them against evidence from trusted sources or knowledge bases. For example, the fact extraction and verification (FEVER) benchmark demonstrated a pipeline of claim extraction, evidence retrieval, and classification [5]. These methods typically employ pre-trained transformers (e.g. bidirectional encoder representations from transformers (BERT)) on features or evidence passages [6]–[10].

Recently, large language models (LLMs) (GPT-3/4) have shown remarkable reasoning and knowledge retrieval abilities, suggesting new fact-checking strategies [11], [12]. Unlike classification-only models, LLMs can be prompted to decompose and verify complex claims by generating evidence or

explanations [13], [14]. This work investigates such an LLM-powered pipeline for social news verification. Specifically, social-media articles (which may contain rumors or misinformation) are collected and cluster them together with reliable news reports on the same events. GPT-4 is then used to extract atomic claims from each social post and verify these claims against the cluster's trustworthy news. The main contributions of this paper are: i) a novel clustering + LLM pipeline for verifying social-media news by cross-checking with mainstream sources; ii) demonstration that LLMs can exceed traditional models for this task: our GPT-4 based approach achieves ≈88.9% F1-score on claim verification and 92.1% F1-score on overall news labeling, significantly outperforming a BERT-based baseline; and iii) insights into retrieval strategies: clustering by event is shown to yield more relevant evidence than naive "top-N" search and scales better.

The remainder of this paper is organized as follows. Section 2 reviews related literature. Section 3 describes the proposed methodology. Section 4 presents the dataset, experimental setup, and evaluation results. Finally, section 5 concludes the paper.

## 2. RELATED WORK

### 2.1. Text-based fake news detection

Early work on automated deception detection focused on linguistic cues and classification. For example, Wang [1] introduced the LIAR dataset of 12.8K labeled statements from PolitiFact and showed that neural models (convolutional neural network (CNNs), hybrid architectures) can use surface patterns for detection. Supervised models like BERT have since achieved strong results on such benchmarks. However, these approaches classify news based solely on the text, without grounding in real-world evidence. As Raza *et al.* [4] note that BERT-like models generally outperform LLMs in classification tasks (i.e., text-based classification). In other words, fine-tuned transformers excel when a training set is available, but they are limited when claims require external context. Distantly supervised datasets like NELA-GT or Fakeddit label articles from social sources by source credibility [4], but these can embed bias (e.g., labeling entire outlets as fake) and miss nuance.

### 2.2. Evidence-based fact checking

To more robustly verify claims, researchers have developed fact-checking pipelines that retrieve evidence. The FEVER dataset exemplifies this: it consists of 185K claims from Wikipedia with ground-truth evidence sentences [5]. Solutions typically extract a claim, retrieve relevant documents (via BM25 or neural search), and apply classifier to judge "Supported/Refuted/NotEnoughInfo". PolitiFact and other fact-checkers use similar processes with news or knowledge bases. For example, Yu *et al.* [15] fused text and retrieved fact-checking signals in a dual-perception model, improving on earlier BERT baselines. But even these enhanced architectures rely on matching surface text or shallow inference.

### 2.3. Large language model-driven fact verification

Recently, LLMs have been applied to fact verification in novel ways. One strand uses prompting or chain-of-thought to have an LLM analyze a claim. Zhang and Gao [12] introduced the hierarchical step-by-step (HiSS) prompting: they decompose complex claims into sub-claims and iteratively verify each using GPT-4, achieving state-of-the-art on misinformation benchmarks [12]. Min *et al.* [14] propose FactScore for fine-grained evaluation of LLM answers, highlighting that LLM accuracy hinges on context and prompt design. Metropolitansky and Larson [16] developed Claimify, an LLM-based extractor that breaks an output into precise factual statements. Experiments show that using high-quality extracted claims is crucial, since faulty claim phrasing can undermine verification. A major trend is to combine LLMs with the retrieval of external knowledge. Khaliq *et al.* [17] exemplify this with retrieval-augmented generation-augmented reasoning (RAGAR), which uses a multimodal LLM plus retrieval to fact-check political claims in tweets and images. They introduce chain-of-retrieval-augmented generation (RAG) and tree-of-RAG reasoning, whereby the model alternates between retrieving evidence and generating new sub-questions, achieving a weighted F1-score of 0.85 on a political claims test set. FactCheckGPT [18] implements an eight-step pipeline on LLM outputs. In experiments on the Factcheck-Bench (678 claims from ChatGPT responses), FactCheckGPT found that state-of-the-art checkers (FacTool, FactScore, and Perplexity) struggle – the best F1-score on false claims was only ~0.53. Another line treats LLMs as agents. Li *et al.* [13] propose FactAgent, which emulates human expert workflows: the model performs research subtasks sequentially (e.g., search, summarization, cross-checking) before giving a verdict. Quelle and Bovet [11] demonstrate that GPT-4 with retrieved context and source-citation can fact-check with promising accuracy, though accuracy varies based on query language and claim veracity. In summary, LLMs can leverage broad knowledge and complex reasoning, but their unchecked outputs can be inconsistent. Therefore, many recent systems combine LLMs with retrieval: e.g., integrating search results into LLM prompts for evidence [11].

