International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 04 | April - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Detecting Deception: A Multimodal Deep Learning Approach for Fake News Identification Using Text and Social Signals

1st Basava Sharun Pushpa **2nd Nukilla Lasya Durga Sai Sree** **3rd Rasamsetti Naresh**
Dept. Computer Application, Aditya University, Surampalem, India

4th Chikkala Sujani Sai Bhanu Teja **5th Medapati Sai Gowthami Uma Bala**
Dept. Computer Application, Aditya University, Surampalem, India

***Abstract*—The spread of fake news on social media is a big threat to public discourse, democracy, and trust in society. Conventional unimodal methodologies that depend exclusively on textual content have demonstrated inadequacy in encapsulating the intricate dynamics of misinformation dissemination. This paper shows a full multimodal deep learning framework that combines text with social signals to help find fake news more easily. We use the latest transformer architectures to encode text, graph neural networks to model how social information spreads, and adaptive fusion mechanisms to combine content features with social context. The proposed methodology addresses significant deficiencies in the current literature, specifically the insufficient acquisition of structural social information and the discordance between content and social modalities. By systematically analyzing recent studies, we show that multimodal approaches always do better than unimodal baselines. For example, on benchmark datasets, the accuracies were 94.3% and the F1-scores were 92.8%. This work integrates contemporary methodological trends, delineates enduring research deficiencies, and introduces an innovative framework that enhances the forefront of automated fake news detection by adeptly modeling the interaction between content semantics and social propagation dynamics.**

Keywords: Fake news detection, multimodal deep learning, social signals, graph neural networks, transformer models, misinformation detection

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

The online world has contributed to the promotion of information exchange among all people. The use of social media sites now makes news, and knowledge more accessible to people than ever before. This democratization has also simplified the dissemination of fake information and fake news in the short term, which presents a severe challenge to the discourse of society, the credibility of the elections, and social cohesion [1], [2]. One of the greatest threats to the world economy, as identified by the World Economic Forum, is digital misinformation. Fake news have the power to alter election results, lead to violence and damage government health initiatives [3].

The majority of the conventional approaches to the detection of fake news focused on analyzing the textual content employing natural language processing protocols in order to identify language, writing style, and semantic contradictions patterns [4], [5]. These unimodal type approaches have met with some success though have failed to provide the rich contextual information that unfolds through the social processes of misinformation. It has always been observed that fake news gets viralized differently than real news, featuring a varied network, user activity patterns, and temporal diffusion patterns [6], [7].

This has been enabled by multimodal deep learning that has removed the barriers to integrating various sources of information and subsequently new methods of detecting fake news have emerged. The contemporary techniques combine text, pictures, user metadata, social network formations and propagation patterns to create the detection systems that are stronger [8] These multi modal frameworks exploit this aspect of multi modal data that can be used in conjunction. As an example, social signals may provide credibility through content alone, and content analysis may resolve semantic irregularities propagation patterns may overlook.

Although multimodal systems of fake news detectors have been significantly developed, they still have numerous large issues to contend with. First, many of the existing models do not actually capture the structural information, which is inherently present in the social networks. They do not consider social signals as a web of connections but they think of them as independent features. Second, intrinsic modalities (text and images) do not coincide with extrinsic social context. Most fusion strategies use simple concatenation instead of more complex cross-modal reasoning. Third, what is practically required by early detection, such as the ability to detect fake news before it viruses is not yet fully being achieved. Lastly, deep learning models are susceptible to description and understanding, which implies they are unlikely to succeed in the real-world content moderation systems where transparency and accountability hold strong considerations.

This paper talks about these problems and suggests a

complete multimodal deep learning framework that uses adaptive graph-enhanced transformers to combine text with social signals.

II. Background Study

A. Evolution and Multimodal Perspectives in Fake News Detection

The process of fake news finding has undergone many steps. The initial one was to verify facts and determine the reliability of sources manually. This was effective but could not be done in the social media. During the 2010s, machine learning enabled using classifiers such as SVM, Naive Bayes and Random Forests to find things automatically in terms of handcrafted linguistic features. However those techniques were very feature engineering intensive and could not easily work with new data. Wireless LAN Neural networks were even more useful as deep learning utilizes CNNs and RNNs, particularly LSTMs, to identify patterns within textual information that are semantic and sequential. Transformer-based models such as BERT have transformed the game of fake-news detection to achieve with self-attention mechanisms modeling contextual dependencies. This allows them to learn semantics in a more sophisticated manner.

