Explainable Automatic Claim Detection for Journalists Augmentation

Abstract

STILL NEEDS TO BE DONE We propose an Explainable ACD framework that is explicitly grounded in empirically studied journalistic cognition. A qualitative user study with professional fact-checkers validates the artifacts and confirms and is evaluated as a human-centred decision-support tool through qualitative studies with professional fact checkers.

1 Introduction and Motivation

Professional fact-checking has emerged as a critical socio-technical response to the large-scale circulation of misleading and harmful claims in digital media. It has been shown that the median time for human fact-checkers to produce explanation articles debunking disinformation narratives is 4.5 days, whereas such narratives reach peak virality in just 1.5 days [Wack *et al.*, 2024]. Polarization in social networks is formed when disinformation is left unaddressed for a given period [Altoe *et al.*, 2024]. Therefore, by the time fact-checkers publish their work, polarization has likely set in, making many individuals less open to their corrective information. Automatic Fact-Checking (AFC) shows potential to be a valuable tool for journalists, as it promises substantial time savings in the journalistic workflow. AFC executes two high-level tasks, Automatic Claim Detection (ACD) and Automatic Claim Verification (ACV). ACD is the process of automatically detecting claims that are worth fact-checking, and ACV is centered on the evidence retrieval, multi-level claim classification at various levels of "fakeness", and ultimately, explanation generation. Historically, AFC research has focused almost exclusively on the AI/ML techniques that surround the two tasks. Explainable AFC [Tan *et al.*, 2025] has gained momentum. However, explainability in this context has primarily focused on improving the human readability of classification models' decisions. Recently, a critical step was taken toward the purposeful design of an explainable ACV framework that closely mimics the tasks performed by professional fact-checkers and, consequently, augments journalists in real-world scenarios [Altoe *et al.*, 2025].

On the ACD front, while recent advances in Large Language Models (LLMs) have produced increasingly accurate systems for claim checkworthiness classification [Majer and Šnajder, 2024] and automatic dataset annotation [Ni *et al.*, 2024], many such approaches remain poorly aligned with how human fact checkers actually reason about which claims warrant verification in real-world newsroom contexts. As a result, existing systems often optimize technical metrics while failing to support the cognitive, ethical, and organizational decision-making processes that define professional fact-checking practice. Empirical research in journalism studies has shown that claim selection is not a purely computational problem, but a multi-stage human cognitive process [Soprano *et al.*, 2024] involving sequential judgments under time [Dierickx and Lindén, 2023], information, and attention constraints [Markowitz *et al.*, 2023]. Large-scale interview and content analysis studies demonstrate that professional fact checkers first apply a set of pre-conditioning criteria—such as checkability, practical verifiability, and virality [Cazzamatta, 2025], before making more nuanced judgments about potential harm, relevance to public debate, source prominence, and audience demand [Sehat *et al.*, 2024]. These decision processes reflect bounded rationality, heuristic filtering, and harm-oriented reasoning, shaped by institutional norms, ethical responsibilities, and real-world constraints. From the viewpoint of fact-checking organizations, it has been suggested that organizations should be more transparent in articulating to the public the criteria used to determine whether a claim falls within the scope of checkworthiness [Suomalainen *et al.*, 2025]. This adds yet another cognitive load dimension for the professional journalist, who is already under enormous pressure to sift through the sea of social media noise to identify claims worth verifying.

A literature review combined with a fact-checker-focused user study showed that no existing pipeline offers end-to-end support to fact-checkers, only solutions focusing on a single step from the select claims to check -> search for evidence -> decide on claim veracity verdict -> create -> verdict explanation journalistic workflow [Sawiński *et al.*, 2024]. This is indeed the case. The ACD research field focuses on the "checkwothiness" of the claim, assuming the claim has already been selected as important enough to check. There is a fundamental problem with that approach: journalists typically rely on virality signals to select which claim to check. In practice, that means the polarization clock has already started, as stated in [Altoe *et al.*, 2024]. In the digital media use case, the work focuses on anomaly detection and virality predic-

