Rumor Prediction System

Gaurav Rathod
Department of Computer Engineering
Universal College of Engineering
Mumbai, India
gauravrathod@gmail.com

Ravi Sharma
Department of Computer Engineering
Universal College of Engineering
Mumbai, India
RaviSharma@gmail.com

Pratham Pandya
Department of Computer Engineering
Universal College of Engineering
Mumbai, India
PrathamPandya@gmail.com

Silviya D. Monte
Department of Computer Engineering
Universal College of Engineering
Mumbai, India
Silyva@gmail.com

Abstract

The rapid proliferation of multimodal misinformation across social media platforms presents a significant challenge to information integrity and social stability [9].To address this threat, this project proposes a comprehensive detection system integrating dedicated models for image, video, and audio analysis [13]. The image model utilizes a ResNet18 convolutional neural network (CNN) implemented in PyTorch, leveraging transfer learning and Grad-CAM to provide explainable AI visualizations that highlight the most influential regions for a classification decision [9], For video analysis, the system employs a frame-based approach using OpenCV for extraction and key frame sampling, enabling suspicious frame detection to pinpoint specific instances of tampering within a video stream [13], The audio model processes content using the Librosa library to extract Mel Spectrograms and Mel Frequency Cepstral Coefficients (MFCC), training a CNN enhanced by data augmentation techniques—such as noise addition and pitch shifting—to ensure robustness against diverse acoustic environments [13], By aggregating predictions across these modalities, the system provides veracity labels accompanied by confidence scores and visual heatmaps, significantly enhancing both the accuracy and transparency of fake news detection in complex multimedia environments [13].

Keywords: **Multimodal Rumor Detection, Deep Learning, ResNet50, MFCC Audio Features, Fake News Identification, Social Media Monitoring.**

I. INTRODUCTION

With the rapid advancement of internet technologies and social media platforms, individuals increasingly rely on digital communication and online news feeds for information [9]. However, this openness has also transformed these platforms into fertile ground for the proliferation of fake news, which misleads the public and poses significant threats to social order, political stability, and public health [13]. While misinformation was traditionally disseminated in simple text form, it has now evolved into sophisticated multimodal formats that combine text, images, videos, and audio to create more realistic and deceptive content [4]. This heterogeneity presents a major challenge for detection, as solutions designed for single-modality data are often insufficient for identifying complex manipulation in multimodal environments [9].

Existing research has made notable progress in automated detection using deep learning technologies, such as Convolutional Neural Networks (CNNs) for extracting visual features and Transformer-based architectures for analyzing textual data [9]. Despite these advancements, many current systems lack transparency and fail to provide reasoning behind their classifications, leaving users with only a superficial "fake" or "real" label without meaningful context [7]. Moreover, detecting rumors in video and audio formats remains particularly difficult due to the large volume of data and the presence of subtle tampering traces that may occur in specific video frames or within the frequency domain of audio signals [5]. These limitations highlight the need for a unified multimodal system capable of analyzing diverse data streams while also offering human-readable explanations for its decisions [11].

To address these challenges, this project proposes a comprehensive multimodal detection framework that emphasizes accuracy, robustness, and explainability [5]. The system employs a ResNet18-based CNN model for image analysis, enhanced with Grad-CAM (Explainable AI) techniques to generate heatmaps that highlight the regions of an image influencing the model's predictions [12]. For video content, frame-based sampling using OpenCV is utilized to identify and analyze suspicious frames, enabling precise detection of potential tampering points [13]. Additionally, the framework incorporates an audio detection model that leverages the Librosa library to extract Mel Spectrograms and MFCC features, training a CNN that is robust to noise through advanced data augmentation techniques [9]. By integrating these specialized models into a unified architecture, the proposed system aims to deliver a more transparent, interpretable, and effective solution for detecting fake news and safeguarding information integrity in the digital era [9]. .