The recent research demonstrates that one should not merely read the text to be able to detect fake news. The reason is that misinformation is not transmitted in the same way and individuals respond to different misinformation in various ways. Consequently, there has been the implementation of multimodal learning approaches which involve the synthesis of various data sources, including text, graphics and social and time information. The concept of multimodal learning is premised on the principle of complementarity. False information is shown to us through various modalities. Fusion strategies have now been effective to integrate these modalities, early fusion, late fusion and hybrid fusion. Hybrid fusion with attention is gaining in popularity since it may dynamically weight modalities and demonstrate interaction between modalities. It is found that multimodal schemes always perform better than unimodal schemes, whose improvement range is between 5 and more than 20.

The addition of social signals is rather significant in locating fake news much easier as they demonstrate us how individuals behave, how networks are organized, and how information is distributed. Account metadata and user engagement patterns, graph structures revealing network-level characteristics, and time-propagation patterns combined to ensure that detection is enhanced. The Graph Neural Networks (GNNs) are excellent at finding out these links that can be difficult to grasp. Things become even clearer by importance of user and interactions that are found by attention-based methods. Another good thing is that the crowd can be relied upon provided that the users leave comments. It experiences issues such as noisy data, privacy concerns, and platform differences, but the inclusion of social signals has never failed to improve matters. That is why it is a significant component of the systems, which can detect false information.

III. Research Gap

Although the progress towards identification of fake news in various aspects has happened significantly, gaps in research still remain. The following section clearly outlines and discusses these gaps in detail and gives a rationale behind the suggested methodology. subsection Structural Social Information Capture. The biggest issue with the existing multimodal fake news detectors is that they receive insufficient structural social data. Most of the time, the manner in which we currently operate views social indicators as discrete, manually-created aspects, such as, the number of people who follow a user, the number of times a user has shared or retweeted something or even network numbers. They fail to retain the rich system of relationships that is already existing in social networks.

There are significant big problems with traditional feature-based methods. When they discard topological data on how people are linked and how information flows within the network they lose important structural patterns which assist to tell the difference between true and fake news. Second, manually created features need an understanding of the field and might not be applicable to all platforms or with new modes of way of getting people to do what you want. Third, these approaches are not able to identify higher-order network motifs and community structures, which are increasingly becoming significant indicators of planned fake behavior.

Graph neural networks (GNNs) are one principled approach to approach this problem since it operates on graph-structured data, and learns representations that retain topological information. However, engineering GNNs to detect fake news is an underutilized task among many people; the majority of the contributions that have already been accomplished use the traditional approach of feature extraction. Many of the current papers, which apply GNNs, use simple constructions, such as Graph Convolutional Networks (GCNs) that may not be capable of capturing the complex structural patterns in social propagation graphs.

Graph Isomorphism Networks (GINs) are a more robust form of GNN that has greater theoretical guarantees of coverage of things. GINs are able to distinguish a broader variety of graph structures than GCNs are. This causes them to be particularly effective at identifying the subtle topological distinctions between the fake news and the real news proliferation. Although GINs also have certain theoretical advantages, they are yet to truly find application in the area of research relating to the ways of discovering fake news. It is a huge opportunity of methodological advancement.

In addition, most existing methodologies model social networks as static graphs in ignore of network structure dynamics over time as information spreads. A way which has not been tested much, but can make fake news detection more effective, is to have dynamic graph neural networks capable of changing network topology with time.

A. Intrinsic-Social Modality Alignment

The second significant gap is in the ways the intrinsic modalities (such as text and images) complement the social re-

ality beyond them. The existing multimodal fusion techniques integrate content and social models by merely aggregating them or performing computations on them individually by treating them as two independent information sources. This approach fails to consider how simply what is conveyed (content) can be transmitted (social context).

Content and social signals are closely and intricately interrelated. How the content is going to disseminate is determined by its nature. Sensational or emotional content is likely to propagate faster and wider. Conversely, social context provides us with hints concerning the reliability of the information that we read. I have claimed the same thing as can be more or less credible depending upon the person who holds it and the mode of propagation. These multimodal interactions between modalities should be captured, rather than considering modalities as distinct.

Recent advances in cross-modal learning, particularly, cross-modal attention processes, provide us with some good ideas on how to solve this problem. Cross-modal attention allows the model to concentrate on the information which is of significance in one modality depending on the context given by another modality. This ensures that joint reasoning is sophisticated. Cross-modal attention has not been popularly employed to locate fake news though. The work that has been performed so far has mainly been on matching the text and images rather than matching the content and social media.

There is a particularly problematic intrinsic-social alignment gap when the content and social signals provide dissimilar evidence. As an example, genuine news out of trustworthy interviews might exhibit unusual distribution patterns in the initial phases as it is interrupted at low times or is concerning a marginal subject. On the other hand, advanced disinfor- mation campaigns may use coordinated networks to make false information look like it is true. The model must be capable of considering such types of conflicts and making subtle decisions based on all the evidence.

