SHORTCHECK: Checkworthiness Detection of Multilingual Short-Form Videos

Henrik Vatndal
University of Stavanger
henrik@factiverse.ai

Vinay Setty
Factiverse AI and University of Stavanger
vsetty@acm.org

Abstract

Short-form video platforms like TikTok present unique challenges for misinformation detection due to their multimodal, dynamic, and noisy content. We present SHORTCHECK, a modular, inference-only pipeline with a user-friendly interface that automatically identifies checkworthy short-form videos to help human fact-checkers. The system integrates speech transcription, OCR, object and deepfake detection, video-to-text summarization, and claim verification. SHORTCHECK is validated by evaluating it on two manually annotated datasets with TikTok videos in a multilingual setting. The pipeline achieves promising results with F1-weighted score over 70%. The demo can be accessed live at http://shortcheck.factiverse.ai.

1 Introduction

The popularity of short-form video platforms such as *TikTok*, *YouTube Shorts* and *Instagram Reels* has transformed how information is produced, consumed, and spread. With billions of monthly active users, these platforms create fertile ground for the spread of misinformation on sensitive topics including politics, health, and social issues. Unlike traditional text or image-based content, these videos may include multiple modalities such as speech, text overlays, music, and visuals, often edited in ways that obscure meaning or context, though not all of these elements are always present; for example, some videos contain only on-screen text without audio, while others show just a speaker without any additional graphics or overlays. For example, Figure 1 shows a TikTok screenshot where the overlay text makes the claim, while the audio transcript (translated to "it's a shame") does not provide any useful information. The video summary notes an urban explosion, indicating potentially contentious content for fact-checkers.

The multimodal complexity of short videos makes automated fact-checking technically challenging and manual efforts are increasingly unsustainable, especially for under-resourced fact-checkers facing unprecedented content scale and funding cuts. In this demo, we present a prototype designed to automate the identification of potentially checkworthy videos, significantly reducing the time required by human fact-checkers. Our prototype is easy to use and can predict checkworthiness in over 30 major languages.[1]

Field	Value
overlay_text	Someone captured the | missile in the Beirut blast
transcript	عيبه اي اي منلو ارهى مريك عيبه
video_summary	The video captures footage of the 2020 Beirut blast, showing destruction and chaos in an urban area, with explosions visible throughout.
buzzword_detected	False
transcript_verdict	hostile
summary_verdict	contentious-issue
overlay_verdict	hostile
Checkworthiness	True

Figure 1: TikTok video with overlay text claiming "Someone captured the missile in the Beirut blast," and noisy Arabic audio and explosion visuals. Features below show why SHORTCHECK marked it *Checkworthy*.

Most existing misinformation detection systems are designed for structured, single-modality content such as news articles, social media posts, or deepfake detection, with a primary focus on either text, audio, transcriptions or visual modalities. We summarize the existing fact-checking systems and their modalities in Table 1. However, these approaches

[1]Intersection of languages supported by Meta Llama3 https://ai.meta.com/blog/meta-llama-3/ and OpenAI Whisper https://github.com/openai/whisper

Proceedings of The 14th International Joint Conference on Natural Language Processing and The 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: System Demonstrations, pages 77–85
December 20-24, 2025

Table 1: Fact-checking systems categorized by input modality, granularity, multilinguality, and fact-checking support.

