WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal Homepage: www.ijrpr.com ISSN 2582-7421

Claim Verification with Tiered Trusted Sources: Why Style-Based Fake News Classifiers Fail to Generalize

Ananya Singh Baghel, Tanish Kumar Sahu, Shweta Dubey*

(UG students), Ananya Singh Baghel and Tanish Kumar Sahu
*(Assistant Professor), Department of CST, Shri Shankaracharya Professional University, Bhilai

ABSTRACT

Many fake news detectors are text classifiers trained on labelled news articles. They score very high on their own test data, but we suspected that they learn how a publisher writes and not whether a statement is true. This paper tests that idea on LIAR, a set of short political claims, and compares the classifier with a claim-checking pipeline that uses web evidence.

A Random Forest trained on the ISOT dataset reaches 99.76% on ISOT, yet it scores only 44.5% on 200 LIAR claims and calls every one of them fake. The pipeline extracts claims with a large language model (LLM), searches the web with three query variants, keeps only pages from a list of trusted domains, and gives a verdict only when two independent trusted domains agree. We compare it with a plain LLM and with an LLM that uses raw web search.

The plain LLM was the most accurate system (70.5%). The pipeline reached 65.4% on the 52 claims it answered (26.0% coverage), while the plain LLM scored 78.8% on those same 52 claims. The extra query variants raised coverage from 17.0% to 26.0%, but accuracy did not change in a measurable way. We conclude that style classifiers do not carry over to new data, and that a strict source rule gives checkable evidence and caution, not better accuracy than a strong LLM.

Keywords: misinformation detection, claim verification, evidence retrieval, source credibility, ISOT, LIAR

1. Introduction

Fake news spreads quickly and manual fact-checking cannot keep up, so many tools try to detect it automatically [1], [2]. The usual method is to train a classifier on a collection of real and fake articles. Such a model often gets a very high score on its own test set. The problem is that the score may come from small cues, like agency tags, formatting and writing style, and not from the facts. In the ISOT corpus [3], most real articles come from one news agency and carry its location and agency tags. The fake ones come from other websites. A model can separate these two groups without knowing anything about the truth.

Another way is claim verification, where each claim is checked against evidence found elsewhere [4], [5], [6]. Large language models can do this with web search, but they can also give a confident answer with no support. So a useful tool should show where its answer comes from, and it should say so when it cannot find enough evidence.

We became interested in this while building a misinformation checker as our college project. Our first version used an ISOT classifier and it looked excellent on paper. When we tried real news from outside ISOT, it failed. This paper reports what we found. We built a claim-level pipeline with trusted source tiers and a two-domain rule, and we compared it with the classifier and with two LLM baselines on 200 claims from LIAR [7].

The rest of the paper is organised as follows. Section 2 reviews related work. Section 3 explains the methodology, the experimental setup and the evaluation metrics. Section 4 describes the system architecture and the technologies used, and Section 5 walks through the working of the algorithm. Sections 6 and 7 cover the implementation, the results and the limitations, and Sections 8 and 9 give the future scope and the conclusion.

2. Related Work

Shu et al. [2] survey fake news detection that uses article content and social context, and FakeNewsNet [8] is a data repository for such work. FEVER [4] moved the field toward evidence-based checking by pairing claims with Wikipedia sentences. AVeriTeC [5] uses real claims and evidence from the open web, which is closer to our setting. Surveys on automated fact-checking [6], [9] describe the common steps: find claims, retrieve evidence, predict a verdict.

Retrieval-augmented generation [10] gives a language model retrieved documents to base its answer on, and our verdict step follows the same idea. LIAR [7] contains short political statements with six truth labels. We use it as data from outside the ISOT domain, so that a classifier trained on ISOT is tested on text it has never seen in style.

3. Methodology

The system follows the four stages shown in Fig. 1. The input is plain text or an article URL. In the first stage an LLM splits the text into short claims. In the second stage each claim is searched on the web with three query variants. In the third stage the results are filtered by a list of trusted domains, and in the fourth stage the LLM gives a verdict that is accepted only if two independent trusted domains support it.

