Fake News Detection Using Stance Extracted Multimodal Fusion-Based Hybrid Neural Network

Sudhakar Sengan, *Member, IEEE*, Subramaniyaswamy Vairavasundaram, Logesh Ravi, Ahmad Qasim Mohammad AlHamad, Hamzah Ali Alkhazaleh, *Member, IEEE*, and Meshal Alharbi

***Abstract*— Public and governmental concerns over online rumors' widespread diffusion and deceptive impact on social media have increased. For users to obtain accurate information and preserve social peace, finding and controlling social media rumors is challenging. Automatically, identifying fake news (FN) is a critical yet challenging topic that is still little understood because the consequences are so high. The text, visual features, the acceptance of the user's reply, stance, and social context are a few aspects of FN that are universally acknowledged. Current research has concentrated on modifying results to one specific trait, which has been partially the reason for their success. This article proposes Fakefind, a convolutional neural network (CNN) + recurrent neural networks (RNNs) hybrid model that integrates multimodal features for efficient rumor detection (RD). Additionally, the stance is extracted from indirectly implied postreply pairs using a CNN-based knowledge extractor (KE), and the stance representations are integrated for FN detection (FND). Extensive research findings are based on three multimedia rumor datasets from Weibo, Fakeddit, and PHEME. The outcomes show how well the recommended Fakefind identifies rumors with multimodal content.**

***Index Terms*— Convolutional neural network (CNN), fake news detection (FND), multimodal fusion, natural language processing, recurrent neural networks (RNNs).**

I. Introduction

FAKE news (FN) has been widely used in recent years to spread political propaganda, impact election results, or harm an individual (or) group. It has brought FN propagation increasingly closer to the spotlight. For amplifying FN on social media in the format of images, texts, audio, or video, large, sophisticated applications (bots) are set up in networks and distributed. Frequently, these botnets were collected by foreign state actors looking to hide their origins. FN has spread fast in one decade, with the 2016 United States elections as a notable example [1]. Multiple challenges have arisen due to the typical spread of sharing articles online in politics and areas, such as sports, health, and science [2]. The stock exchange is one such area impacted by FN [3], where a rumor can have extreme values and even put the call to a complete stop. This work's information significantly impacts feasibility studies for decision-making and shapes our horizons. Users have increasingly shown signs of acting stupidly in response to news that becomes fake next. FN about the novel coronavirus' origin, nature, behavior nature, and behavior was one recent instance, and they spread online. More people were alert of the fake online content, which degraded the issue. It can be a daunting task to find such news online [4].

FN detection (FND) is done by checking the news' consistency concerning multiple domains, like the technical background, determining the actual sender (or) social/judicial background (e.g., what would be the intention of the FN, such as harming a person/group); as a result, fact-checking requires the information of some contexts and the availability of reliable sources. Even the human eye finds it difficult to tell the transformation between true news (TN) and FN; for instance, one study found that when participants were shown an FN article, they believed it to be "'*somewhat' OR 'very' accurate*" 75% of the time" and additionally discovered that 80% of high school pupils found it difficult to tell whether an article was fake [5].

The text of an article is its most natural characteristic. Advice for the media ranges from figuring out if this article's headline and body make sense to figuring out how clear and reasonable it is. Efforts to systematize the assessment of text have taken the form of contemporary natural language processing (NLP) and artificial intelligence (AI) algorithms that use manual and data-specific textual features to classify a part of the text as true (or) false. Because FN language behaviors are unclear, these methods are confined.

The risks associated with FN are becoming more prevalent. It can significantly change all businesses, not just those on online social networks (OSN) like Twitter and Facebook. The risks of FN can indeed impose hidden costs on businesses. "*Trust is a necessity in business. It decides brand success. Without it, brands will struggle to market and protect themselves*" [6].

Manuscript received 23 December 2022; revised 5 March 2023 and 10 April 2023; accepted 18 April 2023. Date of publication 5 May 2023; date of current version 2 August 2024. *(Corresponding authors: Sudhakar Sengan; Subramaniyaswamy Vairavasundaram.)*

Sudhakar Sengan is with the Department of Computer Science and Engineering, PSN College of Engineering and Technology, Tirunelveli, Tamil Nadu 627152, India (e-mail: sudhasengan@gmail.com).

Subramaniyaswamy Vairavasundaram is with the School of Computing, SASTRA Deemed University, Thanjavur, Tamil Nadu 613401, India (e-mail: vsubramaniyaswamy@gmail.com).