System	Text	Audio	Image	Video	Granularity	M.lingual	Full FC
BRENDA Botnevik et al. (2020)	✓	✗	✗	✗	Long Text	✗	✓
FLEEK (Bayat et al., 2023)	✓	✗	✗	✗	Single Claim	✗	✓
QACheck (Pan and et al., 2023)	✓	✗	✗	✗	Single Claim	✗	✓
ClaimLens (Devasier et al., 2024)	✓	✗	✗	✗	Single Claim	✗	✗
TruthReader (Li et al., 2024)	✓	✗	✗	✗	Long Text	✗	✓
OpenFactCheck (Iqbal et al., 2024)	✓	✗	✗	✗	Long Text	✗	✓
FactCheckEditor (Setty, 2024a)	✓	✗	✗	✗	Long Text	✓	✓
Loki (Li et al., 2025)	✓	✗	✗	✗	Single Claim	✓	✓
AudioCWD (Ivanov et al., 2024)	✗	✓	✗	✗	Single Claim	✗	✗
LiveFC (Venktesh and Setty, 2025)	✓	✓	✗	✗	Long Text	✗	✓
PodFC (Setty, 2025)	✓	✗	✓	✗	Long Text	✓	✓
Fauxtography (Zlatkova et al., 2019)	✓	✗	✓	✗	Single Claim	✗	✓
AVerImaTeC (Cao et al., 2025)	✓	✗	✓	✗	Single Claim	✗	✓
CER (Barone et al., 2025)	✓	✓	✓	✓	Single Claim	✗	✓
COVID-VTS (Liu et al., 2023a)	✓	✓	✓	✓	Single Claim	✗	✗
ShortCheck (ours)	✓	✓	✓	✓	Long Text	✓	✗

are not suited for short-form video content found on platforms like TikTok or YouTube Shorts due to the casual unstructured nature of the content.

Short-form videos pose unique challenges due to their *limited multimodal generalization*. They often blend speech, text, music, and visuals in non-linear, asynchronous ways that traditional unimodal models struggle to interpret (Alam et al., 2022; Yao et al., 2023; Guo et al., 2022; Singhal et al., 2019). These videos also exhibit *noisy or incomplete modality signals*: some lack audio, others feature distorted overlays or rapid cuts, making existing detectors brittle in real-world conditions (Jindal et al., 2020; Venktesh et al., 2024). Furthermore, most models offer *low interpretability*, returning binary predictions without justifications. This hinders adoption in professional workflows that require transparent, evidence-backed reasoning (Schlichtkrull et al., 2023; Guo et al., 2022; Alam et al., 2022; Venktesh and Setty, 2025).

While multimodal fact-checking is gaining traction, the gap between research prototypes and deployable, interpretable tools for short-form video remains substantial. Bridging this divide requires not only improved multimodal understanding but also system outputs that align with the needs of human fact-checkers in high-throughput environments.

Our Contributions. This paper introduces a demonstration system for detecting checkworthy TikTok videos. Our key contributions are:

- A **modular multimodal pipeline** that integrates OCR, transcription, video-to-text captioning, semantic classification, retrieval and fact-checking modules.
- Two **new multilingual annotated datasets** of TikTok videos labeled for checkworthiness.
- An **evaluation and error analysis** showing which modalities contribute most to reliable classification.
- A **demo interface** that allows fact-checkers to upload videos, inspect intermediate results, and link claims to existing fact-checks.

Together, these contributions provide a step toward bridging the gap between state-of-the-art research and practical tools for combating misinformation on emerging short-form video platforms.

2 Related work

Automated fact-checking has advanced through benchmark datasets such as FEVER (Thorne et al., 2018) and subsequent surveys (Thorne and Vlachos, 2018; Guo et al., 2022), along with real-world datasets like MultiFC (Augenstein et al., 2019), AVeriTeC (Schlichtkrull et al., 2023) and QuanTemp (Venktesh et al., 2024). Within checkworthiness detection, early work introduced context-aware ranking (Gencheva et al., 2017) and systems such as ClaimRank (Jaradat et al., 2018) that support real-time prioritization of claims for journalists.

Multimodal research now combines text, audio and visual cues (Alam et al., 2022), enriched by datasets and models such as FakingRecipe (Bu et al., 2024), SpotFake (Singhal et al., 2019), NewsBag (Jindal et al., 2020) and end-to-end video fact-checking systems with explanation generation (Yao et al., 2023). Deepfake detection architectures like MesoNet (Afchar et al., 2018a) further support authenticity analysis within video pipelines.