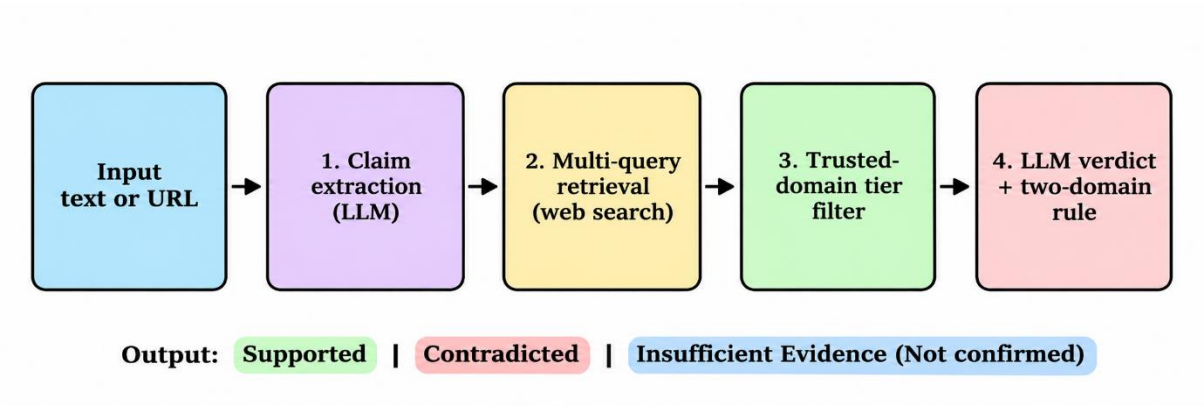


Fig. 1. Pipeline of the claim verification system.

3.1 Claim extraction and verdicts

For a URL, the article text is downloaded first. An LLM (Gemini 3.5 Flash-Lite) then splits the text into short claims, and each claim is checked separately. A search API (Tavily) returns web pages for the claim, and the LLM reads the pages that remain after filtering. It answers Supported, Contradicted or Insufficient Evidence. In the web interface these appear as Likely True, Likely False and Not confirmed, together with the source links.

3.2 Multi-query retrieval

A claim searched exactly as written often brings back poor results, especially when it is short or vague. So we run three searches for each claim: the claim itself, a keyword query made from the main names, the event and the date, and the same keyword query with the word news added. The three result lists are merged and repeated URLs are removed. The keyword query is built by a simple rule, so it costs no extra LLM call.

3.3 Source tiers and the corroboration rule

We keep a list of trusted domains in two tiers. Tier 1 holds official and wire sources such as government, international body and agency sites. Tier 2 holds major news outlets. The system may answer Supported or Contradicted only if at least two independent trusted domains give evidence. With fewer than two, it answers Insufficient Evidence. We added this rule so that a single website can never decide a verdict alone. Two pages from the same outlet count as one domain.

3.4 Experimental setup

We took 200 claims from LIAR [7] with a fixed random seed. After converting the labels to true or false, 111 claims are true and 89 are false. PolitiFact pages were blocked from the search results, because LIAR labels come from PolitiFact and the answer would leak. We compare five systems. The first is the ISOT classifier, a TF-IDF Random Forest trained on ISOT [3]. System A is a plain LLM with no search. System B is an LLM with raw web search, without the trusted list and without the two-domain rule. System C is our full pipeline with one query per claim. System D is the full pipeline with three queries per claim. Systems B, C and D can abstain (Insufficient Evidence or Mixed), and an abstained claim counts as not answered. All 200 claims were run with the same code, trusted domain list and seed.

3.5 Evaluation metrics

Let N be the number of claims, n_{ans} the number of answered claims and $n_{correct}$ the number of answered claims with the right label. We report coverage and accuracy on answered claims:

$$coverage = n_{ans} / N \quad (1)$$

$$acc_{ans} = n_{correct} / n_{ans} \quad (2)$$

We also report overall accuracy, where an abstained claim counts as wrong, macro-F1 on answered claims, and the average time per claim. Because some systems answer only a few claims, we give a 95% Wilson interval [11] for every accuracy:

$$CI = [\ p + z^2/(2n) \pm z\ \sqrt{(\ p(1-p)/n + z^2/(4n^2)\ )}\ ]\ /\ (\ 1 + z^2/n\ ) \quad (3)$$

Here p is the observed accuracy, n is the number of answered claims and z = 1.96.

