

Multimodal Artificial Intelligence for Social Media Content Verification, Influence Assessment and Risk Analysis

Prof. S. B. Bele¹, Prajwal G. Dhote², Om S. Shriwas³, Akash B. Zalke⁴

Professor, MCA Department¹

Students, MCA Department²⁻⁴

Vidya Bharati Mahavidyalaya, Amravati, India

Abstract: Social platforms now carry news, opinion, commerce and public safety information in a blend of text, pictures, video and audio. Deception has adapted to this format: a genuine photograph can be given a false caption, a synthetic voice can be attached to an old clip, and coordinated accounts can push a post until it looks popular. Detectors that examine only one channel often miss these cases. This paper presents a unified multimodal framework that treats three linked questions together: is the content authentic (verification), who and what is amplifying it (influence), and how much harm could it cause (risk). The framework encodes each modality separately, aligns them through cross-modal attention, and feeds a shared representation to three task heads whose outputs are merged into an interpretable risk score with a human review step. We survey existing work, describe the architecture, datasets, training approach and evaluation protocol, discuss applications in journalism, public health, brand safety, elections and disaster response, and examine ethical constraints such as privacy, bias and over-blocking. The contribution is a design and evaluation blueprint rather than a new benchmark result, and we close with open problems and a research agenda.

Keywords: Multimodal Learning, Fake News Detection, Influence Analysis, Risk Scoring, Explainable AI, Social Media

I. INTRODUCTION

Online platforms have changed how information spreads. A single post can reach millions of people within hours, and its persuasive force often comes from the combination of a striking image, a short caption and visible engagement counts. Large-scale studies of Twitter cascades found that false stories travelled farther and faster than accurate ones, mainly because people, rather than automated accounts, chose to pass them on [23]. This makes fast, scalable screening valuable, but it also makes the task hard, because what is false is often the relationship between elements rather than any single element.

Consider three ordinary examples. A photograph from an earlier disaster is reposted with a caption claiming it shows a current event; the image is real and the words are plausible, yet the pair is misleading. A short clip carries an AI-generated voice that imitates a public figure; the visuals are ordinary while the audio is fabricated. A modest claim is boosted by a cluster of newly created accounts that like and share it within minutes, giving it a false sense of consensus. A text-only classifier misses the first case, an image-only forensic tool misses the second, and neither notices the third. Multimodal machine learning offers a way forward because it learns from several signal types and from the way they agree or conflict [2]. Recent advances in transformers [21], large pre-trained language models [6], vision transformers [7] and contrastive image-text pre-training [16] have made strong joint representations widely available. Yet most published systems address only verification. Practitioners, however, must decide what to do about a post, and that



decision depends on more than whether it is false: it depends on how far it is likely to spread, who is spreading it, and what damage it could do.

A. Problem statement

Given a social media item that may contain text, one or more images, video or audio, together with metadata and a local propagation network, estimate (i) a veracity label with supporting evidence, (ii) an influence profile of the item and its promoters, and (iii) a calibrated risk level that can be used to choose a moderation or communication response.

B. Objectives and contributions

- To review verification, influence and risk research and show where the three strands have been kept apart.
- To propose an end-to-end multimodal architecture that shares one representation across all three tasks.
- To define a transparent risk-scoring rule that combines veracity, reach, harm severity and source credibility.
- To set out an evaluation protocol covering accuracy, robustness, calibration, fairness and explanation quality.
- To discuss deployment issues, including human oversight, privacy and adversarial adaptation.

II. LITERATURE REVIEW

A. Misinformation and fake news detection

Early surveys organised the field around content, social context and propagation features [18], and later reviews broadened it to cover the knowledge, style, source and user perspectives on false information [26]. Text benchmarks such as LIAR, built from fact-checked political statements, allowed comparison of linguistic models [24]. Work on data collection has stressed that content alone is not enough; FakeNewsNet, for example, pairs news articles with the social engagement that surrounds them [19]. These resources established that context improves detection, but they were mostly used for text or text with simple metadata.

B. Multimodal fake news detection

Researchers soon observed that posts with images behave differently from text-only posts. An attention-based recurrent network combined textual, visual and social features for microblog rumour detection [9]. EANN added an event discriminator so that the model would learn cues that transfer to new events rather than memorising event-specific words [25]. A multimodal variational autoencoder learned a shared latent space for text and image and used it for classification [10]. SpotFake fused BERT text features with image features from a pre-trained convolutional network in a light, practical pipeline [20]. The Fakeddit dataset later supplied more than a million Reddit samples with several label granularities, allowing training at scale [15]. A common limitation across these studies is that fusion is often simple concatenation, which cannot express fine-grained mismatch between a caption and the specific region of an image it refers to.