Practical systems such as BRENDA (Botnevik

et al., 2020), FACTCHECKEDITOR (Setty, 2024b), PODFC (Setty, 2025) and LIVEFC (Venktesh and Setty, 2025) have begun bridging research and real-world fact-checking needs, though they typically focus on single-modality inputs or long-form content. SHORTCHECK differs by targeting short-form video platforms and integrating text, audio, image and video signals within a modular and interpretable pipeline designed to support professional fact-checkers.

3 System Overview

Given that the checkworthiness of a TikTok video can be inherently subjective, we base our approach on established best practices followed by professional fact-checkers. In particular, we consulted the guidelines of Faktisk.no[2] This also aligns with the definition of fact-checkworthiness in the literature(Jaradat et al., 2018; Barrón-Cedeño et al., 2024)

Short videos up to ten minutes may contain multimodal claims, and are checkworthy only when they pose potential public harm in areas like politics, health or society. Content such as celebrity gossip, sports or advertisements is excluded.

We propose a modular, inference-only pipeline for detecting potentially misleading or checkworthy content in short-form videos, particularly those published on platforms like TikTok. The system assigns each video one of two categorical labels: `Checkworthy`, or `Not_Checkworthy`. Unlike prior work that builds monolithic or end-to-end models, our design emphasizes modularity, interpretability, and adaptability. Each component in the pipeline can be independently replaced, which makes the system robust to failures in specific modalities and easier to maintain in production settings.

3.1 Pipeline Components

The pipeline comprises feature extraction modules tailored to speech, text, visuals, and metadata, whose outputs are aggregated by a rule-based engine for final video classification. Additional modules, such as object detection for weapons, were tested but excluded due to limited contribution.

Optical Character Recognition (OCR): The first stage involves extracting visible on-screen text through Optical Character Recognition (OCR), using the EasyOCR library[3]. This module captures embedded captions or textual overlays, which are common in TikTok videos. However, OCR performance is often challenged by stylized fonts, rapid transitions, and visually noisy frames.

DeepFake Detection: To detect synthetic media, we incorporated a deepfake detection module into the pipeline. Since many methods rely on identity-specific data or high-quality frontal imagery, we evaluated their generalization to TikTok's unconstrained, user-generated content (Abbas and Taeihagh, 2024). We tested three zero-shot models: *MesoNet* (Afchar et al., 2018b), a mesoscopic CNN; *EfficientNet* (Bonettini et al., 2021; Dolhansky et al., 2020), an attention-based model from the DFDC challenge; and *Wvolf/ViT*[4] (Bonettini et al., 2020), selected for its plug-and-play accessibility. Their outputs contributed as signals in the rule-based decision logic.

Speech Transcription: Speech transcription is handled by OpenAI Whisper (Radford et al., 2022), which offers robust multilingual transcription and performs well in noisy audio environments. However, overlapping music or sound effects, which are prevalent in entertainment-oriented videos, can still degrade transcription quality.

Video Summarization: To obtain a visual semantic summary, video frames are sampled and passed through LLaVA (Liu et al., 2023b), a vision-language model that generates frame-level captions describing people, objects, and scenes. These captions are subsequently summarized and contextualized using Meta's LLaMA 3 model, which also performs high-level semantic classification. The model predicts whether a video is political, hostile, benign, or promotional in nature. All models are hosted via Ollama, a lightweight, local model serving platform that supports REST-based inference.[5]

Ideological Language Detection: In addition to these core modules, we incorporate a rule-based system for detecting ideological buzzwords and coded language, often referred to as dog whistles. This module operates on both OCR and transcript outputs, scanning for terms known to encode political or ideological meaning in subtle ways. The

[2] A Norwegian non-profit organization accredited by the International Fact-Checking Network (IFCN).