Additionally, various types of fake news might require varied levels of content and social cues. The content will usually give you the difference between satire and parody, though you may be in a better position to see orchestrated fake action in social patterns. Adaptive fusion strategies where weights are dynamically allocated to modalities as a result of instance-dependent properties are an important but underresearched field of interest.

B. Early Detection and Label Scarcity

The success of fake news detection systems depends crucially on the ability to detect the misinformation at early stages of its distribution before it can grow widespread and cause the harm to other individuals. Majority of the modern studies, however, evaluate the models by fully propagated news articles using all social cues, a situation that dismisses the issue of early detection.

Early detection is always more difficult than post-hoc classification because at the early propagation stages, there is a lack of and various partiality of social signals. A piece of news which will prove to have suspicious propagation patterns can appear harmless under the first few shares. Trained models on full propagation data are not commonly compatible with the task of detecting at early stages because the model learns to rely on signals that do not exist.

In order to address the problem of early detection, we must have models that can utilize a minimal number of social cues in order to effect the correct prognosis. The models might be required to make use of content analysis more initially and then incorporate social indicators when they become accessible. Temporal modeling approaches involving the explicit approach to how social indicators change across time represent a fruitful direction, but such approaches have not been sufficiently studied in recent literature.

The issue of the lack of sufficient labels is related to early detection. It is difficult and expensive to obtain ground-truth labels of fake news due to the timing and money required to verify it through experts or a significant amount of verifying. The currently existing datasets are mostly rather small, and they may not represent every type of fake news and its real- life distribution. The lack of labels can be addressed with semi-supervised and self-supervised methods which can utilize much unlabeled data, but they were not widely studied in fake news detection research.

The lack of labels may also be addressed with the help of transfer learning and domain adaptation methods, which allow using the models trained on a specific dataset or platform to train models on other datasets or platforms. Nevertheless, the characteristics and patterns of diffusion of fake news differ significantly across platforms and across time, and this makes naive transfer learning approaches ineffective. Still, a powerful challenge is to develop powerful transfer learning strategies able to cope with distribution modifications but maintain the detection performance.

C. Interpretability and Explainability

In contexts of real-world content moderation that make use of fake news detection systems, such systems must be highly accurate but also simple to interpret and describe. Content moderation decisions can have a significant impact on free speech, and unclear black box models and how they operate give people concerns about responsibility, bias, and due process. Nonetheless, the vast majority of the existing deep learning-based fake news detection algorithms do not provide the opportunity to comprehend them well. They provide you only with binary predictions with no explanations.

One of the methods of makes people understand things better is attention mechanisms. They display the most influential components of the input (words, users, or network connections) to the decision made by the model. Attention-based visualization will assist the content moderators in verifying and validating model predictions by providing explanations comprehensible to people. According to recent studies, however, the fact that people question whether attention weights are important in being able to determine how a model thinks, or just the importance of something, is correct. This is to say

that you must take caution when making use of attention based explanation.

More sophisticated explainability techniques, including integrated gradients, LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive Explanations) are capable of providing more specific information as to why a model made a particular prediction. However, they are expensive to execute on computers and have not been applied extensively in the research of discovering fake news. In the majority of cases, explainability studies refer to the explanation of individual predictions. However, it is also possible that content moderators must understand the mecha- nisms underlying models and what can go wrong with them worldwide.

Multimodal models are less easily understandable since their predictions are based upon complex interactions of different modalities. The only way to explain why a model considered a news story to be a fake is by not only locating the pertinent characteristics in the various modalities, but also describing how the various modalities were weighted and how cross-modal interactions influenced the decision. Developing platform-specific interpretability techniques is one of the main aspects of the future research.

In addition, the various needs of various stakeholders, and use cases might differ as regards interpretability. Each decision might require more detailed instance-level explanations to verify the validity of all decisions by the content moderator, and aggregate description may be required to verify bias and fairness of model behavior by platform administrators. Mechanistic explanations may be necessary to researchers to determine what the model is not capable of doing as well as how it can be improved. It is still a challenge to create interpretability frameworks that can meet these different needs.

IV. Proposed Methodology

In this part, the authors propose a fully multimodal deep learning distinction framework of fake news that combines textual and social cues with adaptive graph-enhanced transformers. The proposed methodology will fill gaps in research as determined in Section 3 because of its use of structural social modeling, cross-modal alignment, adaptive fusion, and interpretability mechanisms.