Logesh Ravi is with the Centre for Advanced Data Science, Vellore Institute of Technology, Chennai, Tamil Nadu 600127, India (e-mail: LogeshPhD@gmail.com).

Ahmad Qasim Mohammad AlHamad is with the College of Business Administration, University of Sharjah, Academic City, Sharjah, United Arab Emirates (e-mail: aalhamad@sharjah.ac.ae).

Hamzah Ali Alkhazaleh is with the College of Engineering and IT, University of Dubai, Academic City, Dubai, United Arab Emirates (e-mail: halkhazaleh@ud.ac.ae).

Meshal Alharbi is with the Department of Computer Science, College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, Al-Kharj 16278, Saudi Arabia (e-mail: mg.alharbi@psau.edu.sa).

Digital Object Identifier 10.1109/TCSS.2023.3269087

The characteristics vary among distinct types of FN subjects, including social media platforms (SMPs). The reaction that such a news article was supposed to elicit is a second characteristic. Advice columns exhort readers to reflect on the way a story helps them feel that it evokes hate or another intense emotion. Advice is based on the comment that FN frequently uses discriminatory and aggressive language to attract clicks (or) confuse [7]. The New York Times provided instances of people making money by publishing FN stories online; the more distressing the report, the better the response and the profit. Automated response detection frequently applies manual features, such as Facebook likes, in combination with conventional classifiers or simulates the spread of FN as an epidemic on the social graph. However, accessibility to the social graph may not be practical, and manually selecting features is labor-intensive.

The different information modalities present more proof of the occurrence of news events, and new possibilities for detecting features in FN are made available: 1) to begin, keeping the unique features of each modality while fusing the relevant data across multiple modalities is a challenge for the multimodal FN detection task; and 2) second, the performance of models is negatively impacted by noise data created during the information fusion between multiple modalities for certain types of news.

Deep neural networks (DNN) were created to address the problems associated with traditional methods because of their powerful and practical capacity to detect complex patterns automatically. Using textual features to detect FN is the primary focus of most current research experiments. The fact that OSN news now increasingly includes data in multiple media formats, including text messages, images, and video files, is a concept that is not forbidden. The relationships between the multiple modalities are characterized as complementary and have improved. A larger audience is likely to pay much more attention to news that includes video data, which increases its overall scope. However, a minimum of research has been done to exploit visual data to confirm the veracity of the news.

Due to other factors, FND from OSN seems to be quite challenging.

1) Collecting FN data and labeling the FN manual is challenging. Personal data include the news in the information feeds on Facebook/Twitter. Few large-scale FND public datasets exist for this context yet. Few news datasets that are accessible online only include a few instances, which are lacking for training the generalized model to the application.

2) FN is created through humans. To avoid being detected, most liars use crafty language techniques. Despite efforts to keep their speech under control, language "leakage" happens with verbal features that are difficult to check, such as pronouns, harmful emotions, word usage, and conjunction usage patterns and frequency.

3) A bottleneck for FDN is the limited data representation of texts. The bag-of-words (BOW) method aggregates and analyses word frequencies or "*n-grams*" (multiwords) to find deceptive clues.

OSN user responses have auxiliary data, which helps analyze FN. User responses and information spreading patterns are seen as having more excellent identifying signals than the information's content because, primarily, they are harder to use as indicators of veracity than the information's content. Interactions of users (likes, responses, shares, or comments) are a form of secondary knowledge that has a wealth of data that has been clustered in a broadcast tree, which proves the direction of data flow, timestamp data on interactions, textual data on communications, and user profile data on the users who participate in communications. The most common classification includes the following: help, question, reject, help, and analysis [unrelated (or) neutral].

The performance measures of models created just from news content are unsatisfactory, and it recommended that users' public behaviors are added as auxiliary data to enhance the FND task. To measure users' trust in FN, researchers created real-world datasets and chose distinct classes of "experienced" and "naive" users.

Researchers [8] also compared implicit and explicit user profile features. News content-based research is well-researched. However, selecting news content for FND presents too many problems. One of the key roles of this method is that news articles are purposefully written for fake readers. Therefore, this article contends that attacking the mitigation of FND from the perspective of news content is not easy. The result of the social context is supplying auxiliary information that might lead to higher performance results. However, this work provides an alternative method of using social context data in the FND process (i.e.,) using user response data as the primary input for FND. The input data are shown as text and visual emotions.