[3] https://github.com/JaidedAI/EasyOCR

[4] https://huggingface.co/Wvolf/ViT_Deepfake_Detection

[5] ollama.com

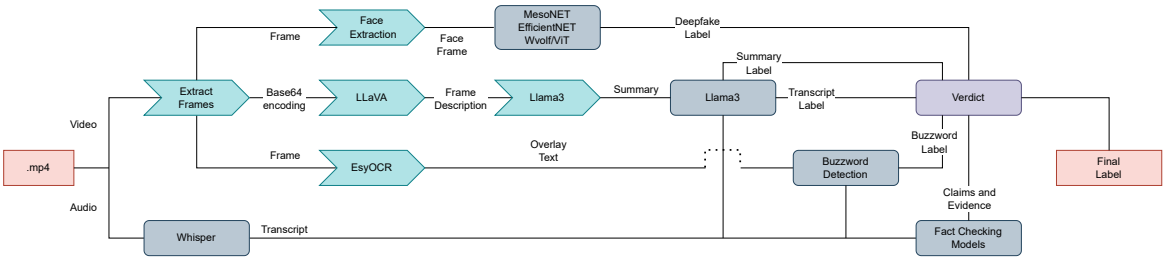

Figure 2: Modular pipeline for fact-checking TikTok videos.

detection rules are informed by prior literature on dog-whistle communication (Albertson, 2015) and curated datasets from organizations such as Faktisk.no.

Text-based fact-checking To further refine the decision-making process, we incorporate a claim detection and external fact-checking module. This component leverages fine-tuned transformer models for claim detection and natural language inference (NLI), applied to both the transcript and the visual summary of the video. Following the approach proposed by (Setty, 2024b), the module identifies declarative, factual statements and attempts to verify them against a fact-checking evidence database. While this does not fact-check the entire content of the video, it provides fact-checkers early signals indicating whether the video may contain verifiably false or misleading information.

Finally, the system aggregates the outputs from all modules using a rule-based logic engine. A scoring system considers multiple module outputs, including the presence of claims, ideological language, or political classification. If the cumulative score exceeds a threshold, the video is marked as `Checkworthy`. Advertisements are detected and filtered early in the pipeline and override all other scores. Videos that do not meet either criterion are labeled as `Not_Checkworthy`. This decision process is entirely transparent and configurable, enabling future refinement without retraining models.

3.2 Verdict Evaluation

Our system produces a final binary label using a lightweight rule-based scoring function that aggregates signals from multiple modules. Each module contributes a fixed weight to a cumulative *Checkworthiness score*.

The scoring scheme assigns concrete weights to each signal: detecting ideological buzzwords adds +2, and transcripts labelled as a contentious issue add +2 while other checkworthy transcript categories add +0.5. Summary and overlay-text classifiers contribute +1.5 for contentious-issue labels or +0.7 for other checkworthy labels. Evidence-supported claims add +0.5 when at least one such claim is present, with an additional +0.25 for each extra supported claim. The total score of over 2 is considered checkworthy otherwise not.

The weights and threshold were manually selected based on findings from the initial thesis work. At the current stage, no automated parameter optimization has been implemented.

3.3 Model Configuration and Deployment

All models used in the pipeline are inference-only and require no task-specific fine-tuning. We use prompt engineering to adapt general-purpose models to specific sub-tasks. The LLaMA 3 and LLaVA models are deployed using Ollama, which allows for lightweight, local hosting and fast prototyping. Custom prompts, temperature settings, stop sequences, and token limits are adjusted to ensure consistent outputs across modules.

3.4 Interpretability and Modularity

A key design goal of our approach is interpretability. Each module exposes intermediate outputs that are human-readable and can be inspected by fact-checkers. This transparency builds trust and enables feedback-driven improvement of the system. The modular design also ensures that any individual component, such as the OCR engine or the semantic classifier, can be replaced or updated without disrupting the entire pipeline. This is especially important for deployment in evolving information ecosystems where content formats and threat types change rapidly.