C. Manipulated and synthetic media

Forensic research examines pixels, compression traces and physiological inconsistencies to detect edited or generated media. FaceForensics++ gave the community a large collection of manipulated face videos produced by several methods, along with detection baselines [17]. Broader overviews describe the arms race between generators and detectors and note that detectors degrade on manipulation types they have not seen [22]. The societal stakes of realistic synthetic media, from harassment to election interference, have been analysed from a legal and policy angle [4]. Forensic detectors are typically evaluated in isolation from the caption and account context that a moderator would actually see.

D. Influence, bots and coordinated behaviour

Measuring influence has long been more subtle than counting followers. An analysis of Twitter found that audience size is a weak predictor of the ability to trigger reshares, and that engaged, topically focused users matter more [3]. Earlier work showed that a small group of influential users drives much of the diffusion, though ordinary users collectively contribute a large share of it [1]. Automated and semi-automated accounts add another layer, since they can inflate apparent popularity and shape early momentum [8]. Graph neural networks, such as the graph convolutional network [12], provide a natural way to learn from follower, reshare and co-activity structure.



E. Harmful content and risk

Hate speech and offensive language research showed how difficult it is to separate abusive content from ordinary strong language [5]. The Hateful Memes challenge demonstrated a multimodal problem in its purest form: individually harmless image and text can become hateful together, and unimodal models perform poorly on such items [11]. Risk in a broader sense, covering health harm, financial fraud, public order and reputation, has mostly been handled by policy teams rather than modelled, which leaves a gap between detection outputs and decisions.

F. Explainability and fairness

Because moderation decisions affect speech, explanations matter. Attribution methods such as SHAP [13] provide feature-level reasons that a reviewer can examine. Fairness surveys document how bias can enter through data, labels and deployment, and how it can produce unequal error rates across languages, dialects and communities [14]. These concerns apply with special force to systems that score both content and the people behind it.

TABLE I
Summary of representative prior work and the gap addressed here.

Work	Modalities	Main idea	Main limitation
Jin et al. [9]	Text, image, social context	Attention-based recurrent fusion for rumour detection	Shallow fusion; verification only
EANN [25]	Text, image	Event-adversarial learning for transfer to new events	No propagation or risk estimate
MVAE [10]	Text, image	Shared latent space via variational autoencoder	Limited cross-modal alignment
SpotFake [20]	Text, image	Pre-trained BERT plus image features	Ignores account and network signals
Hateful Memes baselines [11]	Text, image	Benchmark for joint reasoning on memes	Targets harm class only, not veracity
FaceForensics++ [17]	Video	Manipulated face video detection	No caption or context modelling
Cha et al. [3]	Graph, activity	Influence measured beyond follower counts	No content understanding
This framework	Text, image, video, audio, metadata, graph	One shared representation for verification, influence and risk	Proposed; requires empirical validation

III. METHODOLOGY

A. Overview

The proposed framework has five layers, shown in Fig. 1: input acquisition, modality-specific encoding, cross-modal fusion, task heads, and a decision layer with human review. Each layer is deliberately modular so that individual encoders can be replaced as better models appear.