In earlier research, learning intricate and scalable textual or visual features was limited to manual-rafted features. However, current fusing methods are hypothetical and may not correctly combine many modalities. This work proposes "Fakefind," an end-to-end hybrid convolution neural networks (CNNs) + recurrent neural networks (RNNs) with an attention mechanism to fuse features from the text, image, and social context for the RD challenge, considering these restrictions and this research aims to use multimodal content. In the proposed method, this work first learns the shared text and social context (TSC) data using an RNN, followed by a CNN-text embedding. This research provides a knowledge extractor (KE) to model the postures implicitly. This research creates postreply to pairs and creatively uses a text-social context-opinion encoder to obtain the stance representations without manual labeling. This work recommends applying an opinion convolution network (OCN) layer after the text social context-opinion (TS-O) encoder creates a simultaneous TSC opinion embedding to extract stance-indicative features [9].

High-dimensional data are encoded using TS-O and have encoders, codes, and decoders. Encoders generate code from compressed input. Decoding the code replaces the input data. Learned generative data models were the pioneers of this method. Many unsupervised learning tasks, such as high-dimensional data analysis, FE, efficient coding,

sequence-to-sequence modeling, wavelet transform, anomaly or outlier detection, and so on, require multiple uses of the TS-O encoder. Like a single-layered autoencoder with a linear activation function, principal component analysis (PCA) minimizes the dimensionality of big datasets.

The stance-indicative features are extracted using an OCN layer after the TS-O encoder creates a simultaneous TSC opinion embedding. This article uses the bi-directional recurrent neural network (Bi-GRU), an RNN version of the model, to show how graphical features relate to textual or social data. Next, a multilayer perceptron (MLP)-based binary classifier is used to classify the findings from the study. This study provides a KE for implicit modeling postures. This study generates postreplies to pairs and implements a novel text-social context-based opinion encoder to obtain stance representations without manual labeling [10].

As a result, a thorough stance representation is reached for detection. Then, they are combined with image visual attributes that a pre-trained deep CNN represents. To capture the relationships between graphical features and combined textual or social information, this article uses the bidirectional gate recurrent unit (Bi-GRU), an RNN version of the model. The findings are then classified using an MLP-based binary classifier.

This research article is structured as follows. Section II presents the problem definition of the work, Section III presents the proposed Fakefind model, Section IV presents the tested result and discussion, Section V presents the performance analysis, and Section VI concludes the work.

II. Background Studies

Many computational methods exist to find articles as fake based on their documentary content. Most of these methods use websites for fact-checking, like "PolitiFact" and "Snopes." Lists of websites considered ambiguous and fake are available in other repositories maintained by researchers [11]. However, detecting articles (or) websites as fake requires a basic level of skill (or) knowledge, which is the weakness of these resources. More significantly, the fact-checking web contents only feature articles from respective fields, like politics, and are not designed for FND from distinct locations, such as sports, entertainment, and technology [12].

Several research on OSN sites like Facebook, Twitter, and other similar ones has focused on identifying and classifying FN. The concept of typing FN into different forms has been developed; the information is then expanded to generalize machine learning (ML) models for other fields. The study extracted language features from textual articles, such as n-grams, and trained a standard ML, such as K-nearest neighbor (KNN), support vector machine (SVM), logistic regression (LR), linear SVM (LSVM), decision tree (DT), and stochastic gradient descent (SGD), with SVM and LR achieving the highest accuracy (92%) [13]. The study found that overall accuracy declined as the sum of "n" in the n-grams generated for a specific thing. The phenomena have been noted for classification-based learning methods and incorporating textual features using supporting information, including user social activities on OSN, that improved the accuracy of multiple models. This article also covered psychological and sociological theories and showed how to use them to spot FN online. This work also covered data mining methods for developing models and presented feature extraction (FE) methods. These models use information, like script style, and public contexts, like posture and spread.

Most automatic rumor detection (ARD) techniques are text and social-based. Classification and graph-based optimization methods prove online textual posts using manually created textual and social context features. Only recent research has tried to find web-based rumors. Such works' visual features were developed manually, but they consisted of pre-existing factors that were combined, and feature-chained (early fusion), and averaged (late fusion). Manual features in older works help to learn both hard-to-read text and scalable graphics. Since current fusing methods are still experimental, different techniques may not be compatible [14].

Two types of FND techniques are now in use: propagation and content-based. Earlier works that used content-based methods considered posts having images, replies, external knowledge, or stances suggested in postreply to pairings. There are six postreplies to pairs whose indicated viewpoints are classified as supporting, denying, or commenting. Earlier research studies [15] have shown that stance information can enhance FND performance by manually labeling stances, followed by training the stance classification task and the FND task concurrently by multitask learning, but identifying the perspective by manually labeling takes considerable time and effort.