4. System Architecture

The system has a web interface and a backend service. The interface is built with React. It sends the text or URL to the backend, shows a five-step loading view while the check runs, and then shows an overall result card, one card for each claim with its verdict and source links, and a short explanation of how the check was done. The backend is written in Python with FastAPI and controls the whole pipeline. It calls the LLM and the search API, and it keeps a SQLite cache so that repeated runs do not call the services again. The technologies are summarised in Table I.

TABLE I. TECHNOLOGIES USED

Layer	Technologies	Responsibility
Web interface	React	Input box, loading steps, verdict cards, overall result, source links
Backend	FastAPI (Python)	API endpoints and control of the pipeline
LLM	Gemini 3.5 Flash-Lite	Claim extraction and verdict from evidence
Search and text	Tavily API, trafilatura	Web search, and article text from URLs
Storage	SQLite	Cache of search and model results
Baseline	scikit-learn	TF-IDF and Random Forest trained on ISOT

5. Algorithm and Working

The checking of one input can be written as the following steps:

1. Read the input. If it is a URL, download the article text.
2. Ask the LLM to split the text into short claims.
3. For every claim, search the web with the claim itself, with a keyword query, and with the keyword query plus the word news.
4. Merge the results, remove repeated URLs, and keep only pages from Tier 1 or Tier 2 domains.
5. Let the LLM read the kept pages and give a verdict. If fewer than two independent domains support it, change the answer to Insufficient Evidence and give the reason.

The interface then collects the verdicts. If any claim is Likely False, the overall result says that the text contains likely false claims. If at least half of the claims are Likely True and none is false, it says mostly supported. Otherwise it says mostly unverified. No percentage is shown, because the system has no calibrated probability. The demo system also runs a fourth search limited to the trusted domains, which was added after the evaluation.

6. Implementation

The following modules have been implemented and are working:

- Claim extraction from pasted text and from article URLs.
- Multi-query web retrieval with merging of results and removal of repeated URLs.
- A trusted domain list with two tiers and the two-domain rule.
- Verdicts with source links, and a reason when a claim is not confirmed.
- A web interface with a loading view, verdict cards, an overall result card and an explanation of the checking steps.

- An evaluation script that runs all systems on the 200 claims and saves per-claim results, a summary file and confusion matrices.

7. Results and Limitations

7.1 Main results

Table II gives the main numbers, Fig. 2 shows coverage and accuracy together, and Table III gives the confusion counts. The ISOT classifier called all 200 claims fake. Its accuracy of 44.5% is lower than the 55.5% we would get by always answering true. Since it is almost perfect on ISOT itself, it has clearly learned the style of its training sources and not the truth of a claim.

The plain LLM did best, with 70.5% on all 200 claims. Raw web search (B) did not help. It answered 68.5% of the claims and was right on 65.7% of them. The trusted list and the two-domain rule (C) cut coverage to 17.0%, with 64.7% accuracy on the 34 answered claims. With three queries (D), the system answered 52 claims (26.0%), which is 52.9% more than C, and accuracy on those was 65.4%. This came with a time cost: 5.83 s per claim for D and 2.13 s for C, because D runs three searches.

The confidence intervals of B, C and D overlap a lot with each other and with A, so we cannot say that any of these systems is truly more accurate than another. Only two things are clear in our data: the ISOT classifier fails, and D covers more claims than C. Because C and D answer only some claims, we also scored A on the same 52 claims that D answered. There A reaches 78.8% (41 of 52, CI 66.0 to 87.8), which is higher than the 65.4% of D. On the 148 claims where D abstained, A scores 67.6%. So the plain LLM is better even on the claims that our pipeline felt sure enough to answer.