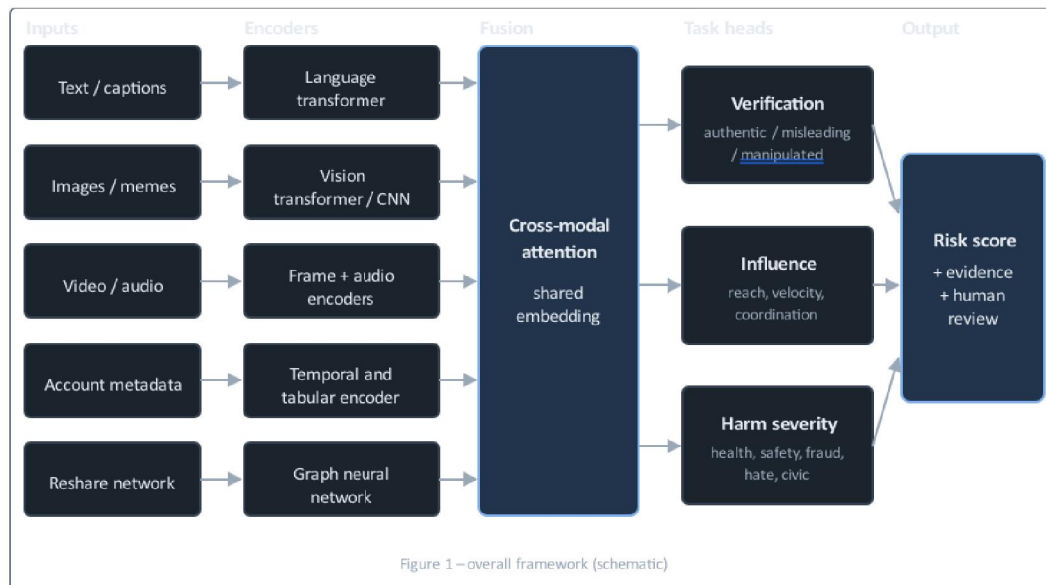


Fig. 1 Proposed multimodal framework. Five input types are encoded separately, fused with cross-modal attention, and used by three task heads whose outputs feed the risk score and the review queue.

B. Data acquisition and preprocessing

Posts are collected through official platform interfaces or licensed research datasets, respecting each platform's terms and applicable data-protection law. Text is normalised for case, URLs, mentions and emoji, and language is identified so that a multilingual encoder can be selected. Images are resized, and both the original and any embedded text (extracted by optical character recognition) are retained. Videos are sampled into keyframes with the audio track separated and transcribed. Account metadata such as account age, posting frequency and profile completeness are converted into numeric features, and the reshare structure is stored as a time-stamped directed graph. Personal identifiers are pseudonymised before storage.

C. Modality-specific encoders

Text is encoded with a pre-trained transformer language model [6, 21]. Images and video frames are encoded with a vision transformer [7], and image-text similarity features can be taken from a contrastive model [16]. Audio is converted to spectral features and encoded by a sequence model, while a separate small network handles temporal and tabular metadata. The reshare graph is processed by a graph convolutional network [12], producing an embedding for the post and for each promoting account. Every encoder outputs a vector of the same size so that they can be fused directly.

D. Cross-modal fusion

Fusion strategy strongly affects how well inconsistency between modalities can be detected. Fig. 2 contrasts the three common options. Early fusion concatenates raw or low-level features, late fusion combines per-modality decisions, and attention-based fusion lets one modality query another. We adopt attention-based fusion, because it allows a caption phrase to attend to the specific image region or audio segment it refers to, and because the attention weights offer a first level of explanation. An auxiliary image-text agreement score, computed from contrastive embeddings, is passed to the fusion block as an additional token to help detect recycled images with new captions.



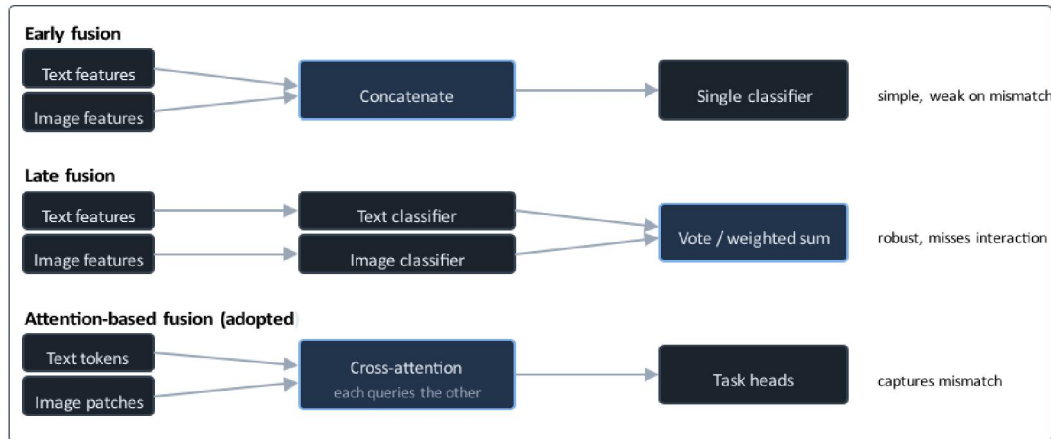


Fig. 2 Fusion strategies compared. Attention-based fusion allows fine-grained interaction between modalities.

E. Task heads and multi-task training

Verification head. A classifier predicts one of three classes: authentic, misleading (true media with false context) and manipulated (edited or synthetic media). A fourth output, "unverifiable", is allowed so that the system can abstain when evidence is insufficient. An evidence retrieval step can attach fact-check articles or reverse image search matches for reviewers.