tion, but most approaches operate at the level of individual social media posts [Jin *et al.*, 2024]. We argue that work is needed to extend anomaly detection and virality prediction to the claim level. In practical terms, this means real-time streaming and normalization of posts, clustering of all posts that reference the same claim, followed by temporal analysis of the claims, leading to automatic claim virality prediction and the ability to signal a checkworthiness analysis. In this paper, we argue that next-generation ACD systems should not only include claim-level claim anomaly detection and virality prediction but also be designed as human-centred cognitive support tools, rather than as autonomous classifiers. Such ACD systems can be combined with human-centered ACV systems, as presented in [Altoe *et al.*, 2025], to address this gap. To this end, we introduce an Explainable ACD Framework explicitly grounded in empirically validated journalistic cognition. The proposed framework operationalizes key human decision criteria as modular, machine-interpretable components that address checkability, verifiability, virality dynamics, and harm potential, purposefully aligning its design with the pre-conditioning criteria used by professional fact-checkers to select claims for verification. Its explainability is intended to align with professional reasoning practices.

Feature-level explanations are often faithful but hard for non-technical users to act on, whereas high-level textual rationales are user-friendly but difficult to validate rigorously [Athira *et al.*, 2023]. Crucially, our approach treats explainability not as a post-hoc technical add-on, but as an interactive cognitive interface between AI systems and human decision-makers. We evaluate the framework not only against established computational benchmarks but also through qualitative interviews with professional journalists, examining how generated explanations support trust, accountability, and informed decision-making in naturalistic newsroom settings. By integrating computational methods with empirical insights into human judgment, ethics, and professional practice, this work contributes to the design of responsible, transparent, and human-centred AI systems for digital media and disinformation mitigation.

This work makes the following human-centred AI contributions:

1. Human-driven AI system design: To the best of our knowledge, our work is the first to translate journalistic cognition into a modular AI architecture, where each component corresponds to a specific human judgment (checkability, verifiability, virality, harm potential), ensuring conceptual transparency and interpretability by design.

2. Explainability as a cognitive interface: We conceptualize explainability artifacts as human-facing cognitive scaffolds that support accountability, trust, and professional judgment, rather than as purely technical feature attributions. The report focuses on reducing the cognitive load on professional journalists needed to meet the growing transparency requirements imposed on fact-checking organizations.

3. Empirical human-in-the-loop evaluation: Few studies evaluate whether explanations actually help journalists or fact-checkers prioritize claims better or trust the system appropriately, despite growing work on evaluating explanation usefulness in AI-assisted fact-checking interfaces [Kotonya and Toni, 2020]. We conduct journalist-based qualitative evaluations to assess how explanations and classifications align with professional reasoning, expectations, and ethical standards in naturalistic newsroom contexts.

Additionally, we offer the following technical contributions to the ACD research community:

1. Claim level anomaly detection and virality prediction: To the best of our knowledge, our work is the first to propose a real-time streaming-like anomaly detection and virality prediction approach for claims that naturally appear in the open, in high-volume digital media.

2. New ACD benchmarks: There is a lack of datasets where claim spans, check-worthiness labels, and the reasons why a claim is considered check [Guo *et al.*, 2022]. We offer a new dataset that combines streamed tweets and their corresponding time sequences, the clustered claims normalized from the tweet streams and their corresponding time sequences, and virality predictions, the model prediction percentages for the four key cognite decisions used by fact-checkers to make claim checkworthiness decisions, and the final checkworthiness classification for the claims. To the best of our knowledge, this is the first dataset specifically designed to support further research on digital media claim anomaly detection and checkworthiness classification.

ADD OTHERS HERE AS THEY BECOME APPARENT:

Section 2 includes related work in explainable AFC and claim-decomposition-based AFC. Section 3 describes the methods used to create and validate the proposed pipeline and the user study design. Section 4 presents the framework pipeline and the user study results. Section 5 offers a discussion of the findings, known limitations, and future work opportunities. Section 6 concludes the work.

2 Related Work

Since we propose extending the explainable ACD state of the art to include claim-level virality prediction, this section presents the state of the art for each of these two fields.