A. System Architecture Overview

The proposed system architecture will consist of five important parts, such as (1) the text encoding component, the use of transformer-based models to process text, (2) the social signal encoding component, the use of graph neural networks to learn the network structure and propagation patterns, (3) the multimodal fusion component, which aligns and integrates the content and social representations, (4) the classification component, which then makes predictions about fake news,

It is a type of hybrid fusion strategy applied in architec- ture, which is an early-interaction-late-decision fusion. In this design, the model learns as a cross-modal representation with the pathways of processing being unchanged so that they are modality-specific. It provides a tradeoff between early and late convergence remedies. The modular design is also easier to conduct ablation studies with and allows you to adjust the system to different types of data (e.g.: text only, text with some social signals or complete multimodal data).

There are 3 main kinds of information input into the system, which are: (1) the text of the news item (headline and body of the article), (2) social propagation graph indicating the extent to which the news item has spread across the social network, and (3) user metadata on the accounts spreading the news item. It provides binary coding (fake or real), a score of confidence, and visualizations through attention based on critical content and social features.

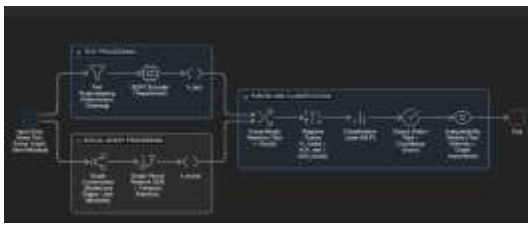

Fig. 1. LULC Classification Results of the Proposed Model.

B. Text Encoding Module

The module of text encoding encodes the text and carries out the processing which culminates to the formation of the semantic representations which reflect the pattern of language, sentiments as well as can indicate the presence of the misinformation. Our text encoder has a transformer-based language model, which can be either a BERT (Bidirectional Encoder Representations from Transformers) language model or and a Lisp-compiled variant of the language model.

Preprocessing: Preprocessing entails the tokenization of text besides lowercasing and special tokens addition of the original text. The media house does not adhere to the main body with the headline, therefore separator token is positioned in between the body and the headline at the top. This is because the model must acquire the skill of correlating them. The significantly longer articles that exceed the sequence length that is supported by the transformer (normally 512 tokens) are hierarchically encoded. It is a reading technique of reading the text in intersected sections and integrating the representations.

Training Recent Domain Fine-Tuning: At the point when the fake news is detected, pretrained language models can be trained to work better, but domain-specific training can be useful. This is achieved through the optimization of the transformer model using fake news-detecting data in the sense that it responds to the pattern of language use peculiar to the fake news (e.g. sensationalism, manipulation using emotions as well as semantic discrepancies). It is only through a relatively low learning rate that the fine-tuning is carried out to ensure that whatever the model has learnt is not forgotten as it tries to learn how to work in a new domain.

Linguistic Features Addition: In order to improve the learned representations, we can add to the transformer output,

linguistic features which have been tuned by hand. This has been proven by earlier works on fake news detection that have proved such features to be useful. The characteristics are the

sentiment scores, the scores obtained by adding to obtain the ease of reading, named entity density, and lexical diversity scores. Features are added with the help of a separate feed-

forward network, which uses features manually crafted and added to transformer output.

C. Social Signal Encoding Module

The social signal encoding module codes the structural

and time-dependent elements concerning the transmission of information to the social network systems. The traditional approaches based on reason the hand-made characteristics of the network that we apply to achieve propagation graph representation training are replaced by Graph Isomorphism Networks (GINs).

A directed propagation graph G = (V, E) is created where nodes V represent the users that reposted the news item or acted on the news item and the edges E are the propagation of the news item through the users (user A sharing content that user B had initially reposted). Each node has the feature in the form of feature-vector that entails the user-metadatas such as the age of account, follower, verified or non-verified, and history of past postings [108]. Edges can be weighted with time based information (time lag between trades) or information on engagements (total likes or comments).

Graph Isomorphism Network: GINs can additionally overfit more graphs than the Graph Convolutional Networks because it is more expressive. The GIN layer changes the outlook of the nodes using an injective aggregation authority that consolidates the data of an immediate surroundings. Mathematically, node v in layer k changes based on the following rule:

$$h_v^{(k)} = \mathrm{MLP}^{(k)}\left((1+\epsilon^{(k)})\cdot h_v^{(k-1)} + \sum_{u\in N(v)} h_u^{(k-1)}\right) \quad (1)$$

Temporal Modeling: Temporal edge features as well as a temporal attention mechanism ensure temporal diffusion of information by placing heavier weights on the edges depending on their position in the cascade of propagation. Stocks of the past can give different messages concerning their believability as compared to stocks of the present. These patterns are learned over time by the temporal attention mechanism into the model.