A second family augments LLMs with structured knowledge or symbolic reasoning. GraphCheck [19] is a prime example: it extracts a knowledge graph (KG) from the document and claim, encodes it with a graph neural network, and feeds the resulting embeddings into an LLM prompt. Empirically, GraphCheck improved overall accuracy by ~7.1% over baselines on seven general and medical-domain benchmarks. Another hybrid system, HybridFC, is a hybrid KG-based fact-checking method that ensembles multiple approaches (text-based, path-based, rule-based, embedding-based) to decide the veracity of assertions (triples) in KGs. It reports improvements of 0.14 to 0.27 in area under the curve (AUC) over the prior state-of-the-art on the FactBench dataset [20].

Recent work has extended fact-checking to multimodal claims (text + images/videos). The FakeMark system [21] tackles fact-checking of news videos. FakeMark uses a question-driven pipeline: it passes the video, title, and transcript through a vision language model (VLM) and LLM (GPT-4o-mini and InternVL-8B) with tailored prompts. FakeMark's training-free pipeline achieved ~62.9% accuracy (F1≈59.4%) on binary veracity. Another multimodal approach is LRQ-Fact [22], designed for multimodal misinformation (images+text). It generates relevant fact-check questions by prompting a VLM + LLM pipeline. A third example is multimodal automated fact-checking via textualization (MAFT) [23], a general multimodal fact-checker for text, images, video, and audio. MAFT first textualizes all input media (e.g., generating captions/transcripts), then extracts claims, and collects evidence (via web search or KB), and finally generates a report.

The proposed method differs from these prior approaches in several ways. First, social news articles are pre-clustered by event before retrieval. This dramatically narrows the search space: instead of retrieving from an entire news corpus, we verify each post against a small set of event-related articles, similar in spirit to the clustering used in FASTTRACK [24]. This reduces noise and computational cost while preserving relevant evidence. Second, a plaintext LLM pipeline (GPT-4) is used without any fine-tuning or end-to-end training, relying on its zero/few-shot capabilities to extract claims and check them. In contrast, many prior systems either fine-tune models (e.g., Raza *et al.* [4] fine-tune BERT) or train classification heads. Third, the approach targets social media news posts in Vietnamese, a domain where few fact-checking studies exist. By operating in Vietnamese and aligning social posts with mainstream news clusters, a low-resource multilingual setting (comparable to the domain adaptation in ViFactCheck [25]) is addressed rather than mostly English political news. Finally, unlike monolithic classifiers, the pipeline outputs a "suspicious" label only if there is conflicting evidence between the post and its cluster; this cross-validation step leverages both the LLM's knowledge and concrete retrieved facts, an aspect similar to RAG but framed through event clusters. Overall, by combining LLM reasoning with cluster-constrained retrieval, the method achieves higher accuracy than baselines while remaining efficient.

### 2.4. News clustering and retrieval

Another key idea is clustering related news before verification. Instead of retrieving top-K documents for a claim, one can cluster the entire news corpus into events. Panchendrarajan *et al.* [26] introduced MultiClaimNet, clustering similar claims across languages to reduce redundancy and improve retrieval efficiency [26]. FASTTRACK [24] takes a two-stage approach: it builds an offline clustering index of the corpus, then for each query retrieves only relevant clusters and uses an LLM for fine-grained evidence scoring. This dramatically shrinks the search space; FASTTRACK reported >33× speed-ups and much higher accuracy than naïve retrieval. Recent clustering-based RAG frameworks such as CRAVE [27] further demonstrate that organizing documents into coherent event or narrative clusters improves evidence consistency and contextual grounding compared to flat top-K retrieval.

In the pipeline, social posts and trustworthy news are similarly grouped by event, so that each claim is checked against the most relevant cluster of sources (rather than millions of documents). As other works note, clustering by event "improves claim retrieval and validation" by reducing redundancy [26]. This approach also addresses practical issues: selecting a fixed top-N results can miss context or include noise, whereas clustering adapts to the data distribution.

## 3. METHOD

To ensure the news verification process is both systematic and transparent, the system developed is structured into four main, interconnected stages. These stages are include: i) news collection and event clustering, ii) claim extraction, iii) claim verification, and iv) final news verification.