Anomaly Detection and Virality Prediction: Typical research in the topic covers post-level prediction, which we have shown does not apply to our use case. Recently, the community has recognized the potential benefits and applications of event detection in social media data [Karimiziarani, 2022]. This encompasses anomaly detection from a topical perspective, followed by real-time clustering of similar topics to enable temporal analysis. Previous work has focused on predicting emerging events [Steuber *et al.*, 2023]. This is a first step toward predicting claim virality; however, an event is a superset of the claim set, as a given event may contain multiple claims that could be candidates for verification. Some works focus on disinformation. A diffusion study compares the sentiment on the virality of real news versus misinformation /citeking2023diffusion. A model has been proposed to explain the spread of false beliefs [Rabb *et al.*, 2022], whereas a study focused on the impact of content features on virality [Esteban-Bravo *et al.*, 2022]. A common thread among these disinformation-focused works is their focus on how disinformation propagates across digital media;

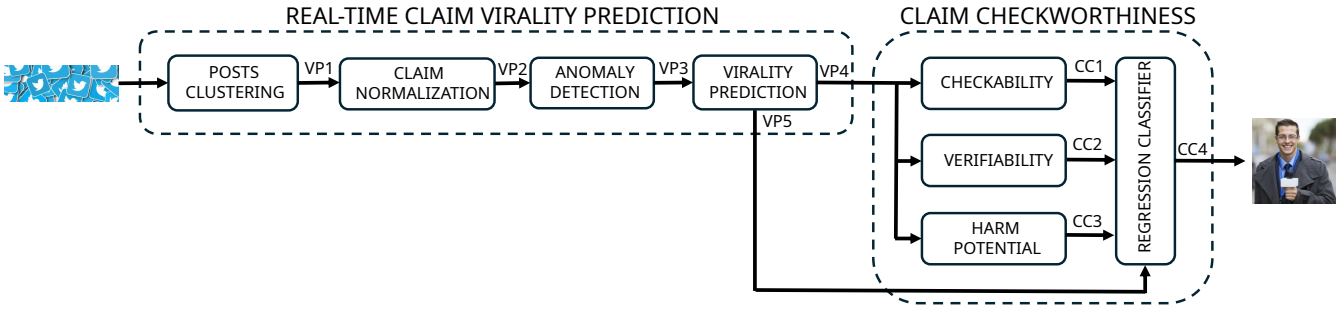

Figure 1: Complete Explainable ACD Pipeline (VP: Virality Prediction, CC: Claim Checkworthiness)

however, this does not extend to the intervention side of fake-news research.

Explainable Automatic Claim Detection: Recent systems for explainable fact-checking increasingly build multi-stage LLM pipelines that reason over claim–evidence correlations and produce explanations as a primary output rather than an afterthought [Tan *et al.*, 2025; Zhao *et al.*, 2024; Altoe *et al.*, 2025]. However, they focus on the ACF side of AFC rather than ACD. Typically, explainable ACD research combines check-worthiness prediction with interpretable components such as stance, evidence selection, and argument structure [Augenstein, 2021]. A notable recent line uses argumentation schemes and critical questions to make the identification of misinformation more interpretable at the argument level [Ruiz-Dolz and Lawrence, 2025]. However, it focuses the reasoning on the final classification of check-worthiness rather than on the intermediate steps leading to the verdict. We propose embedding high-level explanations of claim virality and checkworthiness within an end-to-end pipeline in which each step's decisions are inspectable. Lack of datasets is a clear issue in the field. Some approaches have begun using LLMs to annotate datasets [Ni *et al.*, 2024; Majer and Šnajder, 2024]. However, they primarily focus on claim annotation for claim checkworthiness classification, leaving the intermediate decisions unaddressed.

3 Methodology

This section covers the methodology followed for the framework pipeline design and the user study.

3.1 Framework Pipeline

This section details the two high-level components that together implement our proposed extended Explainable ACD pipeline: the real-time claim virality prediction (VP) and the claim checkworthiness classifier (CC). The VP module comprises four submodules: post-clustering, claim normalization, anomaly detection, and virality prediction. Each of the four modules is presented below.