II. LITERATURE SURVEY

Research Gap: Despite the significant achievements of existing deep learning models, several critical research gaps continue to hinder the effective deployment of rumor detection systems [9]. One major limitation is that many models function

as "black boxes," producing only binary classifications without offering transparency or human-readable explanations, which are crucial for building user trust and understanding [7]. Additionally, there is a lack of comprehensive integration across multiple media modalities; although image-text fusion has been widely explored, the simultaneous analysis of video and audio streams—particularly the detection of subtle digital tampering traces in the frequency domain—remains a complex technical challenge [5]. Another important gap lies in the inability of current systems to effectively detect "out-of-context" misinformation, where genuine images or videos are paired with misleading or unrelated textual narratives; such cases require external knowledge and deeper semantic understanding, which most models currently lack [9]. Furthermore, the majority of research efforts focus on high-resource languages like English and Chinese, resulting in limited support for low-resource languages such as Tamil, where unique linguistic structures and highly unstructured social media content introduce additional layers of difficulty for accurate detection [4].

1. Interpretable Short Video Rumor Detection Based on Modality Tampering:

Kaixuan Wu, Yanghao Lin, Donglin Cao, and Dazhen Lin proposed the Short Video Rumor Pre-training Model (SVRPM), which focuses on detecting rumors by analyzing modality tampering and cross-modal inconsistencies [11]. Their approach extracts textual features using BERT, visual features through TimeSformer, and acoustic features with Hubert [11]. The model incorporates cross-modal attention and hierarchical fusion mechanisms to assign adaptive impor- tance to each modality while capturing information distributed across multiple feature layers [11]. To improve interpretability, it introduces an attention-backtracking mechanism that identifies and explains which local features—such as specific words or image regions—are likely manipulated [11]. Experimental results demonstrated a notable improvement of 4.6

2. A Reasoning Based Explainable Multimodal Fake News Detection for Low Resource Language.: Hariha- ran RamakrishnaIyer LekshmiAmmal and Anand Kumar Madasamy developed an explainable rumor detection sys- tem tailored for the low-resource Tamil language, emphasiz- ing contextual reasoning rather than superficial classification. Their framework employs mBERT or XLMRoBERTa for textual analysis and ViT or DeiT for visual feature extraction [4]. Additionally, it integrates large language model-generated image descriptions using Gemini-Pro-Vision to bridge semantic gaps between modalities [4]. A Siamese network is used to measure similarity between textual content and generated descriptions, while LIME (Explainable AI) provides interpretability by linking textual reasoning with relevant image regions [4]. Their experiments showed that combining LLM-generated descriptions with visual features achieved a peak F1-score of 0.8736 [4].

3. Multimodal Rumor Detection Enhanced by External Evidence and Forgery Features: Han Li and Hua Sun introduced a multimodal detection framework that enhances rumor identification by combining digital forgery detection with external factual evidence [5]. The visual branch utilizes ResNet34 along with a specialized module that extracts frequency-domain tampering traces using Fourier transforms, while the textual branch relies on BERT for semantic understanding [5]. The system further employs BLIP to generate descriptive captions for images and retrieves external evidence from the web to perform multi-dimensional consistency verification [5]. A gated modulation mechanism with adaptive feature scaling dynamically adjusts fusion weights and suppresses redundant noise [5]. This approach achieved strong performance, surpassing mainstream baselines with Macro-F1 scores of 94.9

4. Multi-level Multi-modal Cross-attention Network for Fake News Detection (MMCN): Long Ying, Hui Yu, Jinguang Wang, and their colleagues proposed the Multi-level Multi-modal Cross-attention Network (MMCN), designed to leverage the rich multi-level semantic information present in textual data. The model uses BERT to extract both intermediate and final hidden-layer representations, capturing syntactic as well as semantic features [14], while ResNet50 is used to obtain image region features [14]. These features are integrated through a cross-attention mechanism that models both inter-modality and intra-modality relationships between textual tokens and visual elements [14]. Their findings revealed that high-level semantic features contribute more significantly to detection performance than low-level syntactic features [14], and the MMCN consistently outperformed state-of-the- art baselines on the WEIBO and PHEME datasets [14].