### 3.1. News collection and clustering

A dataset of Vietnamese news articles from late 2024 was gathered, including social-media posts (e.g., viral Facebook, YouTube news) and mainstream outlets (e.g., VnExpress, Vietnamnet). In total, 1,765

articles were collected, covering domains like politics, disasters, sports, and entertainment. Social posts are treated as "items to verify", while the trusted news serves as evidence. The text is preprocessed (cleaning HTML and normalizing fonts) and then cluster articles by event. For clustering, density-based spatial clustering of applications with noise (DBSCAN) [28] is used on document embeddings (with text embeddings extracted using PhoBERT [29]), which groups together articles mentioning the same event or topic. Temporal metadata (publication date) is also incorporated to prevent unrelated stories with similar content from being conflated. After tuning, DBSCAN effectively isolates clusters of arbitrary shape and density, capturing events ranging from natural disasters to press conferences. Each cluster thus contains all social and mainstream articles related to one event.

By clustering first, evidence retrieval is restricted to the cluster of interest. In contrast to "top-N" search (where choosing N is arbitrary and may omit relevant context), clustering adaptively collects all pertinent references. As prior studies show, clustering events can significantly reduce the search space and improve evidence relevance [24]. For example, during verification, we only look at the reliable news articles in the same cluster as a social post, rather than querying a web search engine. This ensures all checked sources are on-topic and of known credibility, which also mitigates the issue of conflicting or spurious retrieval results.

### 3.2. Claim extraction

Given a social-media article, atomic claims are identified within it. An atomic claim is defined as a concise statement that conveys one verifiable piece of information [14]. For instance, "The president will visit Japan in September" is atomic, whereas a sentence listing multiple facts is not. Breaking content into atomic claims simplifies verification and avoids partial truths. GPT-4 is used to extract claims by prompting the model with the article text, asking it to list the key factual statements. In accordance with recent work, a carefully designed prompt (e.g., few-shot examples) is applied to ensure high-precision extraction. Works such as Claimify [16] and fact-in-fragments [30] demonstrate that decomposing long-form content into atomic, verifiable units significantly improves downstream fact-checking reliability. Claimify [16] and related methods have shown that the quality of extracted claims strongly affects downstream fact-checking. Claims that are unverifiable (e.g., opinions or ungrounded speculation) are further filtered out, and only check-worthy factual claims are passed to the next stage. This extraction step aligns with practices in LLM-based evaluation: the FactScore framework also breaks text into atomic facts to assess factuality [14]. By focusing the LLM's effort on each claim independently, we improve consistency and traceability of the verification. In experiments, systems using extracted claims are shown to outperform attempts to verify entire articles at once. Figure 1 illustrates the prompt design for claim extraction.

```
You are an assistant that extracts atomic factual claims from text.
An atomic claim is a concise statement about a single fact that can be checked as true or false.
Instructions:
- Exclude opinions, predictions, or unverifiable statements.
- Break complex or compound sentences into separate claims.
- Each claim must be short, clear, and independent.
- Output as a numbered list.

Text to analyze:
"{ARTICLE_TEXT}"

Task: List the atomic factual claims from the above article.
```

Figure 1. Prompt design for the claim extraction task

### 3.3. Claim verification

Each extracted atomic claim is then verified against evidence. An LLM (GPT-4) is prompted to evaluate the truth of a claim given relevant context. For context, the trusted news articles in the same cluster are used. Specifically, the model is provided with the claim text and the concatenated contents (or summaries) of all cluster news, and ask it to respond "true" or "false" with a brief justification. A multi-step reasoning strategy inspired by hierarchical prompting is adopted: if a claim is complex, the LLM can decompose it and check sub-components (similar to HiSS [12]).

In practice, providing GPT-4 with a few-shot exemplars (claim + evidence + verdict) is found to improve accuracy. The LLM thus acts as a smart comparator: leveraging both its internal world knowledge

and the provided cluster context. This differs from pure retrieval-based checks (which might lack deep inference) and from naive classification (which lacks dynamic evidence). It is also related to agentic approaches: FactAgent [13] demonstrated that breaking verification into subtasks (search, judgment) yields better performance. No model is trained end-to-end; instead, GPT-4's zero-shot/few-shot capability is relied upon. Figure 2 illustrates the prompt design for claim verification.

To assess verification reliability, the LLM is required to provide verbatim evidence citations for each verdict. A verdict is considered evidence-grounded only if all quoted sentences appear verbatim in the referenced articles and logically support the conclusion. The hallucination rate is measured as the fraction of verdicts that either: i) lack valid citations or ii) cite sentences that do not support the verdict. Additionally, a distinguish is made between verdicts grounded in cluster evidence and those inferred primarily from the model's internal knowledge. This design aligns with recent efforts to enforce citation-grounded reasoning to reduce hallucination in LLM outputs [31]. Results show that enforcing citation substantially reduces unsupported judgments, though a small fraction of hallucinated decisions persists, consistent with prior findings on LLM factuality limits.