Post Clustering: The goal of the posts clustering component is to group social media posts that express the same underlying *claim* into a shared cluster, despite lexical variation, informal language, and noise. Following prior work on claim detection and verification [Hassan *et al.*, 2017; Thorne *et al.*, 2018], we define a claim as an assertive statement that conveys propositional content and can, in principle, be verified.

Let $\mathcal{P} = \{p_1, p_2, \ldots, p_t\}$ denote a temporally ordered stream of posts, where each post p_i is associated with a timestamp τ_i and a short text sequence. The task is to assign each post to a claim cluster $c_j \in \mathcal{C}$ such that posts within the same cluster express semantically equivalent or closely related claims. The number of clusters is unknown a priori and may evolve over time. The clustering algorithm is designed to operate under the following constraints: (i) online processing without access to future data, (ii) bounded memory usage suitable for resource-constrained environments, and (iii) low-latency execution without reliance on large language models.

To support streaming operation and limit memory consumption, posts are processed in fixed-size temporal windows of length Δ minutes. Posts are sorted by timestamp and grouped into disjoint intervals $[\tau, \tau + \Delta)$. The clustering state is preserved across windows, enabling incremental processing of large corpora without loading the whole dataset into memory. Social media streams contain a high proportion of non-assertive content, such as questions, emotional expressions, and conversational replies. To reduce noise and computational overhead, we apply a lightweight claim-filtering step that removes posts that are unlikely to express a claim. The filter relies on simple surface-level cues (e.g., minimum length, absence of interrogative punctuation, presence of a finite verb) and is designed to be conservative, prioritizing recall over precision. Similar filtering assumptions are common in prior claim detection pipelines [Hassan *et al.*, 2017].

Each retained post is mapped to a dense semantic representation using a pretrained sentence embedding model. Formally, an embedding function f maps a post p_i to a vector $\mathbf{e}_i \in \mathbb{R}^d$. We use a compact transformer-based encoder optimized for efficiency, producing normalized embeddings suitable for cosine similarity. Embeddings are computed directly from the original post text, without claim rewriting or linguistic normalization, to maintain low latency and support real-time deployment. Clustering is performed using an incremental nearest-centroid strategy. Each cluster c_j is represented by a centroid vector $\boldsymbol{\mu}_j$ and a cluster size n_j. For an incoming post embedding $\mathbf{e}_i$, the algorithm identifies the nearest existing centroid $j^* = \arg\max_j \cos(\mathbf{e}_i, \boldsymbol{\mu}_j)$. If the maximum similarity exceeds a fixed threshold θ, the post is assigned to cluster c_{j^*} and the centroid is updated using an

incremental mean $\boldsymbol{\mu}_{j^*} \leftarrow \frac{n_{j^*}\boldsymbol{\mu}_{j^*}+\mathbf{e}_i}{n_{j^*}+1}$. Otherwise, a new cluster is created with $\mathbf{e}_i$ as its initial centroid. Nearest-centroid search is accelerated using a vector similarity index, ensuring efficient lookup as the number of clusters grows. To support long-running operations and incremental data ingestion, the clustering state (centroids, cluster sizes, and representative posts) is periodically serialized to disk in compressed form. Since only cluster-level statistics are stored, memory and storage requirements scale with the number of distinct claims rather than the volume of posts. Unlike batch clustering approaches commonly used in claim mining, such as k-means or density-based methods, the proposed approach is entirely online, does not require specifying the number of clusters in advance, and is suitable for real-time settings. By clustering posts directly in semantic embedding space, the method effectively groups paraphrased claims while remaining computationally efficient. The resulting claim clusters serve as the input to downstream components, including claim normalization, anomaly detection, and virality prediction.

Claim Normalization: Following posts clustering, the claim normalization module converts semantically grouped posts into a set of canonical, concise claim representations that are suitable for downstream verification and analysis. In the context of fact-checking and misinformation research, claim normalization is defined as the process of transforming noisy, unstructured social media text into simplified, unambiguous propositions that preserve the core assertion of the original statements [Sundriyal *et al.*, 2023].