3.5 Overhead and Concurrency

The video-to-text module is consistently the dominant computational cost, requiring approximately 20 s or more per video regardless of length. Largely

due to the amount of frames being fixed. The `fact_check` module also exhibited substantial variability, between 25.9 s and 1.2 s, depending on the amount of claims present. Audio transcription scaled more directly with video duration, typically accounting for 10–30% of the video length. Lighter modules such as OCR and deepfake detection contributed comparatively little overhead (approximately 2–4 s combined), while buzzword detection, advertisement detection, and final verdict evaluation added negligible cost.

It is important to note that the current prototype processes videos sequentially and does not employ modular concurrency.

4 Experimental Evaluation

4.1 Setup

All experiments were conducted on a local machine with an **NVIDIA A10 GPU (24GB VRAM)**. All models, including vision-language models and LLMs, were served locally via `Ollama` using REST-based endpoints. For full implementation see[6]

4.2 Datasets

We evaluate ShortCheck on two manually annotated dataset in their entirety. Summary of dataset statistics is shown in Table 3

Norwegian influencer data: This dataset includes 249 TikTok videos curated by Faktisk for an emotional analysis study on political trolling via buzzwords like "Stem FRP".[7] We manually annotated each video, with annotator(2) agreement of 96%, for checkworthiness using the guidelines in Section 3.

TikTok Videos from Fact-Checking Websites: This dataset, curated by (Bu et al., 2024), was compiled from fact-checking platforms including Snopes, PolitiFact, FactCheck.org, and Health Feedback. While the majority of content is in English, some posts and modalities appear in other languages. We annotated a sample of 254 videos from this collection, following the same guidelines described before.

4.3 Results

In this section, we present the overall results and ablation studies. In addition, we also present the effectiveness of some of the modules such as DeepFake detection. We plan to evaluate other modules in detail in future work.

Overall Results: The model performs well on both Norwegian and English TikTok videos, with notable differences in class-wise behavior. For Norwegian content, the system achieves strong recall for `Checkworthy` instances (0.91) and high overall accuracy (0.81), indicating robust performance in identifying relevant claims. In contrast, the English dataset shows higher precision for `Checkworthy` (0.82) but lower recall (0.58), suggesting the model is more conservative in flagging English videos as checkworthy. Combined macro-averaged scores are comparable across languages (F1: 0.73 for Norwegian, 0.74 for English), highlighting the pipeline's cross-lingual generalizability with slightly better balance in the Norwegian case.

4.4 Ablation Studies

The ablation study shown in Table 5 demonstrates that textual modules are the most influential components in determining checkworthiness. The removal of the *Transcript Verdict* and *Buzzword* modules resulted in the largest decreases in recall and F1-score, highlighting the critical role of spoken content and ideological language. In contrast, excluding modules such as *Weapon Detection*, *Fact Check*, or *Video-to-Text Verdict* had minimal impact, indicating their limited standalone contribution. Notably, the *Weapon Detection* module slightly reduced overall performance, likely because such content appears infrequently; as a result, it was omitted from the final pipeline. These findings reinforce the system's reliance on semantic and linguistic features rather than visual or metadata-based cues. Finally since removing individual modules does not show a huge drop in performance, the overall performance is attributed to contribution of modules.

DeepFake detection The models were evaluated in a zero-shot setting, without any fine-tuning on the target dataset. Among them, *EfficientNET* outperformed all others across evaluation metrics, achieving an accuracy of 0.612 and a remarkably high precision of 0.992, indicating highly reliable positive predictions. However, its recall of 0.573 suggests it still misses nearly half of actual deepfakes. *MesoNET* and *Wvolf/ViT* perform very poorly.

[6] https://github.com/factiverse/shortcheck

[7] https://www.faktisk.no/artikkel/faktisk-analyse-av-tiktok-menn-mest-negative/109375

Table 2: Results on TikTok videos in Norwegian and English. CW = Checkworthy, NCW = Not Checkworthy. Combined metrics are macro-averages. Metrics: Precision (P), Recall (R), Accuracy (Acc) and F1-weighted (F1-W)

Dataset	**CW**			**NCW**			**Combined (Macro)**			
	P	R	F1-W	P	R	F1-W	P	R	F1-W	Acc
Norwegian influencer	0.42	0.91	0.58	0.98	0.80	0.88	0.70	0.85	0.73	0.81
Fact-checking websites	0.82	0.58	0.68	0.72	0.90	0.80	0.77	0.74	0.74	0.76

Table 3: Dataset composition by language and checkworthiness class. CW: Checkworthy and NCW: Not Checkworthy.