### 3.4. News verification

Finally, claim-level judgments are aggregated to label each social news item. A strict approach is taken: a social news article is marked verified only if all its check-worthy claims are true; if any claim is false, the article is marked as unverified. This conjunctive rule makes sense because one false assertion is enough to deem the news unreliable. The model's justifications also provide transparency for each claim. In the dataset annotation, human annotators likewise labeled each claim and then derived the article label by the same rule.

```
You are a fact-checking assistant.
You will verify a claim against a set of trusted news articles.
Instructions:
- A fact checker will decompose the claim into subclaims that are easier to verify.
- To verify a subclaim, a fact-checker will go through a step-by-step process to ask and answer a
series of questions relevant to its factuality. If multiple articles conflict, follow the majority
evidence from the reliable sources.
- Quote the exact supporting sentence(s) from the trusted articles. Each quoted sentence must be
copied verbatim and attributed to its source.

Claim:
"{CLAIM_TEXT}"

Trusted News Articles:
{EVIDENCE_TEXT}

Task: Is the claim True or False? If True, provide the quoted sentence(s) with its SourceID.
```

Figure 2. Prompt design for claim verification task

## 4. RESULTS AND DISCUSSION

### 4.1. Dataset

A custom dataset is constructed for evaluation. It contains 1,765 Vietnamese news articles: 723 social-media items to verify, and 1,042 "trusted" mainstream news serving as evidence. Articles span November–December 2024 and cover events in politics, disasters, sports, and entertainment. For each social article, atomic claims are manually annotated (3,373 claims in total) and verified their truth by consulting cluster news. In annotation, a claim was marked true if supported by evidence, or false otherwise. It is ensured that each claim had, on average, 3 supporting articles (≥300 words each) in the cluster. The richer evidence set contrasts with some public datasets where only one or two sources are given per claim. The annotations yielded 2,634 true claims and 730 false ones; correspondingly, 352 social articles became verified (all claims true), and 371 became unverified (≥1 false claim).

It is emphasized that existing Vietnamese or multilingual fact-check datasets lack this combination of social news with multiple high-quality references. Hence, the dataset is created specifically for this study. Table 1 summarizes its key statistics. For reproducibility, the articles are ensured to come from late 2024 so they were not in the LLM's training data.

In addition, two other datasets, ViWikiFC and ViFactCheck, are also chosen as evaluation benchmarks for the method. The reasons for choosing these datasets are Vietnamese fact-checking datasets in contexts similar to the task and well-established in the literature. ViWikiFC is an open-domain Vietnamese Wikipedia fact-checking corpus (over 20,000 claims) [32], while ViFactCheck is a Vietnamese multi-domain news fact-checking dataset with 7,232 claims [25]. Both are manually curated and in Vietnamese, matching the language and style of the data. By contrast, popular English benchmarks (e.g., FEVER on English Wikipedia) operate on different languages and domains [5] and would not reflect Vietnamese-specific linguistic phenomena.

Table 1. Dataset statistics

| Type | Quantity |
|---|---|
| Social news | 723 |
| Number of extracted claims | 3,373 |
| Number of claims verified as "true" | 2,643 |
| Number of claims verified as "false" | 730 |
| Number of news classified as true (all claims are true) | 352 |
| Number of news classified as false (at least one claim is false) | 371 |
| News from trusted sources | 1,042 |
| Total news | 1,765 |

### 4.2. Experimental setup

Experiments are structured around key research questions:

i) RQ1: are LLMs effective at social news verification? the LLM pipeline is compared against a strong baseline: a BERT-based model (a reimplementation of a Vietnamese fake-news detector [10]) trained to classify articles directly. This simulates "traditional" natural language processing (NLP) verification. Precision, recall, and F1-score are reported for claim-level and article-level labels.

ii) RQ2: does claim extraction help? the same LLM verification verification is tested when applied to full articles (without claim extraction) versus atomic claims. Prior work suggests verifying concise claims yields more reliable results [16].

iii) RQ3: clustering vs. top-N retrieval: the clustering-based approach is compared to a simpler scheme: for each social article, use an online search application programming interface (API) to retrieve the top-3 potentially related news (ignoring clustering), then verify claims against those. This evaluates the benefit of offline clustering of all collected news.

In all LLM experiments, the GPT-4 model is used. Claims and cluster news are fed via a LangChain interface. For the claim-extraction pipeline, a few-shot prompt with GPT-4 is used for atomic claims (similar to the Claimify method [16]).

### 4.3. Results

To assess the quality of LLM-based claim extraction, extracted claims are manually evaluated on a random subset of 200 social news posts, stratified across topics. Each post was annotated independently by two human annotators with experience in misinformation analysis. For each post, annotators first identified the set of gold atomic factual claims present in the text. Extracted claims were then matched against gold claims using semantic equivalence criteria. Precision, recall, and F1-score are reported in Table 2. These results indicate that the LLM extracts most verifiable claims while maintaining high precision, which is crucial since downstream verification cost grows linearly with the number of extracted claims.