The data are true (or) false. The specific features of FN set it apart from the scope of TN by analyzing the linguistic features of articles from different FN websites. This article then compared those characteristics to articles from reliable journalistic websites. Its findings imply that FN reports are extended, have more capitalized phrases, and hold fewer stop words. The FN articles used more social terms, temporal, and spoken words. These words advise whether the text is more concerned with the present and the future than becoming truer and aim. According to [16], deceptive stories had fewer phrases, lower semantic intelligence, negative words about feelings, and more words of sequence. It alleged that 13% of 1600 news articles had incoherent headlines and content.

In each investigation, features are assumed from news content being used for FND. This article identified a few research works that analyzed user reactions when solving the FND task, despite users' replies being true responses toward news content with no desire to two-time others. Features taken from user responses were used as an added source of complete data that took user responses into account. Users are engaged with news posted on OSN sites (Twitter and Facebook). Comments express emotions about the post.

Machine translation technological advances have difficulty maintaining sentence definition, grammar, and structure. The statistical ML collects as big data as it can find that initially appeared parallel between two languages and crunches it to decide the likelihood that something is in OSN news: 1) relates to something in OSN news and 2) the new machine translation system (MTS) based on artificial neural networks (ANNs) and

deep learning (DL) was founded by Google in 2016 about the latter [17]. Recently, machine translation quality has been reviewed automatically by correlating hypothesis translations to match.

Moreover, SMP propounds a richness of multimodal information in texts, images, and videos used in OSN research. Microblogs' social connection features provide rise to posts with rich social context. Some social context features have been developed to track microblog communications, such as the number of retweets and replies. OSN features like hash-tag topics (#), mentions (@), and uniform resource locator (URL) have influenced the development of other social context features. Compared to pure text, certain current studies try to confirm the legitimacy of multimedia content. According to a survey by Morris et al. [18], a user's profile picture can help to decide their posts' credibility. Some primary features for images connected to tweets have been implied in the literature.

According to the literature reviewed above, it is critical to consider the position of the post and response aspects of news feeds to decide whether the information is accurate, but finding the attitude by hand takes much time and effort. This article presents a problem for automatically modeling the stance information implied in postreply pairs and entirely using stance representations, thus helping to detect FN. A first attempt to find rumors based on multimedia data that include social context has only been completed in a few recent research. Additionally, the manually created visual features employed in these works are deeply fused with already-existing elements using feature concatenation (early fusion) or outcomes averaging (late fusion) [19].

III. Proposed Fakefind Model

A. Problem Statement

This article is a multimodal news article like a tuple $N = (P, R)$, in which "P" stands for a post and "R" stands for replies. This article then displays the position as just a set of images, text, and social context tuples $P = \{(P_i^T, P_i^I, P_i^S)\}_i^m$. $R = \{(R_1^T, R_1^S), (R_2^T, R_2^S), \ldots, (R_N^T, R_N^S)\}$, where "$N$" $\rightarrow$ post response $i \rightarrow$ iteartions, $T \rightarrow$ text, $I \rightarrow$ image, and $S \rightarrow$ social context. This work requires finding if the news articles in "N" are FN or not in the FND problem. The label set is represented as $Y = \{[1,0], [0,1]\}$, where [1,0] shows TN and [0,1] means FN. A set of features are derived from the text, image, social context information in the post, and a reply depending on the news articles.

In contrast to the traditional multimodal methods that study only textual, visual, and social context features from posts and replies are involved in the proposed model, in addition to text and image. As shown in Fig. 1, the proposed Fakefind model takes the TSC from posts (P^T, P^S) and reply (R^T, R^S) and images from the post (P^I) as an input. The words and sentences in the post and reply are embedded separately using a *"hierarchical convolutional attention network* (HCAN)." The embedded text is then fused with the corresponding post and reply social contexts, and the resultant is a joint representation of these modalities in the post (FP^{TS}) and replies (FR^{TS}). The post and reply representations are paired

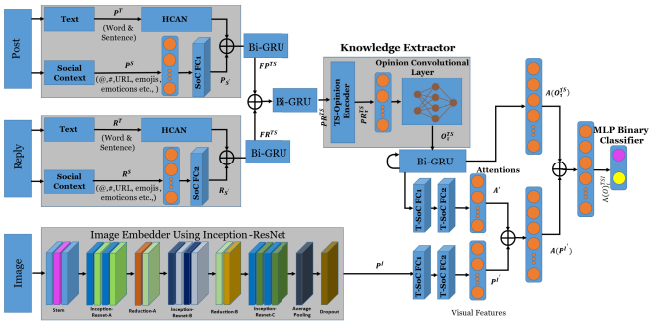

Fig. 1. Fakefind framework.