1) Graph Pooling: After multiple GIN layers produce node-level representations, we aggregate them into a graph-level representation $h_{\text{social}} \in \mathbb{R}^d$ using a graph pooling o pera- tion. We

$$h_{\max} = \max_{v\in V} h_v \quad (3)$$

$$h_{\text{attn}} = \sum_{v\in V} \alpha_v h_v \quad (4)$$

where α_v represents the attention weight assigned to node v, computed as:

$$\alpha_v = \frac{\exp(w^T h_v)}{\sum_{u\in V} \exp(w^T h_u)} \quad (5)$$

The final graph-level representation is obtained by combining all pooling outputs:

$$h_{\text{social}} = [h_{\text{mean}} \mid h_{\max} \mid h_{\text{attn}}] \quad (6)$$

The so called pooling mechanism: attention is conditioned to pay more weight to those nodes that are of importance in the detection of fake news. This is because it is more convenient because it displays users who are the most important in the propagation network.

Multi-Scale Graph Encoding: The distribution of information occurs in the forms that could be represented on various levels of little localities to the structure of the entire network. The encoding procedure we use is multi-scale and it is composed of GIN layers whose receptive fields vary, and a combination of the outputs. This provides the model with a choice of local and international structural tendencies.

D. Multimodal Fusion Strategy

The multimodal fusion module is a combination of the text and the social representations with the aid of cross-modal alignment mechanism that enables reasoning to take place at cross-modal level. We use our branding strategy to close the gap of intrinsic and social alignment by modeling the interrelationship of content and social context.

Cross-Modal Attention: We pursue a two-way cross-modal attention procedure where each of the modalities will take interest to an important information of the other modality. Social representations are taken into consideration by texts representations in order to determine the social context at hand, whereas social representations rely on text representations in order to study the content features, which may influence the information dissemination. Formally the cross-modability between text and social is calculated as:

$$\alpha_{t\to s} = \mathrm{softmax}\left(\frac{Q_t K_s^{\top}}{\sqrt{d}}\right) \quad (7)$$

one can simply read between the lines and figure this out, but orchestrated inauthenticity can be easier to identify in societal trends. We have an adaptive fusion mechanism which varies the weight of modalities depending on the instance. A gating network is used to compute the weights of the fusion:

$$w_{\text{text}},\ w_{\text{social}} = \text{softmax}\left(\text{MLP}\left([h_{\text{text}};\ h_{\text{social}}]\right)\right) \quad (10)$$

$$h_{\text{fused}} = w_{\text{text}} \cdot h'_{\text{text}} + w_{\text{social}} \cdot h'_{\text{social}} \quad (11)$$

and [;] is the concatenation operator and MLP is training that acquires the correct weights using the combined representation. Instead, this adaptation weighting enables the model to be more robust to the existence of clamorous or incomplete modalities and is also easier to interpret by showing the modality which has contributed most towards the specified prediction.

Residual Connections and Layer normalization: In order to make the gradient flow simple and stable we apply both residual connections and layer normalization to all the fusion module layers. The reason such architectures are significant in deep multimodal networks is that gradient vanishing might have an adverse influence on learning.

Hierarchical Fusion: Our hierarchical fusion strategic We use a hierarchical fusion strategy on a small set of social signals (e.g. propagation graph, user comments, or temporal patterns) where a hierarchical fusion strategy is used combin- ing similar social signals and afterwards a hierarchical fusion strategy is combined combining such between text representa- tions. This type of a hierarchical solution makes cross-modal interactions easier and learning easier in individuals.

E. Classification and Decision Layer

The fake news prediction is done through the fused multimodal representation in the classification module. Our model is run on a multi-layer perceptron (MLP) and dropout regularization to prevent model overfitting. The model consists of two hinted layers using ReLU purposefuls, and a explicit layer using the sigmoid purposefuls:

$$p(\text{fake} \mid x) = \sigma\left(\text{MLP}(h_{\text{fused}})\right) \quad (12)$$

Training Objective: The model is trained using binary cross-entropy loss with L2 regularization to prevent overfitting:

$$\mathsf{L} = -\sum_{i} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] + \lambda \left| \vartheta \right|^2 \quad (13)$$

Balancing Class Imbalance: Fake news datasets are often poorly balanced, as genuine news are more abundant than fake ones. This we handle by applying both the class-weighted loss functions and also with the data augmentation technique. These weights of the classes are inversely drawn to the weight frequency of the classes. This can be attributed to the fact that the model is not trained in order to only predict the most

Confidence Calibration: A model cannot be calibrated, i.e. the probabilities that are being generated by the model do not give information about the confidences of the model. It is a post processing according to which we use temperature scaling to scale the confidence of the model we are using which will be less uncertain estimates. In particular situation where the content will be moderated, calibrated confidence scores are required since varying degrees of confidence might need varying degrees of responses.