Dataset	**CW**	**NCW**	**Total**
Norwegian influencer	33	204	237
Fact-checking websites	114	140	254
Total	147	344	491

Table 4: Evaluation scores of DeepFake detection models. Precision (P), Recall (R), Accuracy (A), and F1-weighted (F1-w)

Model	**A**	**P**	**R**	**F1-W**
MesoNET	0.114	0.808	0.019	0.038
Wvolf/ViT	0.100	0.000	0.000	0.000
EfficientNET	**0.612**	**0.992**	**0.573**	**0.727**

Table 5: Ablation study showing the performance change when individual modules are removed. Values represent the change from the full system (Baseline). Red indicates a drop in performance. Metrics: Precision (P), Recall (R), Accuracy (Acc) and F1-weighted (F1-W).

Removed	**P**	**R**	**Acc**	**F1-W**
Weapon detection	+0.005	+0.002	+0.004	+0.004
Video summary	+0.001	-0.008	0.000	-0.002
Transcript	+0.027	-0.076	0.000	-0.024
Buzzword	-0.024	-0.050	-0.021	-0.033
OCR	-0.005	-0.030	-0.003	-0.004
Fact Check	+0.013	-0.040	+0.004	-0.009
All modules	**0.737**	**0.727**	**0.884**	**0.720**

5 Demonstration

Our demo system integrates the pipeline into a web interface (Figure 3). The UI allows users to upload videos, inspect transcripts, OCR text, and visual captions, view the final classification with explanations, access linked fact-checks, and examine errors through confusion matrix visualizations.

6 Conclusion and Future Work

We presented ShortCheck, a modular, multilingual pipeline for detecting checkworthy content in short-form videos. The system integrates multiple modalities—text, audio, video, and image—and achieves strong performance on manually annotated TikTok datasets in both Norwegian and English.

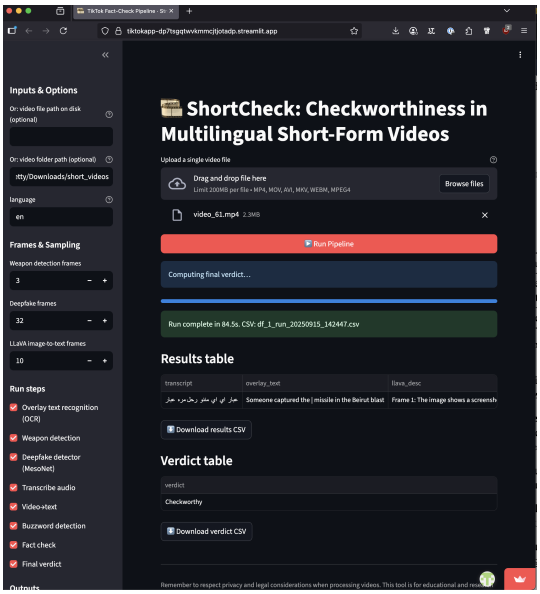

Figure 3: Demo interface for fact-checkers.

Through ablation studies, we showed that transcript features and ideological language signals contribute most strongly to checkworthiness, while visual features such as deepfake detection and object recognition provide limited standalone utility.

The system can be extended to broader fact-checking settings, tougher video and text conditions, learned fusion, and user-centered evaluation to improve robustness and interpretability. Replacing handcrafted weights with data-driven tuning would strengthen generalization across languages and domains. Expanding ideological language resources beyond Norwegian and English would support more accurate detection across diverse linguistic contexts.

