RAVE: RETRIEVAL AND SCORING AWARE VERIFIABLE CLAIM DETECTION

Yufeng Li, Arkaitz Zubiaga

Queen Mary University of London, London, UK

ABSTRACT

The rapid spread of misinformation on social media underscores the need for scalable fact-checking tools. A key step is claim detection, which identifies statements that can be objectively verified. Prior approaches often rely on linguistic cues or claim check-worthiness, but these struggle with vague political discourse and diverse formats such as tweets. We present RAVE (Retrieval and Scoring Aware Verifiable Claim Detection), a framework that combines evidence retrieval with structured signals of relevance and source credibility. Experiments show that RAVE achieves competitive results on CT22-test and PoliClaim-test.

Index Terms— Automated Fact-Checking, Verifiable Claim Detection, Information Retrieval

1. INTRODUCTION

On social media, misinformation spreads more quickly and widely than in traditional formats, and the content is highly diverse, ranging from news and political debates to online posts such as tweets [1, 2]. Such content can shape public opinion and cause societal harm, making fact-checking increasingly important. Traditional fact-checking, however, is labor-intensive and time-consuming, which has led to growing interest in Automated Fact-Checking Systems (AFC) that leverage artificial intelligence to support and scale this process [3]. Given the scale and speed of misinformation, claim detection serves as a crucial first step. It identifies factual statements that require verification, filtering verifiable claims for the downstream task of claim verification, where they are ultimately fact-checked as true or false.

The task of claim detection has been defined in different ways, with most studies focusing on claim check-worthiness [4], where claims are prioritized based on criteria such as public importance or interest [5]. More recent work shifts toward verifiable claim detection, which defines a claim as an assertion that can be objectively checked [6, 7]. In line with this view and inspired by the downstream task of claim verification, we define **a verifiable claim as one that contains at least one factual statement that can be objectively verified through external evidence**.

Early approaches to claim detection relied on traditional machine learning systems such as ClaimBuster [8], ClaimRank [9], and the CNC system [6]. With the rise of deep learning, the CheckThat! shared tasks introduced LSTM-based models [10], but transformer-based architectures have dominated since 2020 [11]. More recently, large language models (LLMs) have driven further progress: top systems in CheckThat! 2023 used GPT-3 in zero-shot and few-shot settings [12], and the 2024 winner fine-tuned eight open-source LLMs [13]. AFaCTA extended this line of work with a multi-step prompting framework that improved factual claim annotation [7].

These methods have been evaluated on two main datasets: sentence-level political debate transcripts and COVID-19 tweets independently. Debate transcripts are often short, vague, and increasing ambiguity, while tweets are more self-contained and explicit. For example, debates may include statements like “the last couple of years have been especially trying for our medical professionals", which lack clear reference points, whereas tweets such as “Pfizer isn’t Lamborghini. Sinovac isn’t Proton" are more concrete. Most research has concentrated on debates [6, 14, 7], with fewer studies addressing tweets [15, 16].

To address these limitations, we propose RAVE (Retrieval-and-Scoring-Aware Verifiable Claim Detection), a framework that improves verifiable claim detection by incorporating evidence retrieval and scoring into the decision process. Our intuition is that the verifiability depends not only on its linguistic form but also on the availability and quality of external evidence that can be used in the downstream task. We evaluate our approach on the CT22-test [16] and PoliClaim-test [7] datasets, comparing it with standard text-only baselines and retrieval-based variants to isolate the role of scoring and selection strategies. Our main contributions are:

- We provide a precise formulation of claim detection as the task of identifying **verifiable claims**, defined as factual statements that can be objectively checked against external evidence, regardless of their truth value.
- We introduce **RAVE**, a retrieval- and LLM-based framework that integrates evidence retrieval with relevance and credibility scoring, using these scores for both evidence selection and verifiability estimation. Code is on GitHub.
- We conduct a systematic comparison of claim detection across datasets with different linguistic and structural properties, providing empirical insights into how dataset characteristics affect verifiability modeling.

Algorithm 1: RAVE: **R**etrieval-and-Scoring-**A**ware **V**erifiable Claim D**E**tection

Input: Claim x

Output: Verifiability label
$y \in \{\text{VERIFIABLE}, \text{NON-VERIFIABLE}\}$

Step 1: Entity-based retrieval

Extract entities $E = \{e_1, \ldots, e_m\}$ from x (PERSON, ORG, LOCATION, EVENT, CLAIM_OBJECT).