Influence head. This head estimates expected reach and propagation speed from early engagement, and it profiles the accounts involved. The influence score for a post p is a normalised combination of audience reach, structural importance and engagement velocity:

$$I(p) = \alpha \cdot \hat{R}(p) + \beta \cdot C(p) + \gamma \cdot V(p), \text{ with } \alpha + \beta + \gamma = 1$$

where $\hat{R}$ is log-scaled predicted reach, C is the centrality of the promoting accounts in the reshare graph, and V is the growth rate of reshares in the first observation window. A coordination indicator, based on synchronised posting and unusually similar account behaviour, is reported alongside $I(p)$ rather than folded into it, so that reviewers can see it separately.

Harm-severity head. A multi-label classifier estimates the harm categories that the content could fall into, such as public health, personal safety, financial fraud, hate and harassment, and civic integrity, and assigns a severity between zero and one according to a policy rubric written with domain experts.

The three heads share the fused representation and are trained jointly with a weighted sum of cross-entropy and regression losses. Sharing helps because harmful content is often false, and highly amplified false content is often coordinated, so the tasks reinforce each other. Pre-trained encoders are fine-tuned with a small learning rate, and class imbalance is handled by re-weighting rather than by discarding data.

F. Risk scoring

The decision layer combines the outputs into a single score in the range zero to one:

$$\text{Risk} = w_1(1 - V) + w_2 \cdot I + w_3 \cdot H + w_4(1 - S)$$

Here V is the model's confidence that the content is authentic, I is the influence score, H is harm severity and S is source credibility derived from account history. The weights sum to one and are set with stakeholders for the deployment context; a health agency, for example, would raise w_3 . The score is then mapped to an action band using both reach and harm, as illustrated in Fig. 3, so that a low-reach, low-harm falsehood is simply monitored while a high-reach, high-harm item goes to an expert. Because the formula is additive and each term is visible, reviewers can see why a post was flagged.



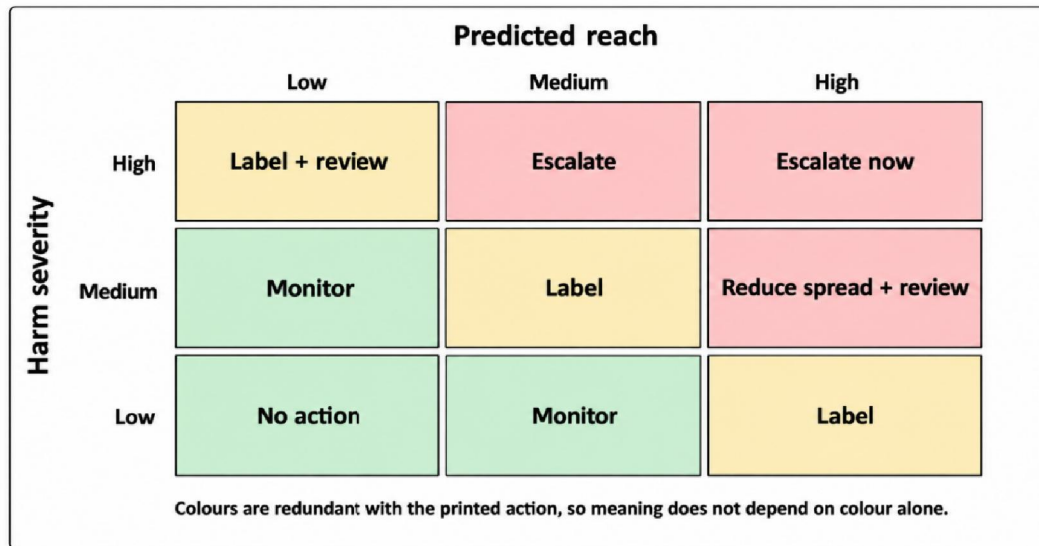


Fig. 3 Illustrative risk matrix that turns reach and harm severity into a suggested action.

TABLE II

Example risk bands and recommended responses.