7 Ethical Considerations

No personal information is stored, and all uploaded files are processed locally within the session. Files are not accessible to other users or administrators and are deleted after the session ends. The system focuses exclusively on automated analysis of user-submitted content and does not collect behavioral data, usage histories, or identifiers.

ShortCheck operates on short-form videos that may contain sensitive political, social, or ideological material. To reduce potential harm, the system is designed to support human fact-checkers rather than replace them. Its outputs are interpretable, allowing users to inspect intermediate signals and avoid unverified automation. The model does not attempt to generate verdicts beyond checkworthiness and avoids producing judgments that could misrepresent individuals or contexts.

As ShortCheck is multilingual, care is taken to avoid bias toward specific languages or cultural framing. However, disparities in transcription accuracy, OCR performance, or ideological language coverage may still introduce uneven behavior. Future work includes expanding lexical resources and conducting user studies with fact-checkers to identify unintended biases, failure modes, or misclassifications across diverse contexts.

References

Fakhar Abbas and Araz Taeihagh. 2024. Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence. *Expert Systems with Applications*, 252(Part B):124260.

Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. 2018a. Mesonet: a compact facial video forgery detection network. In *IEEE Workshop on Information Forensics and Security*, WIFS '18, pages 1–7.

Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. 2018b. Mesonet: a compact facial video forgery detection network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7. IEEE.

Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2022. A survey on multimodal disinformation detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, COLING '22, pages 6625–6643.

Bethany Albertson. 2015. Dog-whistle politics: Multivocal communication and religious appeals. *Political Behavior*, 37(1):3–26.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. 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 '19, pages 4685–4697.

Mariano Barone, Antonio Romano, Giuseppe Riccio, Marco Postiglione, and Vincenzo Moscato. 2025. Cer: Combating biomedical misinformation through multi-modal claim detection and evidence-based verification. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 4025–4029.

Alberto Barrón-Cedeño, Firoj Alam, Tanmoy Chakraborty, Tamer Elsayed, Preslav Nakov, Piotr Przybyła, Julia Maria Struß, Fatima Haouari, Maram Hasanain, Federico Ruggeri, et al. 2024. The clef-2024 checkthat! lab: Check-worthiness, subjectivity, persuasion, roles, authorities, and adversarial robustness. In *European Conference on Information Retrieval*, pages 449–458. Springer.

Pouya Bayat, Sahisnu Mazumder, Niket Tandon, Yash Kumar Lal, and Xiang Ren. 2023. Fleek: Factual error detection and correction with evidence retrieved from external knowledge. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 147–155.

Nicolò Bonettini, Edoardo Daniele Cannas, Sara Mandelli, Luca Bondi, Paolo Bestagini, and Stefano Tubaro. 2020. Video face manipulation detection through ensemble of cnns.

Nicolò Bonettini, Edoardo Daniele Cannas, Sara Mandelli, Luca Bondi, Paolo Bestagini, and Stefano Tubaro. 2021. Video face manipulation detection through ensemble of cnns. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 5012–5019.

Bjarte Botnevik, Eirik Sakariassen, and Vinay Setty. 2020. Brenda: Browser extension for fake news detection. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, pages 2117–2120.

Yuyan Bu, Qiang Sheng, Juan Cao, Peng Qi, Danding Wang, and Jintao Li. 2024. Fakingrecipe: Detecting fake news on short video platforms from the perspective of creative process. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1351–1360.

Rui Cao, Zifeng Ding, Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2025. Averimatec: A dataset for automatic verification of image-text claims with evidence from the web.

Jacob Devasier, Rishabh Mediratta, Phuong Le, David Huang, and Chengkai Li. 2024. Claimlens: Automated, explainable fact-checking on voting claims using frame-semantics. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 311–319.

Brian Dolhansky, Russ Bitton, Ben Pflaum, Juncheng Lu, Ryan Howes, Menglin Wang, Cristian Canton Ferrer, Michael Barajas, Wissam Essifi, Daniel McDuff, and et al. 2020. The deepfake detection challenge dataset.