Table 2. Claim extraction evaluation

| Metric | Score (%) |
|---|---|
| Precision | 89.7 |
| Recall | 87.5 |
| F1-score | 88.6 |

All extraction errors are categorized into four mutually exclusive types. Missing facts account for 42% of errors, where the model captures the main claim but omits secondary verifiable details (e.g., dates or numerical values), primarily affecting recall. Over-splitting or redundancy represents 28%, where a single factual statement is fragmented into overlapping claims, complicating alignment with gold annotations. Including opinions or subjective statements accounts for 18% of errors and negatively impacts precision. The remaining 12% involve contextual incompleteness or entity ambiguity, where extracted claims lack sufficient

specificity. This distribution explains the slightly lower recall (87.5%) compared to precision (89.7%) and aligns with typical error patterns reported in recent LLM-based claim extraction studies. Event clustering quality is evaluated using purity and normalized mutual information (NMI) on a manually labeled subset (300 articles) and report the results in Table 3. High purity confirms that evidence clusters are semantically coherent, which directly benefits verification accuracy.

Macro-averaged precision, recall, and F1-score for claim verification and article (news) verification are reported in Table 4. To prevent information leakage, an event-level train/test split is applied. All articles are first grouped into event clusters, and entire clusters are assigned exclusively to either training or test sets. This ensures that no event appears in both splits. For LLM-based zero-shot evaluation, near-duplicate content between social posts and trusted articles is further removed using string similarity filtering, ensuring that evidence is not trivially copied.

- Baseline (BERT): the BERT model (SemViQA architecture adapted to Vietnamese [10]) achieved an F1-score of only 77.7% on claim-level labels and 74.6% on news-level labels. This relatively low performance highlights the challenge of capturing evidence without context.
- LLM (no extraction): verifying full articles with GPT-4 (using all cluster text) gives an F1-score of 69.8% on news-level labels (no claims are extracted, so there is no result on claim-level labels). The not good result may be suffered from verbosity and confusion in long inputs.
- LLM (with extraction, top-N search): when claims are first extracted and the top 3 relevant cluster articles as evidence (HiSS-style prompt [12]), F1-score increases to 86.7% on claim-level labels and 89.6% on news-level labels. This shows that limiting context to high-quality sources helps.
- LLM (with extraction, clustering): the full pipeline (claims + LLM + cluster evidence) achieved ~88.9% F1-score on claims and 92.1% F1-score on news. This was the best result, far surpassing the BERT baseline and the non-cluster variants. In particular, recall of false claims was high, meaning most false statements were caught.

These results confirm recent findings that well-prompted LLMs can match or exceed supervised models on fact-checking [12]. The usage of GPT-4 with contextual cluster evidence yielded substantial gains. It also validates that clustering is superior to naïvely retrieving fixed top-N sources: clustering ensured that no relevant evidence was missed and that off-topic content was not included. In experiments, clustering improved precision by several points over top-N retrieval for the same number of sources. As expected, the comparison also shows the advantage of claim extraction: verifying concise atomic statements significantly outperformed evaluating whole articles, corroborating the benefit of focused claim decomposition [16].

Table 5 reports results on the test split of each benchmark. A strong baseline model (InfoXLM-large on ViWikiFC [32]; Gemma on ViFactCheck [25]) is compared to the proposed method. Each score is shown with a 95% confidence interval, and we include a p-value from a paired test to check significance. These results show that the method consistently slightly outperforms the baselines on both datasets (albeit by only 1–2 points), while the confidence intervals quantify the reliability of each score.

Table 3. Cluster purity and NMI

| Metric | Score (%) |
|---|---|
| Purity | 91.4 |
| NMI | 86.7 |

Table 4. Experiment results

| Method | Claim verification | | | News verification | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | F1-score | Precision | Recall | F1-score |
| BERT-based baseline (SemViQA [10]) | 75.3 | 80.2 | 77.7 | 63.3 | 90.7 | 74.6 |
| LLM (no extraction) (HiSS [12]) | | | | 63.6 | 77.3 | 69.8 |
| LLM-based with top N trusted news | 93.6 | 80.8 | 86.7 | 95.3 | 84.5 | 89.6 |
| LLM-based with news clustering | 93.1 | 85.1 | 88.9 | 93.6 | 90.7 | 92.1 |

Table 5. Macro-F1 (in %) on Vietnamese fact-checking benchmarks

| Dataset | Baseline model | Baseline F1 | Proposed method |
|---|---|---|---|
| ViWikiFC | InfoXLM-large | 67.0 | $68.2 \pm 0.5$ |
| ViFactCheck | Gemma (LLM) | 89.9 | $90.7 \pm 0.4$ |

The results in Table 5 indicate that the method achieves a slight improvement over the baselines on both Vietnamese benchmarks. On ViWikiFC, InfoXLM-large scored about 67.0% for pipeline (combination of both tasks), and proposed method scores ~68.2% (±0.5). On ViFactCheck, Gemma scored 89.90%,

while the proposed method reaches ~90.7% (±0.4). The differences (≈1.2 and 0.8 points) are small, but the statistical tests confirm they are unlikely to be due to random chance ($p < 0.05$). Reporting confidence intervals helps to see that the intervals barely overlap, supporting the claim of a real gain. These small gains are expected, given that the baselines are already strong. The key point is that proposed approach provides consistently higher performance in the same Vietnamese context. This suggests proposed method better handles Vietnamese-specific linguistic patterns or domain knowledge.