together (PR^{TS}) using a gated recurrent unit (GRU) before feeding an opinion-based KE block, which has text-stance opinion encoder (TSO encoder) layer and an opinion convolutional (OC) layer. The TSO encoder extracts the target ($\text{PR}_\tau^{\text{TS}}$) representation from PR^{TS}. The OC layer then controls the stance label after extracting stance-indicative feature vectors for assorted stances about the feature sequence (O_τ^{TS}). In a parallel stream, the visual features P^I are embedded using a hybrid inception-ResNet layer (IR layer), using an attention layer to refine $A(P^{I'})$. O_τ^{TS} and $A(P^{I'})$ are concatenated as the multimodal representation $[A(O)_i^{\text{TSI}}]$ and are fed as an input to an MLP-based binary classifier on deciding the authenticity of the news.

This study proposes a word embedding-based automatic summarization method for finding, ranking, and concatenating key Top-y sentences that serve as a reference summary. Examining the presence of frequent and pragmatic words is most frequently used in literature to evaluate the sentence's relevance. A sentence including more of these words is therefore regarded as more critical. However, this work contends that this method has two weaknesses: 1) it motivates redundancy in the new summary and 2) does not give meaningful sentences that encompass other words with proper weight. As a result, a redundancy detection mechanism (RDM) and a meaning-oriented sentence relevance assessment method are needed [20].

The first task is removing unnecessary tokens like stop words from the original document. The words from the first sentence were then used in conjunction with other essential words as keywords in this article. A linguistic study shows an explicit statement, recommending that the important words may be present in the first sentence, which is why this work tends to focus on the words of the first sentence. A few words from the title may also appear in the first sentence. Text is among the interacting modalities that encompass multisensory perception. Text modalities are used in a text recognition task, so for the sake of simplicity, this work addresses the following temporal multimodal problem.

1) Take input streams from multiple modalities as an example: the statements $\{X_a = N_1, \ldots, N_T\}$ and $\{X_a = N_n, \ldots, t_n\}$ denote the n-dimensional feature vectors of the $X_a X_a$ modalities existing at the time "t," respectively.

2) Following that, at the time "t," this study integrates the two modalities and assumes the unimodal output distributions at multiple levels of representation.

3) With the ground truth labels $Z = (Z_1, \ldots, Z_t)$ and $Z = (Z_1, \ldots, Z_t)$, researchers choose to train a multimodal learning model M that maps $X_a X_a$ to the same classification set of Z.

4) Where $\chi t_a \in Ri\chi at \in R_i$, each parameter of the input text stream $\chi t_a \chi_a t$ is synchronized differently in time and space, respectively.

5) It allows us to create two distinct unimodal networks from $X_a X_a$, each denoted by $N_a N_a$, where $N_a{:}X_a \rightarrow YN_a{:}X_a \rightarrow Y$. The predicted class label of the training test datasets generated by the output of the created networks is denoted by "Y."

6) By learning a standard description that combines relevant ideas from both modalities, the generated multimodal network "M" can then mention the most discriminating patterns in the streaming data.

This research describes FND as a supervised classification problem to predict the news' label based on the news' contents and its responses (i.e.,) f: $((\mathrm{PR})_i^T, (\mathrm{PR})_i^I, (\mathrm{PR})_i^S)_i^m \rightarrow Y$.

B. Text Embedding Using HCAN

For text embedding, this work implemented the HCAN model based on self-attention that captures linguistic relations over prolonged classifications like RNNs while still being fast to train, like CNNs. Fig. 2 depicts the complete system of the HCAN.

HCAN employs a hierarchical model that divides resources into sentences in the first place. A sentence embedding expressing the substance of that sentence is formed by the word hierarchy after reading in word embeddings from the given text, and the sentence hierarchy generates a document embedding that is the content of the complete text by reading in the sentence embeddings produced by the lower ranking. Later processing makes use of this document embedding.