For each $e_i \in E$, query a web search API and collect candidate snippets $S = \{s_1, \ldots, s_n\}$ with metadata (domain, title, url).

Step 2: Scoring

For each snippet $s_j \in S$:

Compute relevance $r_j = \cos(\mathbf{h}_x, \mathbf{h}_{s_j})$ between embeddings.

Compute credibility c_j from source domain metadata.

Step 3: Scoring-aware selection

Aggregate score: $score_j = \alpha r_j + (1 - \alpha)c_j$.

Select top-K snippets $S_K \subseteq S$ according to $score_j$.

Step 4: Verifiability decision

Build a structured prompt including:

- Claim x
- Selected snippet contents, metadata (domain, title)
- Relevance and credibility scores of each snippet.

Feed the prompt to an LLM.

Obtain output
$y \in \{\text{VERIFIABLE}, \text{NON-VERIFIABLE}\}$.

2. PROPOSED METHOD

To move beyond simple evidence retrieval, we propose a scoring-aware evidence selection and verifiability decision framework for claim detection, called RAVE (Retrieval-and-Scoring-Aware Verifiable Claim Detection), as shown in Algorithm 1.

2.1. Entity-based Retrieval

To retrieve initially relevant evidence, we first perform entity extraction on the input claim. Entities are short and informative units that make a sentence more suitable for linking to external knowledge. We define five types of entities: PERSON, ORG, LOCATION, EVENT, and CLAIM_OBJECT. The latter two types capture specific occurrences and central objects or concepts in a claim, which extend beyond traditional entity definitions. Entities are extracted using a zero-shot learning approach with prompt-based instructions.

Once entities are extracted, we query the Google API with each entity to retrieve candidate snippets, denoted as $S = \{s\_1, \ldots, s\_n\}$, along with their metadata. These snippets are used to construct a context pool for the claim. We retain snippets rather than full web pages, since the goal is not to gather comprehensive evidence for verification but to provide contextual signals that help determine whether a claim is verifiable.

Score	Domain type / Examples
1.00	Highly authoritative (Wikipedia, Reuters, BBC, Nature)
0.95	Government (.gov), educational (.edu)
0.85	Academic and research institutions (universities, institutes)
0.75	Established news outlets (news, times, post, journal)
0.65	Non-profit organizations (.org)
0.50	General commercial (.com)
0.40	Other or unclassified domains

Table 1. Heuristic credibility scores by domain type.

2.2. Scoring

To evaluate the usefulness of retrieved snippets, we assign two scores. Relevance is measured by cosine similarity between a snippet and the input claim, while source credibility reflects the reliability of the snippet's domain following a source-centric view of information quality [17]. Each domain is assigned a normalized score between 0.4 and 1.0 (Table 1), which estimates the general reliability of a source rather than the truthfulness of individual claims, consistent with prior fact-checking and information retrieval research [18].

2.3. Scoring-aware Selection

We then compute a combined score for each snippet:

$$score_j = \alpha r_j + (1 - \alpha)c_j$$

The parameter α is tuned on the CT22-dev dataset through a coarse grid search over $\{0.3, \ldots, 0.8\}$, and $\alpha = 0.6$ is selected using the one-standard-error rule; no additional tuning is applied to other datasets or models. This score balances snippet relevance (r_j) with source credibility (c_j), supporting snippets that are both contextually aligned and reliable. In this way, the retrieved context reduces noise and misinformation while supporting effective claim detection.

2.4. Verifiability decision

In the final decision step, RAVE treats retrieval signals as structured indicators of verifiability rather than as direct evidence for truth assessment. For each selected snippet, the LLM receives i) its textual content, ii) the computed relevance score, iii) the source credibility score, and iv) metadata such

as the domain. Providing these signals explicitly allows the model to evaluate not only the retrieved information but also its reliability and alignment with the claim. Hence, instead of assessing whether the claim is formulated in a way that can be objectively checked given the availability of relevant and credible external information, the decision thus depends on the quality and reliability of potential evidence, aligning RAVE with the goal of detecting verifiable claims.

3. EXPERIMENTS

3.1. Datasets and Baselines

We evaluate on two benchmarks: CT22-test and PoliClaim-test. CT22-test comprises longer, self-contained COVID-19 tweets that are information-rich, with explicit references to entities, events, and time. In contrast, PoliClaim contains debate transcripts segmented into single-sentence units that are often short, ambiguous, and context-dependent, frequently relying on pronouns and rhetorical devices.