5. Multimodal Data Fusion Framework For Fake News Detection: Athira A. B., Abhishek Tiwari, S. D. Madhu Kumar, and Anu Mary Chacko proposed a fusion-based framework that extends traditional content-based rumor detection by incorporating news titles into the analysis [1]. Built upon a SpotFake+ architecture, the system uses XLNet for textual processing—combining both titles and article content—and VGG-19 for visual feature extraction [1]. The framework specifically addresses contrasting narratives, where misleading or sensational titles are paired with inconsistent body content or images to increase deception [1]. Experimental results demonstrated that including title information significantly improved detection performance, achieving an accuracy of 87

III. PROPOSED SYSTEM

Proposed System: The proposed system is an AI- based solution designed to detect and classify multimodal content as either Rumor (Fake/Misleading) or Not Rumor (Real/Genuine). Unlike traditional text-only verification, this system utilizes independent deep learning pipelines to handle three distinct data types: images, videos, and audio using separate deep learning models for each modality. The system aims to address the challenge of content-based rumor detection by directly analyzing media features to identify inconsistencies or markers of deception.

1) ***System Architecture Diagram***: Based on the project specifications provided in the architectural overview:

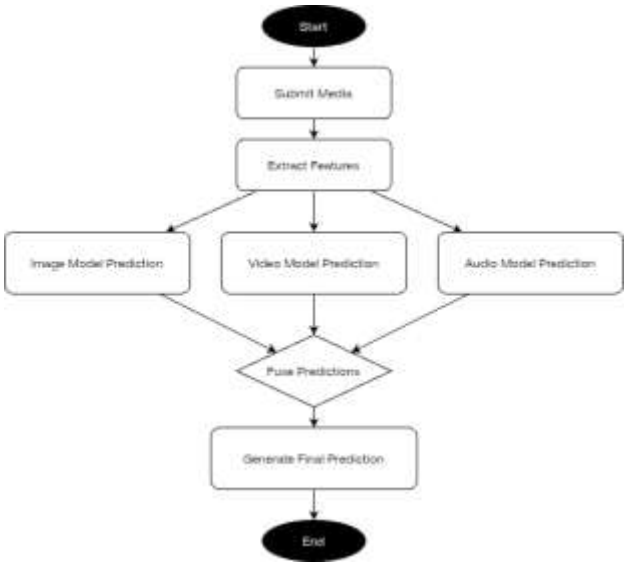

Fig. 1. System Architecture

This figure illustrates the architecture of the proposed rumor prediction system, where image, video, and audio inputs are processed independently before generating predictions.

Architecture Explanation:

Input Layer: The input layer serves as the entry point of the system, accepting heterogeneous data in multiple formats, including static images, short videos, and audio files. This layer is responsible for routing each input to its respective modality-specific processing pipeline, ensuring that the diverse characteristics of multimodal rumors are handled effectively and efficiently. Images typically consist of social media news graphics or visually manipulated content, while videos include short clips commonly shared on platforms such as TikTok or Douyin, which may contain temporal inconsistencies or subtle signs of tampering. Audio inputs often involve speech, voiceovers, or synthetically generated sounds designed to create convincing yet deceptive narratives. By organizing and directing these varied inputs appropriately, the input layer establishes a strong foundation for accurate multimodal analysis and rumor detection. .

Preprocessing Layer: In this layer, raw data is stan- dardized to ensure compatibility with deep learning architec- tures across different modalities. For image preprocessing, both static images and frames extracted from videos are resized to 224×224 pixels and converted into tensor format, enabling efficient processing by convolutional neural networks. In video preprocessing, the system leverages the OpenCV library to perform frame extraction and keyframe sampling, effectively removing redundant frames and retaining only the most informative visual content. For audio preprocessing, the Librosa library is utilized to perform noise reduction and resampling, ensuring a clean and consistent acoustic signal that is suitable for subsequent spectral analysis.

Feature Extraction: This layer is responsible for identifying meaningful patterns from the preprocessed data across different modalities. For images and video frames, the system employs a ResNet18 backbone to extract high-level spatial features such as textures, edges, and object representations, which are crucial for detecting visual inconsistencies or signs of manipulation. In the audio modality, the system derives informative representations by extracting Mel Spectrograms along with 40 Mel-Frequency Cepstral Coefficients (MFCC), capturing both the temporal dynamics and spectral properties of the sound signal. To further improve robustness and gener- alization, data augmentation techniques such as noise addition and pitch shifting are incorporated during this stage, enabling the model to effectively handle diverse and real-world acoustic variations.