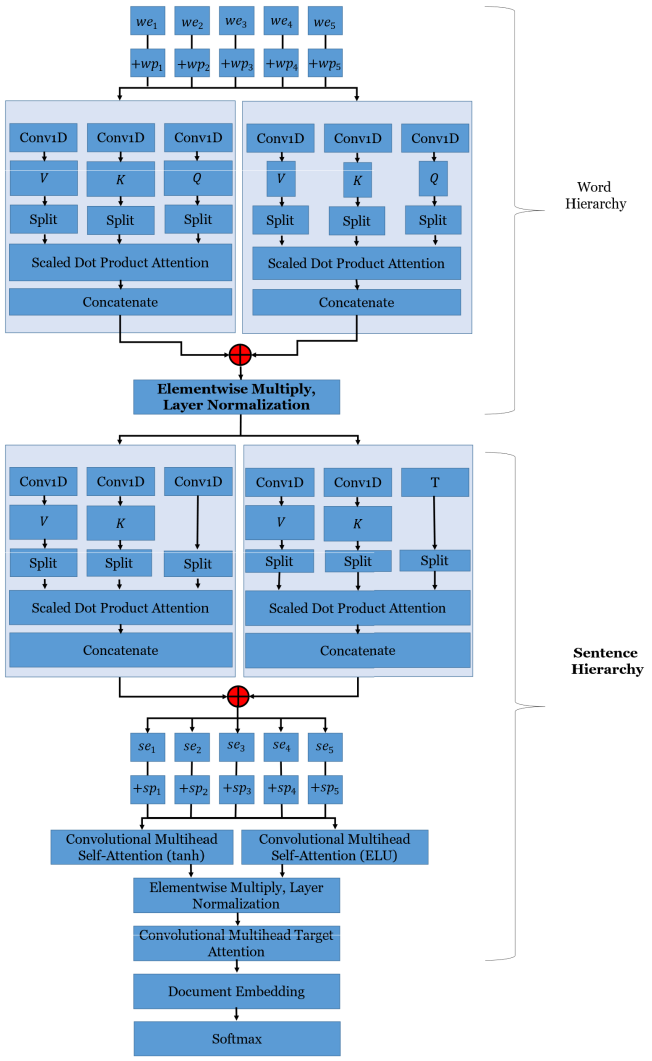

Fig. 2. HCAN architecture.

C. Joint Representation of TSC

Textual and visual/auditory components are typically the foundation of traditional multimodal methods. However, the social context attributes are used in literature for effective rumor detection (RD) for the RD task on microblogs. This research predicts that adding social context to the RD model is valuable in different methods. In order to create the first social context representation separately for post $P_S = [ps_1, ps_2, \ldots, ps_k]^T$ and reply $R_S = [rs_1, rs_2, \ldots, rs_k]^T$, the context created upon microblogs, like mentions, (#) topics, and retweets, in addition to specific textual semantic properties, like an emotional polarity, is being used. Social context features are measured in the dimension "k," and the "i" dimension's scalar value is represented by the letter S_i. The fully connected (FC) layer is used to transform a social context feature P_S and R_S into a representation $P_{S'}$ and $R_{S'}$ that owns a similar dimension as the word embedding vector, as shown in the following equations:

$$R_{S'} = \omega R_S \tag{1}$$

$$P_{S'} = \omega P_S \tag{2}$$

wherein ω is the weight inside the FC layer to the dimension transformation. During every time step, the Bi-GRU takes $P_{T_i S} = [P_{T_i}, P_{S'}]$ and $R_{T_i S} = [R_{T_i}, R_{S'}]$ as an input that seems to be a concatenation of the ith word embedding of P_{T_i} and R_{T_i} and the transformed social context feature of the post $P_{S'}$ and reply $R_{S'}$. Output neurons of Bi-GRU to every word get averaged for forming a joint representation of TSC $\mathrm{FR_{TS}}$ and $\mathrm{FR_{TS}}$. The $\mathrm{FR_{TS}}$ and $\mathrm{FR_{TS}}$ are paired together using BiGRU to generate $\mathrm{PR_{TS}}$.

D. Knowledge Extractor

This KE block has two layers, the TS-O encoder and OSN.

1) TS-O Encoder: As an example, $c \in \mathrm{PR_{TS}}$ contains the TSC representation of post "P" and replies "R" with "n" tokens $(r_1, r_2, \ldots, r_n)$ and an opinion instance "O" with "m" tokens $(o_1, o_2, \ldots, o_m)$. To an opinion "O" that is too related to a stance label "y." TS-opinion (TS-O)-encoder (TS-Oe) uses bidirectional encoder representations from transformers (BERT) to encode text-social representations "R" and opinion "O" concurrently. This work proposes processing R and O along with a single token sequence ([CLS], $r_1, r_2, \ldots, r_n$, [SEP], $o_1, o_2, \ldots, o_m$, [SEP]). Later, TS-Oe builds input embedding by also transforming that into