TABLE II. COMPARISON OF SYSTEMS ON 200 LIAR CLAIMS

System	Answered	Coverage (%)	Acc. ans. (%)	95% CI	Overall acc. (%)	Macro-F1	Time (s)
ISOT classifier	200	100.0	44.5	37.8–51.4	44.5	0.308	0.03
A: Plain LLM	200	100.0	70.5	63.8–76.4	70.5	0.704	1.84
A on D's 52 claims	52	26.0	78.8	66.0–87.8	–	0.788	1.84
B: LLM + search	137	68.5	65.7	57.4–73.1	45.0	0.652	5.66
C: Pipeline, 1 query	34	17.0	64.7	47.9–78.5	11.0	0.642	2.13
D: Pipeline, 3 queries	52	26.0	65.4	51.8–76.8	17.0	0.653	5.83

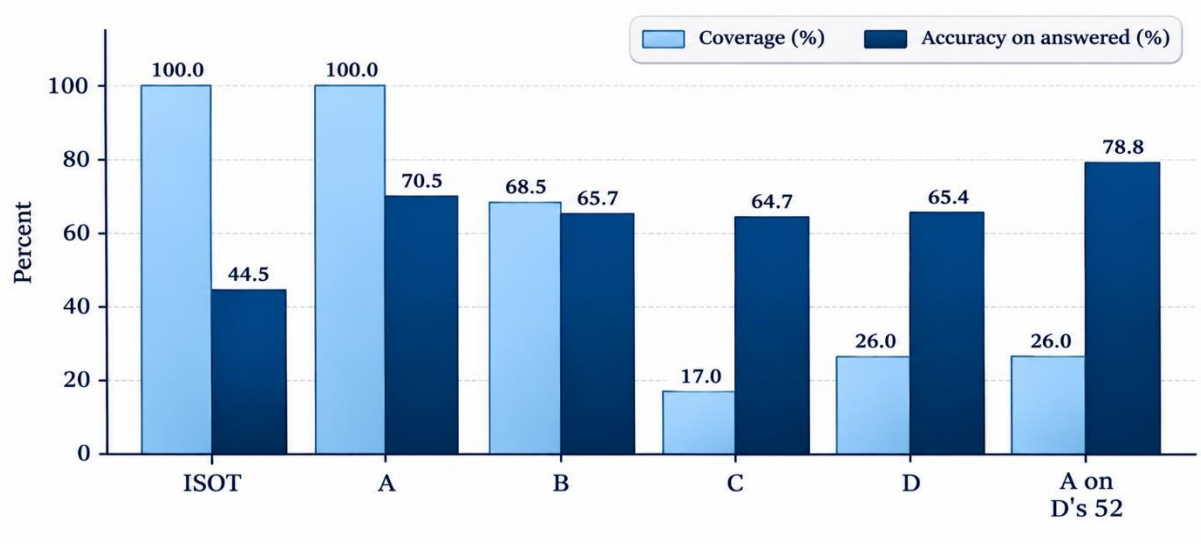


Fig. 2. Coverage and accuracy on answered claims for each system.

TABLE III. CONFUSION COUNTS ON ANSWERED CLAIMS (POSITIVE = TRUE CLAIM)

System	TN	FP	FN	TP	Answered
ISOT classifier	89	0	111	0	200
A: Plain LLM	75	14	45	66	200
B: LLM + search	37	23	24	53	137
C: Pipeline, 1 query	9	8	4	13	34
D: Pipeline, 3 queries	16	10	8	18	52

7.2 Demo cases

We also tried the system by hand on a few inputs from outside LIAR. First, the ISOT classifier labelled a real article from an Indian newspaper as fake with 99% confidence, which matches its poor result on LIAR. Second, a claim about a shooting outside a film actor's home in Mumbai in April 2024 first returned Insufficient Evidence, because Indian news sites were not on our trusted list. After we added 14 outlets, it still found only one trusted domain. With the three query variants it found two (a Times of India report and a BBC report) and the verdict became Supported. Third, the false claim that the Great Wall of China can be seen from the Moon with the naked eye was marked Contradicted, using a university page and Snopes. A true claim about the date WHO declared COVID-19 a pandemic was marked Supported, using WHO and NIH pages.