Given a cluster of posts $\mathcal{C}_j = \{p_{j1}, p_{j2}, \dots\}$ expressing a similar assertion, the goal of normalization is to produce one or more normalized claims q_j that capture the essential factual content of $\mathcal{C}_j$. Normalized claims mitigate variability due to colloquial phrasing, slang, and orthographic noise, thereby improving interpretability and verifiability for automated and human fact-checkers [Sundriyal *et al.*, 2025]. In practice, normalization involves three substeps: (i) identification of claim spans within the text, (ii) extraction and simplification of propositional content, and (iii) rewriting into a canonical, grammatically coherent form. Recent work treats normalization as a dedicated task, in which complex, verbose assertions are rewritten into concise, standardized forms [Sundriyal *et al.*, 2023]. This often involves leveraging external models or heuristics to resolve referential ambiguity and prioritize verifiable content. In our pipeline, each cluster $\mathcal{C}_j$ is summarized by selecting representative posts and applying normalization to produce one or more canonical claims q_j. This step may be implemented using lightweight syntactic transformation rules, paraphrase generation techniques, or models trained specifically for claim simplification, to balance efficiency and semantic fidelity. By producing normalized claims, the system ensures that subsequent components, such as anomaly detection and veracity prediction, operate on a compact set of clear, propositionally meaningful inputs. Normalization thereby functions as a critical bridge between raw social media content and structured assertion representations used in truth assessment.

Anomaly Detection: The anomaly detection module identifies claim clusters whose temporal dynamics deviate significantly from historical or expected behavior. In the context of social media monitoring, such anomalies often correspond to emerging narratives, coordinated campaigns, or sudden shifts in public discourse, and are frequently used as early indicators of misinformation or breaking events [Vosoughi *et al.*, 2018]. Let c_j denote a claim cluster produced by the posts clustering module. For each cluster, we construct a temporal activity signal from the volume of associated posts over fixed time intervals. Anomaly detection is formulated as identifying time periods in which the observed activity of c_j exhibits statistically significant bursts or changes relative to its baseline. Following prior work on burst and trend detection in text streams, we model anomalies using lightweight temporal statistics rather than content-based reanalysis [Steuber *et al.*, 2023]. This design choice ensures low computational overhead and allows the method to operate online. Common anomaly indicators include sudden increases in posting frequency, acceleration in growth rate, or deviations from expected temporal patterns learned from historical data.

Anomalies are detected independently for each claim cluster, enabling fine-grained monitoring at the level of individual assertions rather than coarse topics. This claim-centric view has been shown to be more effective for identifying emerging misinformation narratives than document- or topic-level analysis [Shao *et al.*, 2018]. Detected anomalies are propagated to downstream components for further analysis, including virality prediction and fact-checking prioritization. By operating on normalized claim clusters, the anomaly detection module focuses on unusual dynamics in semantically coherent assertions rather than transient noise in raw social media streams.

Virality Prediction: The virality prediction module estimates the future diffusion potential of claim clusters based on their early temporal dynamics and structural properties. Given a claim cluster c_j, the objective is to predict whether its associated content will experience sustained growth or wide dissemination over a future time horizon. Virality prediction is a well-studied problem in social media analysis and is commonly framed as forecasting cascade growth or popularity trajectories from early signals [Cheng *et al.*, 2014]. For each claim cluster, we extract a set of lightweight features derived from its posting activity, including early volume, growth rate, acceleration, and anomaly scores produced by the preceding module. These features capture both the scale and momentum of diffusion, which have been shown to be strong predictors of eventual virality [Esteban-Bravo *et al.*, 2024]. The model operates at the level of claims rather than individual posts, enabling it to reason about the propagation of semantically coherent assertions rather than isolated messages. Virality prediction is treated as both a binary classification problem (viral vs. non-viral), and as a regression task estimating future engagement volume to allow for a virality prediction percentage to be fed to the downstream chekworthiness classifier. The prediction model is designed to be computationally efficient and compatible with streaming inputs, enabling continuous updating as new evidence arrives. By integrating anomaly signals and early diffusion patterns, the virality prediction module prioritizes claim clusters that are both unusual and rapidly spreading. These predictions can be used to rank claims for downstream intervention, monitoring, or fact-

checking, complementing the anomaly detection component with forward-looking impact estimates.