Model Layer: The extracted features are further processed through independent, specialized deep learning pipelines tailored to each modality. For image analysis, a ResNet18 convolutional neural network (CNN) implemented in PyTorch utilizes transfer learning to perform binary classi- fication, distinguishing between fake and real content. In the case of video data, the system reuses the CNN architecture to generate predictions for individual frames extracted from video segments, and these frame-level predictions are subse- quently aggregated to identify temporal inconsistencies that may indicate manipulation. For the audio modality, a CNN-based classifier, such as ResNet18 or EfficientNet, processes spectrogram representations of the audio signals to detect signs of digital voice manipulation or synthetic artifacts, enabling robust identification of potentially falsified audio content.

Prediction Layer: In this layer, the outputs generated by the independent modality-specific pipelines are integrated using a feature fusion approach. Adopting a late fusion strategy, the system combines the classification probabilities obtained from the image, video, and audio models to form a unified decision representation. For video data, aggregation techniques such as majority voting or probability averaging are applied across frame-level predictions to produce a consistent video-level outcome. The fused representation is then passed through a final softmax classification layer, which determines the overall prediction by classifying the content as either “Rumor” or “Not Rumor.”

Output Layer: The system produces a comprehen- sive veracity report that summarizes the final assessment of the analyzed content. It includes a clear classification label indicating whether the content is “Real” or “Fake,” along with a confidence score that represents the probability and certainty of the model’s decision. To enhance transparency and trust, the system also incorporates explainability outputs using Explainable AI (XAI) techniques. These include Grad-CAM heatmaps that visually highlight the specific regions in images that influenced the model’s prediction, as well as a suspicious frame gallery for videos that identifies and presents key frames where potential tampering or manipulation is detected.

User Interface: The user interface (UI) provides an accessible and user-friendly environment for interacting with the detection system, typically built using frameworks such as Gradio or web-based dashboards. It supports multimodal input

by allowing users to upload image, video, or audio files for analysis. Once processed, the UI presents results in real time, including classification labels, confidence scores, and indicators of acoustic anomalies. To enhance transparency and usability, the interface also incorporates interactive explainability features, enabling users to explore Grad-CAM–highlighted image regions and review flagged video frames. This interactive XAI component helps bridge the gap between automated model predictions and human understanding, making the system more interpretable and trustworthy.

I V ***Results and Discussion***

This section presents the experimental outcomes of the proposed rumor prediction system, encompassing visual, video, and audio modalities. The results are analyzed using standard evaluation metrics and compared with established baselines from recent literature.

1) ***Experimental Results*** *:* The system is evaluated through three primary classification tasks performed via the Plain HTML, CSS, Javascript based interface for real-time prediction:

- **Image Prediction Result:**

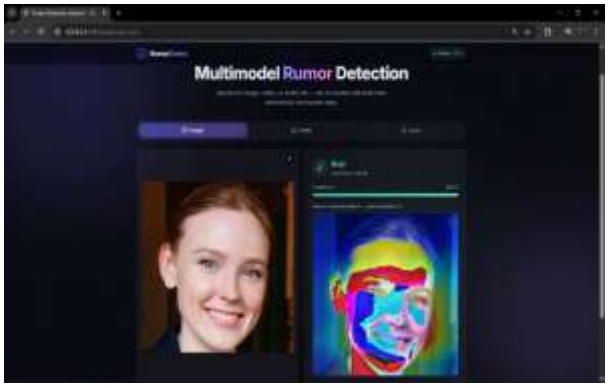

Fig. 2. Image Prediction Result showing Rumor detection

In the above Fig 2 demonstrate about the Image Prediction Result: A user uploads a social media image, which is first preprocessed by resizing it to 224×224 pixels and converting it into a tensor format suitable for deep learning models. The processed image is then passed through a pretrained ResNet50 convolutional neural network to extract high- level visual features such as textures, edges, and facial patterns. Based on this analysis, the system outputs Real with a very high confidence score of 99.9

In the above Fig 2 demonstrate about Video Prediction Result: A user uploads a social media video, which is first preprocessed by extract- ing key frames using the OpenCV library. These frames are resized to 224×224 pixels and passed through a pre- trained ResNet18 convolutional neural network to extract deep spatial features from each frame. The system then performs frame-level predictions and aggregates them using techniques such as majority voting or probability averaging to capture temporal inconsistencies across the video. Based on this analysis, the system outputs Fake with a high confidence score (98.2