### 4.4. Discussion

While the pipeline achieved high accuracy, several practical challenges remain. First, ambiguity in claims can stump the LLM: some statements required nuanced world knowledge or context beyond what was provided. For example, a claim about a policy announcement without named participants needed cross-checking. Integrating external knowledge (e.g., Wikipedia or a KG) could help disambiguate such claims. Indeed, surveys show that LLMs benefit from structured knowledge sources to improve factual grounding [33]. Second, conflicting sources within a cluster can lead to inconsistent verdicts. In a few cases, two reputable outlets quoted contradictory figures for an event. To mitigate this, one could weight sources by reliability (e.g., favoring government or major press). In the experiments, the LLM sometimes arbitrarily chose one source, so human-in-the-loop weighting or an ensemble of LLM judgments could increase robustness. Third, scalability and latency are concerns. Running GPT-4 on many claims is computationally expensive, posing challenges for real-time deployment. Recent industry practice suggests model distillation or quantization techniques to speed up inference. Alternatively, a hybrid approach might use a faster, smaller LLM for easy cases and reserve GPT-4 for ambiguous claims. The FASTTRACK framework shows that offline clustering can dramatically reduce query load, and the results align with their finding that clustering scales to tens of thousands of documents while retaining high accuracy [24].

Finally, multilingualism and knowledge integration are promising future directions. The current work focuses on Vietnamese content, but the architecture can naturally extend: many LLMs are multilingual, and evidence clusters could include sources in any language. For instance, checking a Vietnamese claim against English Wikipedia via GPT's translation ability. This cross-lingual integration could further improve coverage. Surveys note that combining LLMs with domain ontologies and knowledge bases can enhance reasoning [33]; plan to explore connecting the pipeline to medical, financial, or geographic KGs when verifying specialized news. The proposed factuality verification framework can be integrated into newsrooms and social-media moderation systems through APIs that support event clustering, claim extraction, and evidence retrieval. In practice, the model functions as a human-in-the-loop assistant, automatically generating verification summaries and confidence scores that editors or fact-checkers can review before publication. This hybrid design increases efficiency while maintaining editorial control and accountability (a demo version is at https://8am.vn). To ensure ethical and trustworthy deployment, potential language and topic biases are mitigated through balanced dataset fine-tuning and periodic audits of model outputs. All retrieved evidence is traceable to public sources, and user data is never stored. By combining scalable automation with transparency and human oversight, the system supports responsible use of LLM-based fact-checking in real-world news environments.

## 5. CONCLUSION

A robust framework for verifying the factuality of social-media news by leveraging LLMs and event-based clustering has been presented. The contributions include demonstrating that GPT-4, when guided by high-quality evidence clusters and atomic claims, can greatly outperform traditional BERT-based classifiers on a Vietnamese news verification task. News clustering was shown to be a powerful strategy for evidence retrieval, yielding more relevant context than simple top-N search. The high accuracy achieved (≈90–92% F1-score) underscores the potential of AI-driven fact-checking. At the same time, key limitations are identified: LLM inference costs challenge real-time use, and ambiguous or unsupported claims can still slip through. Future research will address these by incorporating efficient model serving (distillation, caching) and richer knowledge sources (ontologies, KGs) for disambiguation. Also planned is expansion to multilingual verification: training with cross-lingual clusters to allow a Vietnamese claim to be checked against English and Chinese content. Another avenue is leveraging LLMs' ability to cite sources explicitly – for greater transparency, the model might provide factual justifications with links or references. In summary, this work demonstrates that LLM-based pipelines, augmented with event clustering, are a promising direction for combating misinformation on social platforms. As LLM technology evolves and becomes more efficient, such systems are anticipated to play an increasingly practical role in real-time news verification, ultimately contributing to a more trustworthy information ecosystem.

**FUNDING INFORMATION**

Authors state no funding involved.

**AUTHOR CONTRIBUTIONS STATEMENT**

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

| Name of Author | C | M | So | Va | Fo | I | R | D | O | E | Vi | Su | P | Fu |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Tran Duc Duong | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | | ✓ | ✓ | |
| Hai Hoan Do | ✓ | | | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ | | | | |

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

**CONFLICT OF INTEREST STATEMENT**

Authors state no conflict of interest.

**DATA AVAILABILITY**

The data that support the findings of this study are available from the corresponding author, [TDD], upon reasonable request.