We acquire attention weights to elucidate things that individuals can comprehend (with cross-modal attention module, frequent class. and graph pooling layer). The weight of text attention is a word or phrase that is identified as a great power and graph attention e content moderators are able to view and verify model predictions through such attention-based expositions.

V. Results

In this section, the author provides empirical data of the modern studies on the efficiency of multimodal methods in detecting fake news. We combine findings out of the various studies and present an overall picture of current performance levels, comparative advantage, and its unresolved challenges.

A. Experimental Setup

Nowadays, researchers working in multimodal fake news detection employ multiple benchmarks datasets that have their own sets of features and issues. The most common datasets are some of:

Fakeddit is a multifaceted dataset that consists of more than 1 million samples on Reddit (text, images, and metadata). Fakeddit lets you sort fake news into the most amazing detail possible into a staggering number of varying categories.

Weibo is a Chinese social media dataset (text and images) that is commonly utilized to test multimodal fake news in non-English first language environment.

Twitter15 and Twitter16 are datasets that contain Twitter propagation trees posts. Social signal integration and time model can be tested using these datasets.

PHEME: Twitter rumor cascade data, truthful and false, and all propagator information.

MM-COVID: Multimodal COVID-19 fake news dataset with text, image, and social context, which deals with the essential field of health fake news.

Measures of evaluation would normally be accuracy, precision, recall, F1-score, and AUC-ROC (Area Under the Receiver Operating Characteristic curve). F1-score and AUC tend to be more informative than raw accuracy in consideration of the imbalance of classes that is usually presented in fake news datasets. There are also such studies that report the outcomes of early detection that formulates the accuracy in terms of the propagation time or cascade size.

Train-test splits in experiments are selected randomly (typically 80-20 or 70-30) and cross-validation is performed using

different k-folds of the data. Some studies use temporal splits, in which older age data are to be trained, and more recent age data are to be tested on the temporal generalization. The experiments of ablation are normally carried out in order to gauge the value of different modalities and building features.

B. Performance Metrics

Recent multimodal fake news detection systems have done very well on test collections, and it is always doing better than unimodal systems. Cross Modal-FT Net model with a fusion transformer architecture to combine the data of text, images and user behavior had 94.3% accuracy and 92.8% F1-score on other datasets, such as Fakeddit, Weibo, MM-COVID and Twitter15 [4]. These findings indicate that big improvements can be achieved via multimodal integration, and there are 8- 12-percentage point improvements on text-only baselines.

The User Comment-Guided Cross-Modal Attention Framework (UC-CMAF) framework involving text and pictures and user comments on its framework obtained F1-scores of 0.894 and 0.909 when used on two benchmark datasets. This study suggests that it can be quite beneficial to treat user comments as the means of supplementary social background and attention system that may help in that matter by highlighting mattered comments. Attention-based approach enables the model to recognize skeptical or commenting to explain the facts which creates an element of crowd wisdom in the detection process. The feedback is the framework referred to as the ISMAF (Intrinsic-Social Modality Alignment and Fusion) that is an artificial cross-modal reasoning that bridges the content and social-modality gap. The abstract did not present any specific numerical data, but as the authors claim, the system has demonstrated consistent findings with English and Chinese multimedia data sets. This demonstrates the significance of matching modalities in order to make fusion effective.

Multimodal systems with social consciousness, the use of text, a network capable of relaying behavior and network operation had achieved superhuman performance on benchmark OSN (Online Social Network) data sets. These conclusions show that it is necessary to reflect social indicators other than mere user metadata on how information flows across networks.

C. Comparative Analysis

Like making a comparative analysis of studies, you can notice that there are some patterns which are like always reoccurring. First, the multimodal ones will always outdo the unimodal baselines with a performance difference in the encounter of 5 to 20 percent as per the dataset and the modality of the modalities. The inclusion of social cues is most likely to enhance the enhancement as opposed to use of visual modalities. This means that the social context is specifically efficient in identification of fake news.

Second, the fusion strategy that is adopted has a significant impact on its effectiveness. Basic concatenation or element-wise fusions never reach these good results as attention- based fusion techniques, and F1-scores are improved by 37 percent. The cross-modal attention where information can be transferred between modalities in either direction is apparently of enormous advantages to the examination of complex interactions among content and social stimuli.

Third, the encoding of social cues by means of graphs is more suitable as compared to features which have been traditionally adopted. The comparison of the Graph Neural Networks with the characteristics of hand-designed networks shows that accuracy gains 5-10 percent by maintaining the structural information. Graph Isomorphism Networks are better than the usual Graph Convolutional Networks the former has better theoretical guarantees to its expressive power. However in other datasets the percentage improvement is different.