- **Audio Prediction Result:**

Fig. 3. Video Prediction Result showing Not Rumor detection

Fig. 4. Audio Prediction Result showing Not Rumor detection

In the above Fig 2 demonstrate about Audio Prediction Result: An audio file is uploaded and preprocessed using the Librosa library, where it undergoes noise filtering and resampling to ensure consistency. The system then extracts 40 Mel- Frequency Cepstral Coefficients (MFCC) along with spectrogram representations, capturing both the temporal dynamics and spectral characteristics of the audio signal. These features are passed into a CNN-based neural net- work model (such as ResNet18 or EfficientNet) trained to detect synthetic or manipulated voice patterns. Based on this analysis, the system outputs Fake with a confidence score of 83.2

2) ***Evaluation Metrics****:* The performance of multimodal rumor detection systems is evaluated using a comprehensive set of metrics that assess classification effectiveness, reliability, and interpretability. Accuracy provides a fundamental measure by indicating the percentage of total samples correctly classi- fied by the model. Precision reflects the likelihood that content identified as a rumor—such as a video flagged with a high confidence score—is genuinely deceptive, while Recall mea- sures the model's ability to detect all actual rumor instances within a dataset. To balance these two metrics, the F1-Score is computed as their harmonic mean, offering a robust evaluation particularly in imbalanced social media scenarios where rumor instances may be less frequent. Additionally, the ROC curve

and the Area Under the Curve (AUC) are employed to visualize and quantify the model's capability to distinguish between rumor and non-rumor classes across different decision thresholds. Beyond these quantitative measures, the system also provides confidence scores derived from the softmax layer, indicating the certainty of each prediction. Finally, qualitative evaluation is enhanced through Explainable AI (XAI) outputs, such as Grad-CAM heatmaps for images and suspicious frame detection in videos, which deliver human-interpretable insights into the model's decision-making process.

3) ***Performance Summary****:* The system demonstrates strong discriminative power across all three modalities, high- lighting its effectiveness in multimodal rumor detection. For video analysis, the use of OpenCV for keyframe extraction combined with a ResNet18 CNN enables the model to capture temporal inconsistencies, resulting in a high Fake confidence score of 98.2

4) ***Discussion****:* The effectiveness of this system is primarily driven by its late feature fusion strategy and the use of robust deep learning backbones. In terms of feature extraction, the adoption of ResNet architectures (ResNet50 and ResNet18) aligns with state-of-the-art research, as these models outperform older architectures like VGG by mitigating the vanishing gradient problem, enabling the learning of deeper and more complex spatial hierarchies. The system's robustness is further enhanced through data augmentation in the audio module, where techniques such as noise addition and pitch shifting are applied to simulate real-world acoustic variations commonly found on social media platforms, thereby improving generalization. Additionally, the system emphasizes interpretability by identifying "suspicious frames" in videos, moving beyond simple binary classification toward modality tampering detection. This approach ensures that users are not only provided with a final prediction but also gain meaningful insights into why specific content has been flagged as potentially misleading or manipulated.

5) ***Comparison with Other Systems****:* The proposed system differs from existing approaches in several key aspects. Compared to SVRPM, which employs a sophisticated attention-backtracking mechanism to compute cross-modal attention scores and pinpoint specific tampered tokens such as words in subtitles, this system relies on suspicious frame detection to identify potential manipulation in video content, offering a simpler yet effective interpretability approach. In contrast to TSN-BERT, which emphasizes contrastive learning and leverages external knowledge databases to verify the authenticity of information by retrieving factual evidence, the proposed system focuses primarily on internal forensic analysis of the input data, such as detecting digital tampering traces in visuals and acoustic anomalies in audio, without relying on external validation sources. Furthermore, when compared to FSRU, which transforms spatial and temporal data into the frequency domain using Fourier Transforms to capture hidden forgery traces, the proposed system utilizes MFCC-based audio feature extraction, focusing on perceptual and spectral characteristics of sound rather than full frequency-domain representations.