Band	Score range	Typical situation	Suggested response
Low	0.00 – 0.30	Authentic or harmless, limited reach	No action; sample for audit
Guarded	0.31 – 0.55	Uncertain claim, modest reach	Monitor; attach context if it grows
Elevated	0.56 – 0.75	Likely misleading with rising engagement	Add label; queue for reviewer
Severe	0.76 – 1.00	Manipulated or dangerous, high reach or coordination	Escalate; limit amplification pending review

G. Workflow

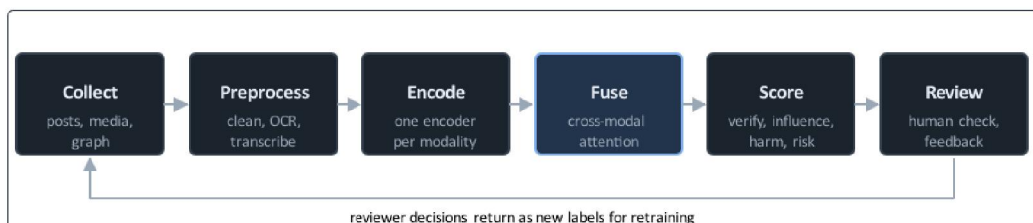


Fig. 4 Operational workflow with a feedback loop from reviewers back into training data.

H. Datasets

No single public resource covers all three tasks, so training combines several sources. Table III lists candidates. Where labels are missing for influence or harm, weak supervision from engagement logs and expert-written rubrics is used, and a smaller, carefully annotated set is reserved for validation.



TABLE III
Public resources that can be combined for training and evaluation.

Resource	Content	Useful for
LIAR [24]	Fact-checked short political statements (text)	Text veracity pre-training
FakeNewsNet [19]	News content plus social engagement context	Verification with propagation signals
Fakeddit [15]	Over one million multimodal Reddit posts, multiple label levels	Large-scale multimodal verification
Hateful Memes [11]	Image-text memes with hate labels	Harm head; joint reasoning
FaceForensics++ [17]	Real and manipulated face videos	Manipulation detection
Platform research data (licensed)	Posts, reshares, account activity	Influence and coordination modelling

I. Evaluation protocol

Accuracy alone is not adequate for a screening tool used on heavily imbalanced data, so we recommend reporting several complementary measures (Table IV). Splits should be made by event and by time rather than at random, because random splits let models memorise event vocabulary and inflate results [25]. Robustness is tested with perturbations such as re-compression, cropping, paraphrase and translation, and with previously unseen manipulation methods [22]. Error rates are broken down by language and community to look for unequal performance [14], and attribution methods [13] are used to check whether explanations point to sensible evidence.

TABLE IV
Proposed evaluation measures.

Aspect	Measures	Importance
Classification Quality	Precision, Recall, F1-Score, ROC-AUC, PR-AUC	Effectively evaluates performance under class imbalance.
Calibration	Expected Calibration Error (ECE), Reliability Plots	Ensures predicted probabilities are trustworthy for risk assessment.
Influence Prediction	Rank Correlation, Predicted Reach Error	Helps prioritize content review and moderation efforts.
Robustness	F1-score degradation under perturbations and unseen generators	Measures resistance to adversarial manipulation and evolving content generators.
Fairness	Error-rate differences across languages and demographic groups	Prevents biased moderation and uneven content blocking.
Explanation Quality	Reviewer Agreement, Decision Time	Supports practical and efficient human oversight.

IV. APPLICATIONS

TABLE V
Application areas and the way each uses the three analyses.

Domain	Verification use	Influence use	Risk use
Journalism and fact-checking	Find recycled images and false captions quickly	Locate the accounts that started a rumour	Rank claims by public impact
Public health	Flag misleading treatment claims with images	Identify accounts amplifying them	Prioritise items that could cause physical harm



Domain	Verification use	Influence use	Risk use
Elections and civic integrity	Detect edited candidate media and cloned voices	Reveal coordinated pushes	Escalate content that suppresses voting
Disaster response	Separate genuine field images from old ones	Find trustworthy local reporters	Prevent panic and misdirected aid
Brand safety	Catch fake endorsements and counterfeit ads	Choose genuine partners, avoid inflated ones	Protect reputation from viral attacks
Finance and fraud	Spot fabricated screenshots and deepfake executives	Detect promotion rings for scams	Warn users before losses occur

In each setting the system acts as a triage aid. It shortens the time between a post appearing and a qualified person looking at it, and it gives that person the evidence and the priority ranking, rather than making final decisions on its own.

A. Ethical and deployment considerations

- **Human oversight.** High-impact actions such as restricting reach or removing content should require human confirmation, with an appeals route.
- **Privacy.** Only the data needed should be collected, identifiers should be pseudonymised, and retention limits should be enforced.
- **Bias.** Training data reflects the languages and communities that were annotated most; regular audits and targeted data collection are needed [14].
- **Freedom of expression.** Satire, opinion and contested claims should not be treated as falsehood; the "unverifiable" output and label-first responses reduce over-blocking.
- **Misuse.** Influence scoring of individuals could be used to profile or target people, so access should be restricted and logged.
- **Adversarial adaptation.** Attackers will modify content to evade detection, which calls for continuous retraining and red-team exercises.