All methods use GPT-4o as the backbone LLM with the same prompt template; only the retrieval strategy varies. The baselines are designed to isolate the contribution of retrieval and scoring to verifiability detection.

- Text-only: The model determines the verifiability using only the input claim, without retrieved context. This serves as the standard baseline in prior claim detection work.
- Rand-K: A random set of k snippets is added as context, providing a lower bound that controls for the effect of extra text without using relevance.
- Search-K: The top-k snippets are selected according to the search engine's default ranking, testing the impact of naive search-based retrieval.
- RAVE-Stats: The model receives only aggregated retrieval statistics, such as entity coverage, snippet coverage, source diversity, and inter-snippet agreement, to assess whether statistical indicators alone can determine verifiability.
- RAVE-Meta: The model is given only snippet metadata, including domain, and associated relevance and credibility scores, without snippet text. This evaluates whether metadata and scoring signals are sufficient.
- RAVE (ours): Our full method, which combines snippet content with relevance and credibility scores, enabling the model to use both textual evidence and structured signals.

3.2. Experiment Setup

For reproducibility, all experiments use GPT-4o-08-06 with temperature set to 0, top-p to 1, and a maximum token length of 500. All baselines and our proposed method follow the same prompt template, differing only in input content and decision strategy, while outputs are consistently formatted in JSON. Performance is evaluated primarily using accuracy and F1-score, with precision and recall reported as supporting metrics. Results are presented for both CT22-test and PoliClaim-test.

Method	Acc.	Prec.	Rec.	F1
CT22-test				
Text-only	0.8207	**0.8824**	0.8054	0.8421
Rand-K	0.8048	0.8333	0.8389	0.8361
Search-K	0.8287	0.8487	**0.8658**	**0.8571**
RAVE-Stats	0.8088	0.8389	0.8389	0.8389
RAVE-Meta	0.8088	0.8435	0.8322	0.8378
RAVE	**0.8327**	0.8690	0.8456	**0.8571**
PoliClaim-test				
Text-only	0.6789	**1.0000**	0.4971	0.6641
Rand-K	0.6998	**1.0000**	0.5298	0.6926
Search-K	0.6985	**1.0000**	0.5278	0.6910
RAVE-Stats	0.6446	**1.0000**	0.4434	0.6144
RAVE-Meta	**0.7010**	**1.0000**	0.5317	0.6942
RAVE	**0.7010**	0.9964	**0.5336**	**0.6950**

Table 2. Results on verifiable claim detection. All methods use GPT-4o as the backbone LLM with fixed K=3; only the retrieval strategy differs. Best scores are highlighted in **bold**.

4. RESULTS AND DISCUSSION

4.1. Main Results

As mentioned in Section 3.1 and shown in Table 2, performance is generally higher on CT22-test, where retrieval provides sufficient context and all methods achieve balanced precision and recall. RAVE attains the best overall F1 by combining high precision with strong recall. PoliClaim-test is more challenging because of its higher entity sparsity. While only 4.7% of verifiable and 24.5% of non-verifiable claims in CT22 lack entities, the proportions in PoliClaim rise to 30.3% and 81.0%. This imbalance leads LLMs to default to the NON-VERIFIABLE label, producing very high precision but low recall for baselines. By incorporating source credibility, RAVE reduces this bias, improves recall while preserving precision, and achieves the highest overall F1. These results demonstrate RAVE's robustness across datasets with different entity densities and contextual richness.

4.2. Ablation Study

4.2.1. Effectiveness of Scoring Components

We conduct an ablation study on CT22-test with K=3, comparing relevance, credibility, raw text, and their combination in RAVE (Table 3). Relevance alone yields high precision (83.9%) but lower F1, as retrieved evidence may not always be reliable. Credibility achieves the highest recall (85.9%) but lower precision, since trustworthy sources are not always directly relevant. Using only input text and snippets remains

Variant	Acc	Prec	Rec	F1	ΔF1
Relevance only	80.9	83.9	83.9	83.9	-1.8
Credibility only	82.5	84.8	**85.9**	85.3	-0.4
Text+Snippets	82.1	85.1	84.6	84.9	-0.8
RAVE	**83.3**	**86.9**	84.6	**85.7**	**+0.0**

Table 3. Ablation study of scoring components on CT22-test under fixed K=3 with identical prompts. We compare models using only *relevance*, only *credibility*, no scores, and their combination (**RAVE**). Best results are highlighted in **bold**.