7.3 Error analysis

We read through the per-claim output and found four common kinds of mistakes in the full pipeline:

- Sources about the topic but not about the claim. A false claim about a government science adviser was marked Supported because the search returned his official biography pages. A false claim about trade surpluses was marked Supported because the pages were lists of trade agreements.
- Missing evidence read as contradiction. In a demo run on a news article, a claim about a 52-week oil price high was marked Likely False because the retrieved pages showed only a lower recent price. They did not state a different value for that period, so Not confirmed would have been the better answer.
- Noise from the extra queries. Some variants returned pages that have nothing to do with the claim, such as pages about other places or about travel rates.
- Claims that cannot be checked. Many LIAR statements are opinions, predictions or just a topic phrase. For these, abstaining is the right answer.

7.4 Limitations

Several limitations should be kept in mind.

- C and D answered only 34 and 52 claims, so their confidence intervals are about 25 to 30 points wide.
- LIAR claims may be in the training data of the LLM, which could help A even on the same 52 claims.
- After the evaluation, a fourth search pass limited to trusted domains was added to the demo system. It was not part of the evaluation.
- We used binary labels instead of LIAR's six classes, one dataset and one LLM.
- Search results change over time, so a rerun may give slightly different numbers.
- The system checks claims, not whole articles. For a long article it may pick background facts and miss the main event.

8. Future Scope

The first step is to test the differences statistically on a larger set of claims and to compare all systems on the same claims. Next we plan to filter out irrelevant pages before the LLM reads them, to choose better claims from long articles, and to treat missing evidence differently from contradicting evidence. We also want to test on newer claims, where the LLM has seen less of the data and retrieval is more likely to find sources, and to try embedding-based retrieval in place of keyword queries.

9. Conclusion

We compared a style-based classifier with evidence-based claim checking on 200 LIAR claims. The classifier trained on ISOT did not carry over. It reached 44.5% and called every claim fake, although it is nearly perfect on ISOT. A high score on one benchmark tells us very little about real fact-checking.

Our pipeline gave verdicts with source links, and it abstained when it could not find two trusted domains. It was not more accurate than a plain LLM (65.4% on answered claims against 70.5% overall and 78.8% on the same 52 claims), and its coverage was only 26.0%. The extra queries raised coverage from 17.0% to 26.0% but made each check slower. So the real value of the two-domain rule is caution and transparency, not accuracy.

References

1. Vlachos and S. Riedel, "Fact checking: Task definition and dataset construction," in *Proc. ACL Workshop on Language Technologies and Computational Social Science*, 2014.
2. K. Shu, A. Sliva, S. Wang, J. Tang and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, pp. 22–36, 2017.
3. H. Ahmed, I. Traore and S. Saad, "Detection of online fake news using N-gram analysis and machine learning techniques," in *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments (ISDDC)*, Springer, 2017.
4. J. Thorne, A. Vlachos, C. Christodoulopoulos and A. Mittal, "FEVER: a large-scale dataset for fact extraction and VERification," in *Proc. NAACL-HLT*, 2018.
5. M. Schlichtkrull, Z. Guo and A. Vlachos, "AVeriTeC: A dataset for real-world claim verification with evidence from the web," in *NeurIPS Datasets and Benchmarks Track*, 2023.
6. Z. Guo, M. Schlichtkrull and A. Vlachos, "A survey on automated fact-checking," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 178–206, 2022.
7. W. Y. Wang, "'Liar, liar pants on fire': A new benchmark dataset for fake news detection," in *Proc. 55th Annual Meeting of the ACL (Short Papers)*, 2017.
8. K. Shu, D. Mahudeswaran, S. Wang, D. Lee and H. Liu, "FakeNewsNet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media," *Big Data*, vol. 8, no. 3, 2020.
9. P. Nakov et al., "Automated fact-checking for assisting human fact-checkers," in *Proc. IJCAI*, 2021.
10. P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
11. E. B. Wilson, "Probable inference, the law of succession, and statistical inference," *Journal of the American Statistical Association*, vol. 22, no. 158, pp. 209–212, 1927.