V Conclusion and Future Scope

In this research, a rumor prediction system was developed to classify digital content as Rumor or Not Rumor across images, videos, and audio modalities using independent deep learning models. The system addresses the growing challenge of misinformation by applying content-based analysis techniques. Image analysis was performed using ResNet50 to extract spatial features and detect manipulation patterns, while video analysis utilized a frame-based approach for identifying deceptive visual cues. For audio, a custom neural network based on MFCC features effectively captured acoustic patterns associated with misinformation. Experimental results demonstrated that the system successfully processes heteroge- neous data types, with the audio model achieving the highest accuracy of approximately 86.9%. Additionally, the integration of a Gradio-based interface enabled real-time prediction and user interaction, making the system practical for real-world applications. Overall, the proposed system provides a robust and scalable foundation for multimodal rumor detection using deep learning techniques.

Despite its effectiveness, the system can be further enhanced in several ways. Future improvements include adopt- ing advanced deep learning architectures such as EfficientNet and Vision Transformers for improved feature extraction. For audio processing, incorporating LSTM or transformer-based models could better capture temporal dependencies.

VI References

[1] AB Athira, Abhishek Tiwari, SD Madhu Kumar, and Anu Mary Chacko. Multimodal data fusion framework for fake news detection. In *2022 IEEE 19th India Council International Conference (INDICON)*, pages 1–4. IEEE, 2022.

[2] Junyi Chen, Leyuan Liu, Tian Lan, Fan Zhou, and Xiaosong Zhang. You only query twice: Multimodal rumor detection via evidential evaluation from dual perspectives. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3415–3427, 2025.

[3] An Lao, Qi Zhang, Chongyang Shi, Longbing Cao, Kun Yi, Liang Hu, and Duoqian Miao. Frequency spectrum is more effective for multimodal representation and fusion: A multimodal spectrum rumor detector. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 18426–18434, 2024.

[4] Hariharan RamakrishnaIyer LekshmiAmmal and Anand Kumar Madasamy. A reasoning based explainable multimodal fake news detection for low resource language using large language models and transformers. *Journal of Big Data*, 12(1):46, 2025.

[5] Han Li and Hua Sun. Multimodal rumor detection enhanced by external evidence and forgery features. *arXiv preprint arXiv:2601.14954*, 2026.

[6] Shiming Li. A multi-modal rumor detection model based on temporal graph attention network. In *2025 10th International Conference on Social Sciences and Economic Development (ICSSED 2025)*, pages 890–905. Atlantis Press, 2025.

[7] Hui Liu, Wenya Wang, and Haoliang Li. Interpretable multimodal misinformation detection with logic reasoning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9781–9796, 2023.

[8] Xin Liu, Mingjiang Pang, Qiang Li, Jiehan Zhou, Haiwen Wang, and Dawei Yang. Mvaclnet: A multimodal virtual augmentation contrastive learning network for rumor detection. *Algorithms*, 17(5):199, 2024.

[9] Jinna Lv, Yuan Gao, Li Li, Lei Shi, and Siyu Li. Multi-modal fake news detection: A comprehensive survey on deep learning technology, advances, and challenges. *Journal of King Saud University Computer and Information Sciences*, 37(9):306, 2025.

[10] Liwen Peng, Songlei Jian, Dongsheng Li, and Siqi Shen. Mrml: Multimodal rumor detection by deep metric learning. In *ICASSP 2023- 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[11] Kaixuan Wu, Yanghao Lin, Donglin Cao, and Dazhen Lin. Interpretable short video rumor detection based on modality tampering. In *Proceed- ings of the 2024 Joint International Conference on Computational Lin- guistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9180–9189, 2024.

[12] Ashima Yadav, Shivani Gaba, Haneef Khan, Ishan Budhiraja, Akansha Singh, and Krishna Kant Singh. Etma: Efficient transformer-based multilevel attention framework for multimodal fake news detection. *IEEE transactions on computational social systems*, 11(4):5015–5027, 2023.

[13] Yuxing Yang, Junhao Zhao, Siyi Wang, Xiangyu Min, Pengchao Wang, and Haizhou Wang. Multimodal short video rumor detection system based on contrastive learning. *arXiv preprint arXiv:2304.08401*, 2023.

[14] Long Ying, Hui Yu, Jinguang Wang, Yongze Ji, and Shengsheng Qian. Multi-level multi-modal cross-attention network for fake news detection. *Ieee Access*, 9:132363–132373, 2021.