TBD: Include information about each module output and the corresponding explainability artifacts

3.2 User Study

4 Results

5 Discussion, Limitations and Future Work

6 Conclusion

Ethical Statement

The user study participants were provided with full disclosure of the study's purpose, how the data would be used in academic research, and that the data could be published. They could choose not to start the survey if they disagreed with the disclosed information.

Acknowledgments

.

References

[Altoe *et al.*, 2024] Filipe Altoe, Catarina Moreira, H. Sofia Pinto, and Joaquim A. Jorge. Online fake news opinion spread and belief change: A systematic review. *HUMAN BEHAVIOR AND EMERGING TECHNOLOGIES*, 2024:1069670, APR 30 2024.

[Altoe *et al.*, 2025] Filipe Altoe, Sérgio Miguel Gonçalves Pinto, and H Sofia Pinto. Explainable automatic fact-checking for journalists augmentation in the wild. In James Kwok, editor, *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 10262–10270. International Joint Conferences on Artificial Intelligence Organization, 8 2025. Human-Centred AI.

[Athira *et al.*, 2023] AB Athira, SD Madhu Kumar, and Anu Mary Chacko. A systematic survey on explainable ai applied to fake news detection. *Engineering Applications of Artificial Intelligence*, 122:106087, 2023.

[Augenstein, 2021] Isabelle Augenstein. Towards explainable fact checking. *arXiv preprint arXiv:2108.10274*, 2021.

[Cazzamatta, 2025] Regina Cazzamatta. The truth game: Verification factors behind fact-checkers' selection decisions. *Journalism*, page 14648849251371952, 2025.

[Cheng *et al.*, 2014] Justin Cheng, Lada Adamic, P Alex Dow, Jon Michael Kleinberg, and Jure Leskovec. Can cascades be predicted? In *Proceedings of the 23rd international conference on World wide web*, pages 925–936, 2014.

[Dierickx and Lindén, 2023] Laurence Dierickx and Carl-Gustav Lindén. Journalism and fact-checking technologies: Understanding user needs. *communication+ 1*, 10(1), 2023.

[Esteban-Bravo *et al.*, 2022] Mercedes Esteban-Bravo, Lisbeth de las Mercedes Jimenez-Rubido, and Jose M Vidal-Sanz. Predicting the virality of fake news in the initial stage of dissemination. *Available at SSRN 4065314*, 2022.

[Esteban-Bravo *et al.*, 2024] Mercedes Esteban-Bravo, Jose M Vidal-Sanz, et al. Predicting the virality of fake news at the early stage of dissemination. *Expert Systems with Applications*, 248:123390, 2024.

[Guo *et al.*, 2022] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206, 2022.

[Hassan *et al.*, 2017] Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, et al. Claimbuster: The first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment*, 10(12):1945–1948, 2017.

[Jin *et al.*, 2024] Ruidong Jin, Xin Liu, and Tsuyoshi Murata. Predicting popularity trend in social media networks with multi-layer temporal graph neural networks. *Complex & Intelligent Systems*, 10(4):4713–4729, 2024.

[Karimiziarani, 2022] Mohammadsepehr Karimiziarani. A tutorial on event detection using social media data analysis: Applications, challenges, and open problems. *arXiv preprint arXiv:2207.03997*, 2022.

[Kotonya and Toni, 2020] Neema Kotonya and Francesca Toni. Explainable automated fact-checking: A survey. *arXiv preprint arXiv:2011.03870*, 2020.

[Majer and Šnajder, 2024] Laura Majer and Jan Šnajder. Claim check-worthiness detection: How well do llms grasp annotation guidelines? *arXiv preprint arXiv:2404.12174*, 2024.