**REFERENCES**

[1] W. Y. Wang, "'Liar, liar pants on fire': a new benchmark dataset for fake news detection," in *55th Annual Meeting of the Association for Computational Linguistics*, Vancouver, Canada, 2017, pp. 422–426, doi: 10.18653/v1/P17-2067.

[2] V. P. Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea, "Automatic detection of fake news," in *COLING 2018 - 27th International Conference on Computational Linguistics,* Santa Fe, New Mexico, 2018, pp. 3391–3401.

[3] I. Augenstein *et al.*, "MultiFC: a real-world multi-domain dataset for evidence-based fact checking of claims," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, 2019, pp. 4684–4696, doi: 10.18653/v1/D19-1475.

[4] S. Raza, D. P. Patterson, and C. Ding, "Fake news detection: comparative evaluation of BERT-like models and large language models with generative AI-annotated data," *Knowledge and Information Systems*, vol. 67, no. 4, pp. 3267–3292, 2025, doi: 10.1007/s10115-024-02321-1.

[5] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "FEVER: a large-scale dataset for fact extraction and verification," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* New Orleans, Louisiana, 2018, pp. 809–819, doi: 10.18653/v1/N18-1074.

[6] A. Soleimani, C. Monz, and M. Worring, "BERT for evidence retrieval and claim verification," in *Advances in Information Retrieval: 42nd European Conference on IR Research (ECIR 2020)*, Lisbon, Portugal, 2020, pp. 359–366, doi: 10.1007/978-3-030-45442-5_45.

[7] C. Kruengkrai, J. Yamagishi, and X. Wang, "A multi-level attention model for evidence-based fact checking," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 2447–2460, doi: 10.18653/v1/2021.findings-acl.217.

[8] F. Yang *et al.*, "XFake: explainable fake news detector with visualizations," in *The World Wide Web Conference*, New York, United States, May 2019, pp. 3600–3604, doi: 10.1145/3308558.3314119.

[9] N. Kotonya and F. Toni, "Explainable automated fact-checking for public health claims," in *2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 7740–7754, doi: 10.18653/v1/2020.emnlp-main.623.

[10] D. X. Tran *et al.*, "SemViQA: a semantic question answering system for Vietnamese information fact-checking," *arXiv:2503.00955*, 2025.

[11] D. Quelle and A. Bovet, "The perils and promises of fact-checking with large language models," *Frontiers in Artificial Intelligence*, vol. 7, Feb. 2024, doi: 10.3389/frai.2024.1341697.

[12] X. Zhang and W. Gao, "Towards LLM-based fact verification on news claims with a hierarchical step-by-step prompting method," in *13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, Nusa Dua, Bali, 2023, pp. 996–1011, doi: 10.18653/v1/2023.ijcnlp-main.64.

[13] X. Li, Y. Zhang, and E. C. Malthouse, "Large language model agentic approach to fact checking and fake news detection," in *27th European Conference on Artificial Intelligence, 19–24 October 2024, Including 13th Conference on Prestigious Applications of Intelligent Systems (PAIS 2024)*, Santiago de Compostela, Spain, 2024, doi: 10.3233/FAIA240787.

[14] S. Min *et al.*, "FActScore: fine-grained atomic evaluation of factual precision in long form text generation," in *2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 12076–12100, doi: 10.18653/v1/2023.emnlp-main.741.

[15] W. Yu *et al.*, "Research on fake news detection based on dual evidence perception," *Engineering Applications of Artificial Intelligence*, vol. 133, Jul. 2024, doi: 10.1016/j.engappai.2024.108271.