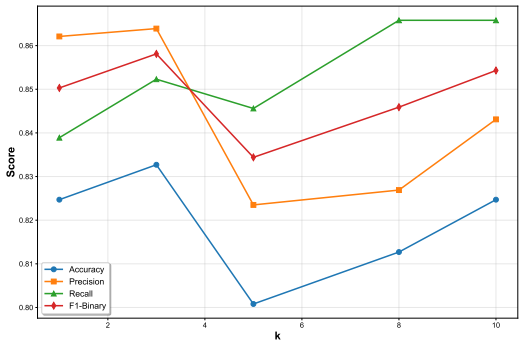

Fig. 1. Sensitivity of model performance to the number of retrieved snippets (K) on CT22-test.

competitive (F1 84.9%), but explicit scoring improves stability. RAVE, which combines relevance and credibility, achieves the best overall results (accuracy 83.3%, precision 86.9%, F1 85.7%), showing that the two signals are complementary and most effective when used together.

4.2.2. K-Sensitivity

We vary the number of retrieved snippets $K \in \{1, 3, 5, 8, 10\}$ for RAVE (GPT-4o) on CT22-test. As shown in Fig. 1, recall increases as K grows, while precision declines slightly, producing a shallow U-shaped F1 curve. The best performance occurs at $K = 3$ (F1=85.7%), with comparable results at $K = 10$ (85.4%). The dip at $K = 5$ indicates added redundancy and noise, though the decline is modest due to the scoring signals in the prompt. For efficiency, we set $K = 3$ for all subsequent experiments.

4.2.3. Generalization to Open-Source LLMs

To verify that improvements are not dominated by GPT-4o's parametric knowledge, we additionally evaluate our framework with Llama-3B, a smaller open-source LLM. On CT22-test, F1 improves from 71.48% (Text-only) to 73.51% (Search-K) and 75.25% (RAVE). This confirms consistent retrieval gains with reduced parametric knowledge, supporting generalizability beyond GPT-4o.

4.3. Error Analysis

To further understand the performance of RAVE, we conducted a qualitative error analysis using representative examples from the CT22 and PoliClaim datasets. The results reveal distinct error patterns shaped by the linguistic features of each dataset and the challenges of verifiable claim detection.

On CT22, the system achieved balanced precision and recall, yielding higher overall accuracy (83.3%). Errors were more evenly distributed:

- False positives often involved non-verifiable claims expressed with factual phrasing. For example, "As president, I will: - Implement nationwide mask mandates - Ensure access to regular, reliable, and free testing - Accelerate the development of treatments and vaccines I won't waste any time getting this virus under control" lists specific actions but describes intent rather than verifiable facts.
- False negatives frequently arose from sarcasm or mixed fact–opinion constructions. For instance, "KFC had 11 secret herbs and spices and nobody questioned it for years but suddenly you want to know what's in the Covid vaccine." contains a verifiable historical reference but was misclassified due to its sarcastic tone.

On PoliClaim, the system struggled more, reflected in lower recall for verifiable claims. Errors showed systematic difficulties with political discourse:

- The only false positive came from rhetorical speech, e.g., "After everything we've faced over the past three years, it's my honor to report that not only is the State of our State Resilient, we're fulfilling our motto of 'North to the Future'." Although symbolic, the model misinterpreted it as fact-based.
- Most false negatives involved broad generalizations lacking specificity. Statements like "When our prison system went unaddressed for decades and resulted in serious challenges, we found a way toward a solution." and "When our roads and bridges were in need for desperate improvements, we found a way to make significant progress all across the state." describe verifiable conditions but were treated as rhetorical because they lacked concrete details such as dates or entities.

5. CONCLUSION

In this work, we propose RAVE, a retrieval and scoring aware framework for verifiable claim detection that integrates external evidence with relevance and credibility signals. Experiments on CT22 and PoliClaim show consistent improvements over text-only and retrieval-based baselines in both accuracy and F1. While performance on PoliClaim is about 10% lower due to ambiguity in debate transcripts, future work will focus on domain adaptation and refined credibility modeling to enhance robustness across diverse datasets.

6. REFERENCES

[1] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata, and R. Procter, "Detection and resolution of rumours in social media: A survey," *Acm Computing Surveys (Csur)*, vol. 51, no. 2, pp. 1–36, 2018.

[2] B. D. Oladokun, J. E. Aruwa, G. A. Ottah, and Y. A. Ajani, "Misinformation and disinformation in the era of social media: The need for fact-checking skills," *Journal of Information and Knowledge*, pp. 1–7, 2024.