the combined embedding like BERT. Accurately, an input embedding being denoted like $X = (x_1, x_2, \ldots, x_l)$, wherein $x_i \in \mathbb{R}^k$ to be the embedding through summating on the token, location embeddings, and segment, and "l" is the determined dimension of a token order. Later, creating the input set X, TS-OE changes that into the joint text and social with opinion embedding. At the jth layer, a text-target set is represented as $H^{(j)} = (h_1, h_2, \ldots, h_l)$, wherein $h_i \in \mathbb{R}^k$ has similar dimensions by an equivalent input x_i, and $H^{(j)}$ is reached as in the following equations:

$$Y = H^{(J-1)} \tag{3}$$

$$H_i(Y) = \text{SoftMax}\left(\frac{\left(YW_{Q_i}\right)\left(YW_{K_i}\right)^i}{\sqrt{d/m)}}\right)\left(YW_{V_i}\right) \tag{4}$$

$$\text{Multi\_Head}(Y) = \text{Concat}(H_1, \ldots, H_m)W_O \tag{5}$$

$$Z = \text{LN}(Y + \text{ Multi\_Head } (Y)) \tag{6}$$

$$H^{(j)} = \text{LN}(Z + \text{MLP}(Z)) \tag{7}$$

wherein $W_{Q_i}, W_{K_i}, W_{V_i} \in \mathbb{R}^{d\times(d/m)}$, $W_o \in \mathbb{R}^{d\times d}$ are learnable variables, LN is layer normalization, and $H^{(0)} = X$. This complete model stacks "J" such layers and the end layer $H^{(J)}$. This article has been selected as our final TSC opinion embedding H'.

2) Opinion Convolution Network: Once the joint TS-O embedding is generated from the TS-O encoder, the OCN layer is applied to extract stance-indicative features. This layer is used to remove structural information from the input sequence. In addition, three CNNs are deployed: neutral opinion convolution (NOC), favor opinion convolution (FOC), and against opinion convolution (AOC). The corresponding stance-indicative features are predicted to be extracted from each OCN. Every OCN uses the combined TSC opinion embedding from the TS-O encoder layer as its input. Let $h_i \in \mathbb{R}^k$ represent a "k" dimensional output embedding corresponding to the ith token within the TS-O encoder's input sequence. The equation for the line of length "l" to be

$$H = h_1 \oplus h_2 \oplus \cdots \oplus h_l \tag{8}$$

wherein $H \in \mathbb{R}^{l\times k}$ and $\oplus$ are the concatenation operators. This work uses different filters to convolution it after that. A convolution operation uses a filter $w \in \mathbb{R}^{h\times k}$ to create a new feature by applying it to the "h" tokens window. A feature $c_i \in \mathbb{R}$, for instance, is produced from a window of tokens $h_{i:i+h-1}$ by the following equation:

$$c_i = f(w \odot h_{i:i+h-1} + b) \tag{9}$$

wherein the function "f" is nonlinear. For producing a feature map, this filter is functional to every potential window of tokens within the sequence in the following equation:

$$c = (c_1, c_2, \ldots, c_{n-h+1}). \tag{10}$$

This work later employs a max-over-time pooling operation upon a feature map by also taking the maximum value $\hat{c} = \max\{c\}$ to be the feature corresponding to such a specific filter. Lastly, the generation of elements from all filters gets cohesive to form a single high-level feature vector.

As a result, to find a feature vector using the expression $\hat{h} \in \mathbb{R}^p$, where "p" is the number of filters for each OCN. The following equation criteria are used to define the stance scores:

$$\begin{aligned} \alpha_0 &= W_0\hat{h}_0 + b_0 \\ \alpha_1 &= W_1\hat{h}_1 + b_1 \\ &\vdots \\ \alpha_n &= W_n\hat{h}_n + b_n \end{aligned} \tag{11}$$

wherein $W \in R^{1\times d}$ and $b \in R$ label a stance-specific linear transformation. The higher the value of α_i, the most probably "i" to be the valid stance label.

Let us assume $\alpha_0, \alpha_1, \ldots, \alpha_n$ to signify an opinion scores vector. This confidence scores "α" to be later normalized for probabilities applying the SoftMax operation in the following equation:

$$\begin{aligned} \alpha &= [\alpha_0, \alpha_1, \ldots, \alpha_n] \\ \hat{y} &= \text{SoftMax}(\alpha) \end{aligned} \tag{12}$$

wherein $\hat{y} \in R^n$ seems to be the vector of forecast probability to the "n" stances.