V. CONCLUSION

Deception on social media is multimodal, networked and consequential, so tools that look at one channel, or that answer only the question of truth, leave important gaps. This paper described a framework in which text, image, video, audio, account and network signals are encoded separately, fused through cross-modal attention, and used to produce verification, influence and harm estimates that are combined into a transparent, tunable risk score. Its design puts weight on the properties that matter in practice: fine-grained detection of mismatch between modalities, explanations a reviewer can act on, calibrated outputs for action bands, and human oversight for consequential decisions. The work presented here is a blueprint and evaluation plan; its claims must be tested empirically on event-split, multilingual and adversarially perturbed data before deployment.

VI. FUTURE SCOPE

- Multilingual and low-resource coverage. Extend to regional languages, code-mixed text and dialects, where labelled data is scarce and error rates are usually highest.
- Generative-media detection. Keep pace with new image, video and voice generators through continual learning and provenance signals such as content credentials and watermarks.
- Evidence-grounded reasoning. Combine multimodal models with retrieval over fact-check archives and trusted sources so that verdicts cite verifiable evidence.
- Temporal and early detection. Improve prediction from the first minutes of spread, when intervention is most effective.



- Privacy-preserving learning. Use federated and differentially private training so that models can learn across institutions without sharing raw user data.
- Causal analysis of influence. Move from correlation to estimates of how much an intervention, such as a label, actually reduces spread.
- Standard benchmarks. Build open, ethically collected benchmarks that carry joint labels for veracity, influence and harm.
- Human–AI collaboration studies. Measure how explanations change reviewer accuracy, speed and trust.

REFERENCES

1. Bakshy, E., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). Everyone's an influencer: Quantifying influence on Twitter. *Proceedings of the 4th ACM International Conference on Web Search and Data Mining (WSDM)*, 65–74.
2. Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2), 423–443.
3. Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, K. P. (2010). Measuring user influence in Twitter: The million follower fallacy. *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 10–17.
4. Chesney, R., & Citron, D. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review*, 107(6), 1753–1820.
5. Davidson, T., Warmesley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, 11(1), 512–515.
6. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, 4171–4186.
7. Dosovitskiy, A., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*.
8. Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, 59(7), 96–104.
9. Jin, Z., Cao, J., Guo, H., Zhang, Y., & Luo, J. (2017). Multimodal fusion with recurrent neural networks for rumor detection on microblogs. *Proceedings of the 25th ACM International Conference on Multimedia*, 795–816.
10. Khattar, D., Goud, J. S., Gupta, M., & Varma, V. (2019). MVAE: Multimodal variational autoencoder for fake news detection. *Proceedings of The Web Conference (WWW)*, 2915–2921.
11. Kiela, D., et al. (2020). The Hateful Memes challenge: Detecting hate speech in multimodal memes. *Advances in Neural Information Processing Systems (NeurIPS)*, 33.
12. Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations (ICLR)*.
13. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems (NeurIPS)*, 30.
14. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), Article 115.
15. Nakamura, K., Levy, S., & Wang, W. Y. (2020). r/Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, 6149–6157.
16. Radford, A., et al. (2021). Learning transferable visual models from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning (ICML)*, PMLR 139, 8748–8763.



17. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1–11.
18. Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1), 22–36.
19. Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2020). FakeNewsNet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data*, 8(3), 171–188.
20. Singhal, S., Shah, R. R., Chakraborty, T., Kumaraguru, P., & Satoh, S. (2019). SpotFake: A multi-modal framework for fake news detection. *Proceedings of the IEEE International Conference on Multimedia Big Data (BigMM)*, 39–47.
21. Vaswani, A., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30.
22. Verdoliva, L. (2020). Media forensics and deepfakes: An overview. *IEEE Journal of Selected Topics in Signal Processing*, 14(5), 910–932.
23. Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), 1146–1151.
24. Wang, W. Y. (2017). "Liar, liar pants on fire": A new benchmark dataset for fake news detection. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers)*, 422–426.
25. Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., & Gao, J. (2018). EANN: Event adversarial neural networks for multi-modal fake news detection. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 849–857.
26. Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys*, 53(5), Article 109.