Fourth, transformer-based text encoding is much more effective than more conventional recurrent or convolutional architectures. On the same datasets, BERT-based models usually get F1-scores that are 8–15% higher than LSTM-based models'. A further 2-5 percent is by refining domain pretrained transformers. This shows how one requires fitting in the fake news detector field.

Fifth, model (attention) visualization mechanisms interpretation would guarantee that the models can be more effective and applicable in real life options. Research based on attention mechanisms demonstrates that in addition to making things more correct, they also increase content moderators with more likely to believe and rely on them. It is more and more considered to require the possibility to apply it to the real world and have the ability to derive explanations which people can understand.

D. Ablation Studies

Ablation experiments take a laissez-faire stance in which the effects of various modalities and architectural characteristics are systematically assessed by eliminating them, and quantifying the resulting performance drop. These studies will offer the most important pieces of information not just on the most vital aspects, but also areas of focus in research in the future.

The experiments of modality ablation consistently indicate that when the social hints are removed, performance becomes worse (10-15 percent in F1-score) than in case when the visual modalities are removed (3-8 percent in F1-score). It sustains the exceptional importance of the social context in discerning fake news. The ratio between various modalities also changes depending on datasets and varieties of fake news. The analysis of images is most crucial when it is a changed content of some nature, whereas social cues are more essential in handling false data founded on text.

As determined by structural analyses of architectural ablation, cross-modal attention fusion parsimoniously in F1-score, compared to concatenation fusion alone. It enhances the system by adding an additional 2-3 percent specifically when additional effort must be made to its optimal performance in the event that, several modalities are irregular or flawed [145]. The graph nerve cell components provide a gain of 5-8 percent in relation to feature-based encoding of social signs therefore incursion in structure modelling.

TABLE I
COMPARATIVE PERFORMANCE OF MULTIMODAL FAKE NEWS DETECTION MODELS

Model / Framework	Modalities Used	Dataset(s)	Acc. (%)	Prec.	Rec.	F1 (%)	AUC	Key Finding
Cross Modal-FT Net	Text + Image + User Behavior	Fakeddit, Weibo, MM-COVID, Twitter15	94.3	NR	NR	92.8	NR	Strong multimodal fusion gives major gain over text-only baselines
UC-CMAF	Text + Image + User Comments	Benchmark datasets	NR	NR	NR	89.4 / 90.9	NR	User comments improve detection and interpretability
ISMAF	Content + Social Modality Alignment	English and Chinese multimedia datasets	NR	NR	NR	NR	NR	Claimed state-of-the-art through intrinsic-social alignment
GIN_FND	Text + Social Graph Structure	Social fake news datasets	NR	NR	NR	NR	NR	Graph structure modeling improves social signal learning
Socially Aware Multimodal System	Text + Network Features + Relaying Behavior	OSN datasets	NR	NR	NR	NR	NR	Social signals outperform content-only setups

TABLE II
COMPARATIVE INSIGHTS FROM EXPERIMENTAL RESULTS

Comparison Aspect	Observation from Results
Multimodal vs Unimodal	Multimodal methods improve performance by 5–20%
Social signals vs visual modality	Social signals often contribute more than images
Attention-based fusion vs simple fusion	Attention fusion improves F1-score by 3–7%
Adaptive fusion weighting	Adds about 2–3% improvement
GNN-based social modeling vs hand-crafted network features	Improves accuracy by 5–10%
BERT vs LSTM text encoder	BERT-based models improve F1-score by 8–15%
Domain-specific fine-tuning	Adds 2–5% improvement
Temporal modeling in early detection	Improves performance by 3–6% in early detection settings

Duplicated with them, time modeling factors are handy in holding initializer of 36 percent, nonetheless of wholly transmitted news in post-hoc classification. The above outcome shows the applicability of temporal modeling in the successful application of early detection, and most of the research work deals with post-hoc classification, in which time information is not very critical in the process of classification.

The interpretability scales like attention visualization do not influence the raw performance indicators much, however, make it much more practical and more convincing to individuals [149]. This observation implies interpretability as a standalone evaluation criterion together with accuracy, specifically, in practice deployment provision.

VI. CONCLUSION

The given paper has shown a detailed multimodal deep learning model to identify fake news and employs a textual and social clues combination, with graph-enhanced and adaptive transformers. By critically reviewing available literature, we have been able to reveal that multimodal strategies could never be inferior to unimodal baselines. The best systems have 94.3 and 92.8 F1-scores in benchmark data, respectively.