[3] P. R. Johnson, "A case of claims and facts: Automated fact-checking the future of journalism's authority," *Digital Journalism*, vol. 12, no. 10, pp. 1461–1484, 2024.

[4] Y. S. Kartal and M. Kutlu, "Re-Think Before You Share: A Comprehensive Study on Prioritizing Check-Worthy Claims," *IEEE Transactions on Computational Social Systems*, vol. 10, no. 1, pp. 362–375, 2023.

[5] R. Panchendrarajan and A. Zubiaga, "Claim detection for automated fact-checking: A survey on monolingual, multilingual and cross-lingual research," *Natural Language Processing Journal*, vol. 7, p. 100066, 2024.

[6] L. Konstantinovskiy, O. Price, M. Babakar, and A. Zubiaga, "Toward Automated Factchecking: Developing an Annotation Schema and Benchmark for Consistent Automated Claim Detection," *Digital Threats: Research and Practice*, vol. 2, no. 2, pp. 1–16, 2021.

[7] J. Ni, M. Shi, D. Stammbach, M. Sachan, E. Ash, and M. Leippold, "Afacta: Assisting the annotation of factual claim detection with reliable llm annotators," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 1890–1912.

[8] N. Hassan, G. Zhang, F. Arslan, J. Caraballo, D. Jimenez, S. Gawsane, S. Hasan, M. Joseph, A. Kulkarni, A. K. Nayak, V. Sable, C. Li, and M. Tremayne, "ClaimBuster: The first-ever end-to-end fact-checking system," *Proceedings of the VLDB Endowment*, vol. 10, no. 12, pp. 1945–1948, 2017.

[9] I. Jaradat, P. Gencheva, A. Barron-Cedeno, L. Marquez, P. Nakov *et al.*, "Claimrank: Detecting check-worthy claims in arabic and english," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 2018, pp. 26–30.

[10] R. Dhar, S. Dutta, and D. Das, "A hybrid model to rank sentences for check-worthiness." in *CLEF (Working Notes)*, 2019.

[11] E. Williams, P. Rodrigues, and V. Novak, "Accenture at checkthat! 2020: If you say so: Post-hoc fact-checking of claims using transformer-based models," in *CLEF (Working Notes)*, 2020.

[12] M. Sawiński, K. Węcel, E. P. Księżniak, M. Stróżyna, W. Lewoniewski, P. Stolarski, and W. Abramowicz, "Openfact at checkthat! 2023: head-to-head gpt vs. bert-a comparative study of transformers language models for the detection of check-worthy claims," *Working Notes of CLEF*, 2023.

[13] Y. Li, R. Panchendrarajan, and A. Zubiaga, "Factfinders at checkthat! 2024: Refining check-worthy statement detection with llms through data pruning," in *CLEF (Working Notes)*, 2024.

[14] F. Alam, A. Barrón-Cedeño, G. S. Cheema, G. K. Shahi, S. Hakimov, M. Hasanain, C. Li, R. Míguez, H. Mubarak, W. Zaghouani *et al.*, "Overview of the clef-2023 checkthat! lab task 1 on check-worthiness in multimodal and multigenre content." in *CLEF (Working Notes)*, 2023, pp. 219–235.

[15] S. Shaar, M. Hasanain, B. Hamdan, Z. S. Ali, F. Haouari, A. Nikolov, M. Kutlu, Y. S. Kartal, F. Alam, G. Da San Martino *et al.*, "Overview of the clef-2021 checkthat! lab task 1 on check-worthiness estimation in tweets and political debates." in *CLEF (working notes)*, 2021, pp. 369–392.

[16] P. Nakov, A. Barrón-Cedeño, G. Da San Martino, F. Alam, R. Míguez, T. Caselli, M. Kutlu, W. Zaghouani, C. Li, S. Shaar *et al.*, "Overview of the clef-2022 checkthat! lab task 1 on identifying relevant claims in tweets," in *2022 Conference and Labs of the Evaluation Forum, CLEF 2022*. CEUR Workshop Proceedings (CEUR-WS. org), 2022, pp. 368–392.

[17] G. Pasi and M. Viviani, "Information credibility in the social web: Contexts, approaches, and open issues," *arXiv preprint arXiv:2001.09473*, 2020.

[18] D. Westerman, P. R. Spence, and B. Van Der Heide, "Social media as information source: Recency of updates and credibility of information," *Journal of computer-mediated communication*, vol. 19, no. 2, pp. 171–183, 2014.