[16] D. Metropolitansky and J. Larson, "Towards effective extraction and evaluation of factual claims," in *63rd Annual Meeting of the Association for Computational Linguistics*, 2025, pp. 6996–7045.
[17] M. A. Khaliq, P. Y.-C. Chang, M. Ma, B. Pflugfelder, and F. Miletić, "RAGAR, your falsehood radar: RAG-augmented reasoning for political fact-checking using multimodal large language models," in *Seventh Fact Extraction and VERification Workshop (FEVER)*, Miami, Florida, 2024, pp. 280–296.
[18] Y. Wang *et al.*, "Factcheck-bench: fine-grained evaluation benchmark for automatic fact-checkers," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Miami, Florida, 2024, pp. 14199–14230.
[19] Y. Chen *et al.*, "GraphCheck: breaking long-term text barriers with extracted knowledge graph-powered fact-checking," in *63rd Annual Meeting of the Association for Computational Linguistics*, 2025, pp. 14976–14995, doi: 10.18653/v1/2025.acl-long.729.
[20] U. Qudus, M. Röder, M. Saleem, and A.-C. N. Ngomo, "HybridFC: a hybrid fact-checking approach for knowledge graphs," in *The Semantic Web (ISWC 2022)*, Cham, Switzerland: Springer, 2022, pp. 462–480, doi: 10.1007/978-3-031-19433-7_27.
[21] B. Yang, D. Xu, K. Niu, W. Liu, Z. Wang, and M. Kankanhalli, "A new dataset and benchmark for grounding multimodal misinformation," in *33rd ACM International Conference on Multimedia*, New York, USA, Oct. 2025, pp. 12571–12577, doi: 10.1145/3746027.3758191.
[22] A. Beigi *et al.*, "Can LLMs improve multimodal fact-checking by asking relevant questions?," in *2025 IEEE International Conference on Big Data (BigData)*, Dec. 2025, pp. 2732–2741, doi: 10.1109/BigData66926.2025.11400846.
[23] K. Kakizaki, Y. Matsunaga, and R. Furukawa, "MAFT: multimodal automated fact-checking via textualization," in *Proceedings of the 39th Annual AAAI Conference on Artificial Intelligence*, 2025, pp. 29646–29648, doi: 10.1609/aaai.v39i28.35354.
[24] S. Chen, F. Kang, N. Yu, and R. Jia, "FASTTRACK: reliable fact tracing via clustering and LLM-powered evidence validation," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Miami, Florida, 2024, pp. 5821–5836, doi: 10.18653/v1/2024.findings-emnlp.334.
[25] T. T. Hoa, T. Q. Duy, K. Q. Tran, and K. V. Nguyen, "ViFactCheck: a new benchmark dataset and methods for multi-domain news fact-checking In Vietnamese," in *Proceedings of the 39th AAAI Conference on Artificial Intelligence*, 2025, pp. 308–316, doi: 10.1609/aaai.v39i1.32008.
[26] R. Panchendrarajan, R. M. Pérez, and A. Zubiaga, "MultiClaimNet: a massively multilingual dataset of fact-checked claim clusters," in *Association for Computational Linguistics: EMNLP 2025*, 2025, pp. 11203–11215, doi: 10.18653/v1/2025.findings-emnlp.599.
[27] A. U. Dey, M. J. Awan, G. Channing, C. S. D. Witt, and J. Collomosse, "Fact-checking with contextual narratives: leveraging retrieval-augmented LLMs for social media analysis," *IEEE Transactions on Computational Social Systems*, vol. 13, no. 3, pp. 3365–3376, Jun. 2026, doi: 10.1109/TCSS.2026.3669799.
[28] R. J. G. B. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in *Advances in Knowledge Discovery and Data Mining (PAKDD 2013)*, 2013, pp. 160–172, doi: 10.1007/978-3-642-37456-2_14.
[29] D. Q. Nguyen and A. T. Nguyen, "PhoBERT: pre-trained language models for Vietnamese," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 1037–1042, doi: 10.18653/v1/2020.findings-emnlp.92.
[30] L. Zheng *et al.*, "Fact in fragments: deconstructing complex claims via LLM-based atomic fact extraction and verification," *Expert Systems with Applications*, vol. 302, Mar. 2026, doi: 10.1016/j.eswa.2025.130572.
[31] K. Wu *et al.*, "An automated framework for assessing how well LLMs cite relevant medical references," *Nature Communications*, vol. 16, no. 1, Apr. 2025, doi: 10.1038/s41467-025-58551-6.
[32] H. T. Le, L. T. To, M. T. Nguyen, and K. V. Nguyen, "ViWikiFC: fact-checking for Vietnamese wikipedia-based textual knowledge source," 2024, *arXiv:2405.07615*.
[33] W. Yang, L. Some, M. Bain, and B. Kang, "A comprehensive survey on integrating large language models with knowledge-based methods," *Knowledge-Based Systems*, vol. 318, Jun. 2025, doi: 10.1016/j.knosys.2025.113503.

## BIOGRAPHIES OF AUTHORS

**Tran Duc Duong** holds a Doctor of Computer Engineering degree from Posts and Telecommunications Institute of Technology, Vietnam in 2017. He also received his B.Sc. from Vietnam National University of Hanoi (Information Technology) and M.Sc. (Information Systems) from University of Leeds, United Kingdom in 1999 and 2004, respectively. He is currently a lecturer at Faculty of Information Technology, Posts and Telecommunications Institute of Technology, Hanoi, Vietnam. His research includes machine learning, deep learning, image and natural language processing, and LLMs. He can be contacted at email: ducdt@ptit.edu.vn.

**Hai Hoan Do** received a Doctor of Economics from the Vietnam Academy of Social Sciences, Hanoi, Vietnam and Master of Project Management from Foreign Trade University, Hanoi, Vietnam in 2018 and 2013, respectively. She is currently a lecturer at the Economic Research Institute of Postal Technology, Posts and Telecommunications Institute of Technology, Hanoi, Vietnam. Her research includes social entrepreneurship, social communications, and social press. She can be contacted at email: hoandh@ptit.edu.vn.