[Markowitz *et al.*, 2023] David M Markowitz, Timothy R Levine, Kim B Serota, and Alivia D Moore. Cross-checking journalistic fact-checkers: The role of sampling and scaling in interpreting false and misleading statements. *Plos one*, 18(7):e0289004, 2023.

[Ni *et al.*, 2024] Jingwei Ni, Minjing Shi, Dominik Stammbach, Mrinmaya Sachan, Elliott Ash, and Markus Leippold. Afacta: Assisting the annotation of factual claim detection with reliable llm annotators. *arXiv preprint arXiv:2402.11073*, 2024.

[Rabb *et al.*, 2022] Nicholas Rabb, Lenore Cowen, Jan P de Ruiter, and Matthias Scheutz. Cognitive cascades: How to model (and potentially counter) the spread of fake news. *Plos one*, 17(1):e0261811, 2022.

[Ruiz-Dolz and Lawrence, 2025] Ramon Ruiz-Dolz and John Lawrence. An explainable framework for misinformation identification via critical question answering. *arXiv preprint arXiv:2503.14626*, 2025.

[Sawiński *et al.*, 2024] Marcin Sawiński, Milena Stróżyna, Włodzimierz Lewoniewski, Piotr Stolarski, Krzysztof Węcel, Ewelina Księżniak, and Witold Abramowicz. Supporting fact-checking process with it tools. *Procedia Computer Science*, 246:2052–2061, 2024.

[Sehat *et al.*, 2024] Connie Moon Sehat, Ryan Li, Peipei Nie, Tarunima Prabhakar, and Amy X Zhang. Misinformation as a harm: structured approaches for fact-checking prioritization. *Proceedings of the ACM on human-computer interaction*, 8(CSCW1):1–36, 2024.

[Shao *et al.*, 2018] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. The spread of low-credibility content by social bots. *Nature communications*, 9(1):4787, 2018.

[Soprano *et al.*, 2024] Michael Soprano, Kevin Roitero, David La Barbera, Davide Ceolin, Damiano Spina, Gianluca Demartini, and Stefano Mizzaro. Cognitive biases in fact-checking and their countermeasures: A review. *Information Processing & Management*, 61(3):103672, 2024.

[Steuber *et al.*, 2023] Florian Steuber, Sinclair Schneider, João AG Schneider, and Gabi Dreo Rodosek. Real-time

anomaly detection and popularity prediction for emerging events on twitter. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, pages 300–304, 2023.

[Sundriyal *et al.*, 2023] Megha Sundriyal, Tanmoy Chakraborty, and Preslav Nakov. From chaos to clarity: Claim normalization to empower fact-checking. *arXiv preprint arXiv:2310.14338*, 2023.

[Sundriyal *et al.*, 2025] Megha Sundriyal, Tanmoy Chakraborty, and Preslav Nakov. Overview of the clef-2025 checkthat! lab task 2 on claim normalization. *Working Notes of CLEF*, 2025.

[Suomalainen *et al.*, 2025] Kari Suomalainen, Nooa Nykänen, Hannele Seeck, Youna Kim, and Ella McPherson. Fact-checking in journalism: An epistemological framework. *Journalism Studies*, pages 1–21, 2025.

[Tan *et al.*, 2025] Xin Tan, Bowei Zou, and Ai Ti Aw. Improving explainable fact-checking with claim-evidence correlations. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1600–1612, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics.

[Thorne *et al.*, 2018] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*, 2018.

[Vosoughi *et al.*, 2018] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *science*, 359(6380):1146–1151, 2018.

[Wack *et al.*, 2024] Morgan Wack, Kayla Duskin, and Damian Hodel. Political fact-checking efforts are constrained by deficiencies in coverage, speed, and reach. *arXiv preprint arXiv:2412.13280*, 2024.

[Zhao *et al.*, 2024] Yilun Zhao, Yitao Long, Tintin Jiang, Chengye Wang, Weiyuan Chen, Hongjun Liu, Xiangru Tang, Yiming Zhang, Chen Zhao, and Arman Cohan. Findver: Explainable claim verification over long and hybrid-content financial documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14739–14752, 2024.