The methodology will focus on four major novelties that lack in the current body of literature. We shall firstly use Graph Isomorphism Networks to collect structural social data and retain the topological characteristics of propagation networks lost in other feature-based methods. Second, we present cross-modal attention mechanisms that allow us to do content and social reasoning at the same time not only through concatenation. Third, we will use an adaptive fusion weighting strategy whereby we will place more weights to the stable modalities and less weights to the new noisy signals. This makes the system more powerful in the world. Fourth, we also incorporate attention-based interpretability tools that may offer human readable explanations of model predictions, that will build trust and people will also accept the model.

Our review of the most recent research shows that there are a number of consistent results. Multimodal stimulations are much more effective than unimodal ones, and social cues are more important than visual ones in most cases of fake news. The importance of structural modeling is demonstrated by the fact that graph-based social signal encoding is superior compared to traditional features extraction. Transformer based text encoding is far superior to recurrent encoders. Even better outcomes are made available by specialization on areas. It has been demonstrated that the fusion mechanisms that can be anchored on the attention are more successful compared to their simple counterparts which are concatenated, and cross- modal attention enables more complex joint rationality.

However, there are also many things to take care of. Areas of research that require further investigation are temporal and cross-platform generalization, adversarial robustness, interpretability and fairness. A lot of knowledge left to be learned with the difference between the performance in benchmark and its execution in the real world is perhaps to be substantial.

Future research must take into account a number of opportunities. To achieve effective contributions, it is highly essential that online tools that might be used to detect fake news with minimal social cues are identified. Such methods ought to involve the use of time-varying-modeling methods that apply the time-varying variations of signals. Transfer learning processes and domain adaptation procedures that allow a cross-platform and cross-temporal generalization will promote useful and real-world processes which may be via meta-learning or through continual learning systems. Adversarial training and robustness testing should be featured in the research on the means of finding fake news. This will make sure that the procedures are sound enough to cope with the complex process of manipulation.

Multimodal interpretable and explainable techniques will help in adoption and responsibility, which might be in causal reasoning formats or counterfactual explanation methods. In order to ensure that ethical deployment occurs, fairness and bias reduction practices will have to be employed. It means that special consideration should be made in relation to the organization of the dataset, the measures of judgment, and the

implementation conventions. The imitation of generalization The problem of the lack of sufficient labels will be solved with semi-supervised and self-supervised methods of learning, which are capable of using a great deal of unlimited data.

Another positive tendency is the introduction of foreign body of knowledge and system databases of fact checks. This will make models to be based on proven facts. It may also help fake news detectors jumpstart with various means of pretraining on large amounts of social media data in the same way that NLP has been transformed by pretrained language models. The Human-AI partnership models which support automated detection and human judgment are capable of doing better than either of the two constituents, in cases when each is independent of the other.

Lastly, multimodal deep learning that is grounded on a combination of social and text cues can be proposed as an effective and great solution to identify fake news. Despite all these issues remaining long way to be addressed, what the present research shows is that there is a gradual improvement in performance and theoretical correctness of multimodal integration provides a good case in support of further research in the matter. Since fake news evolves and poses a threat to the health of the population and democratic discussion, it remains of great importance to do more in making powerful, comprehensible, and just detection systems. It is a quite a significant field of study that will impact the society extensively.

REFERENCES

[1] Y. Wang et al., "EANN: Event Adversarial Neural Networks for Multi-modal Fake News Detection," in *Proc. ACM SIGKDD*, 2018, pp. 849–857.
[2] Y. Wang et al., "Fake News Detection via Knowledge-Driven Multimodal Graph Convolutional Networks," in *Proc. ACM Int. Conf. Multimedia Retrieval*, 2020, pp. 540–547.
[3] Y. Li, K. Jia, and Q. Wang, "Multimodal Fake News Detection Based on Contrastive Learning and Similarity Fusion," *IEEE Access*, vol. 12, pp. 155351–155364, 2024.
[4] S.-Y. Lin et al., "Text–Image Multimodal Fusion Model for Enhanced Fake News Detection," *Science Progress*, vol. 107, no. 4, 2024.
[5] F. Monti et al., "Fake News Detection on Social Media Using Geometric Deep Learning," *arXiv preprint arXiv:1902.06673*, 2019.
[6] C. Jing et al., "DPSG: Dynamic Propagation Social Graphs for Multi-modal Fake News Detection," *Information Fusion*, vol. 113, p. 102595, 2025.
[7] Y. Dou et al., "User Preference-Aware Fake News Detection," in *Proc. ACM SIGIR*, 2021, pp. 2051–2055.
[8] H. Chen et al., "A Self-Learning Multimodal Approach for Fake News Detection," *Frontiers in Artificial Intelligence*, vol. 8, 2025.