Pepa Gencheva, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, and Ivan Koychev. 2017. A context-aware approach for detecting worth-checking claims in political debates. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, RANLP '17, pages 267–276.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10(1):178–206.

Hasan Iqbal, Yuxia Wang, Minghan Wang, Georgi Georgiev, Jiahui Geng, Iryna Gurevych, and Preslav Nakov. 2024. Openfactcheck: A unified framework for factuality evaluation of llms. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 219–229.

Petar Ivanov, Ivan Koychev, Momchil Hardalov, and Preslav Nakov. 2024. Detecting check-worthy claims in political debates, speeches, and interviews using audio data. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12011–12015. IEEE.

Israa Jaradat, Pepa Gencheva, Alberto Barrón-Cedeño, Lluís Màrquez, and Preslav Nakov. 2018. 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*, NAACL '18, pages 26–30.

Sarthak Jindal, Raghav Sood, Richa Singh, Mayank Vatsa, and Tanmoy Chakraborty. 2020. Newsbag: A multimodal benchmark dataset for fake news detection. In *Proceedings of the SafeAI@AAAI 2020 Workshop on Artificial Intelligence Safety*, SafeAI@AAAI '20, pages 1–8.

Dongfang Li, Xinshuo Hu, Zetian Sun, Baotian Hu, Shaolin Ye, Zifei Shan, Qian Chen, and Min Zhang. 2024. Truthreader: Towards trustworthy document assistant chatbot with reliable attribution. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 89–100.

Haonan Li, Xudong Han, Hao Wang, Yuxia Wang, Minghan Wang, Rui Xing, Yilin Geng, Zenan Zhai, Preslav Nakov, and Timothy Baldwin. 2025. Loki: An open source tool for fact verification. In *Proceedings of the 31st International Conference on Computational Linguistics: System Demonstrations*, pages 28–36.

Fuxiao Liu, Yaser Yacoob, and Abhinav Shrivastava. 2023a. Covid-vts: Fact extraction and verification on short video platforms. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 178–188.

Haotian Liu, Pengcheng Lin, Chaoyi Wu, Xiangning Li, Kevin Lin, Ying Zhang, Jingren Du, and Jian Fang. 2023b. Llava: Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Yuchen Pan and et al. 2023. Qacheck: A demonstration system for question-guided multi-hop fact-checking. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 137–146.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Whisper: Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.01234*.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. Averitec: A dataset for real-world claim verification with evidence from the web. In *Advances in Neural Information Processing Systems 36 (Datasets and Benchmarks Track)*, NeurIPS '23, page –.

Vinay Setty. 2024a. Factcheck editor: Multilingual text editor with end-to-end fact-checking. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2744–2748.

Vinay Setty. 2024b. Surprising efficacy of fine-tuned transformers for fact-checking over larger language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2842–2846.

Vinay Setty. 2025. Fact-checking multilingual podcasts. In *Proceedings of the 18th ACM International Conference on Web Search and Data Mining*, WSDM '25, pages 1100–1101.

Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. 2019. Spotfake: A multi-modal framework for fake news detection. In *Proceedings of the IEEE International Conference on Multimedia Big Data*, BigMM '19, pages 39–47.

James Thorne and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, COLING '18, pages 3346–3359.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. 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*, NAACL '18, pages 809–819.

V. Venktesh, Abhijit Anand, Avishek Anand, and Vinay Setty. 2024. Quantemp: A real-world open-domain benchmark for fact-checking numerical claims. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, pages 650–660.

V. Venktesh and Vinay Setty. 2025. Livefc: A system for live fact-checking of audio streams. In *Proceedings of the 18th ACM International Conference on Web Search and Data Mining*, pages 4–7.

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 2023. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, pages 248–258.

Dimitrina Zlatkova, Preslav Nakov, and Ivan Koychev. 2019. Fact-checking meets fauxtography: Verifying claims about images. 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)*, pages 2099–2108.