As a result of the global average pooling's (GAP) large pooling size, which is symbolized by the 7×7 and 8×8 layouts, it also significantly affects the transfer of the gradient. For example, consider the network's final loss $X(Y^a, Y^b)$, where (Y^a, Y^b) denotes the actual and predicted labels, respectively. The SoftMax function decides the Y^p value in the following equation:

$$Y^p = f\left(P_x Y^c + Q_s\right). \tag{13}$$

In (20), $f()$ stands for the SoftMax function, and P_x and Q_y represent the SoftMax layer weights. When the SoftMax layer's gradient is multiplied, this work obtains the SoftMax layer's gradient

$$\frac{\partial Z\left(Y^a, Y^b\right)}{\partial Y^c} = \frac{\partial Z\left(Y^a, Y^b\right)}{\partial Y^b} + \frac{\partial Y^a}{\partial f} + \frac{\partial f}{\partial Y^c}. \tag{14}$$

Like the often-used 77 pooling region, the GAP is typically huge. The losses transmitted from the classification layer are averaged by 49 to ensure that the backpropagation keeps the total gradients before and after pooling.

E. Visual Stream

Image content from news feeds is fed into the graphical subnetwork (bottom layer in Fig. 1), which creates the visual neurons that appear in the images. To extract visual features, this work used inception-ResNet-v2. Convolutional neural network (CNN) inception-ResNet-v2 produced innovative performance on the image classification benchmark. Adding the bypass connection seen in ResNet, inception-ResNet-v2 modifies the prior inception V3 model. The model has 54.36 M parameters to provide the 512-neuron visual representation $P_I = [v_1, v_2, \ldots, v_{512}]^T$ of all images. This work trains the deep CNNs outlined above with pretrained on ImageNet weights to learn visual neurons to recognize specific patterns

for RD. To reduce the spatial dimensions of a 3-D tensor, the image stream extracts visual elements sent to a global average pooling layer.

Additionally, it conducts a more severe form of dimensionality reduction. The FC final layer receives the global middle pooling layer for the three deep CNNs' final layers, which perform classification with a SoftMax layer. However, only the limits of the last two FC layers are updated for more effective training during the cooperative training with the Bi-GRU subnetwork in the following equation:

$$P_I = \omega_{vf_1}\left(\omega_{vf_2} P_{I_p}\right) \tag{15}$$

wherein P_{I_p} to be the visual feature acquired through a pretrained inception-ResNet-v2, ω_{vf_1} to be weights inside a first FC layer, and ω_{vf_2} to be weights inside the second FC layers along with SoftMax.

Visual neurons within such a process were not restricted to getting specific network concepts. Hopefully, provided the training data are labeled as rumors and nonrumors, they can successfully record the important semantic images for RD during the combined training of the complete network. One of the main challenges for directly integrating visual and common social-textual words inside this model is that one representation will dominate the other and lead to biased performance toward this modality. This work must jointly reach many modalities' alignments to optimize the benefits of multimodal features.

F. Attention to Visual Representation

In rumor tweets, this article believes that images would correlate in specific correlations with the text and the social context. This work proposes a neuron-level attention mechanism about visual features under a direction for a neuron that concurrently is TSC in order to describe these interactions. Recent language-vision tasks have used attention mechanisms to map textual and graphical semantic concepts. In the context of RD, this research considers that text content words are connected to certain semantic concepts inside the image. This work's goal would be to discover these types of connections automatically. Remarkably, more weight is given to the visual neurons whose semantic meanings are like words.

The innovative inception-ResNet-v2 network extracts deep spatial features from videos using residual connections and inception layout. Stem, inception ResNet, and reduction layers are included in the deployed inception-ResNet-v2 model. An average-pooling layer, 1000-channel layer, and FC layer follow. Prior convolution functions conducted before entering the inception blocks are included in the stem. The convolution operation and residual connections are part of the inception ResNet layers, while the reduction layers implement the required corrections to the grid's sizes. The model is insensitive to illumination and translation because the pooling layers hold essential features but reduce the feature map's dimensionality. Convolutional layers extract spatial features. In front of the convolutional layers are the batch normalization layer and rectified linear unit (ReLU), a nonlinearity function that reduces training time. Inception-ResNet-v2 uses $299 \times 299 \times 3$ RGB frames to detect each sentence's spatial information. Spatial information is essential because many behaviors are actively linked to objects. The inception-ResNet-v2-Net accepts each frame and finds the local features in it. The global features of a sentence are formed from the aggregate of these local features.

The recommended visual-neuron attention mechanism weighs the contribution of different neurons to different words. This investigation uses an output hidden state h_m at every time step, thus helping to conduct this goal. For an attention vector $A_m \in \mathbb{R}^{512}$ that have a similar dimension to the visual neurons P_V, h_m is to be associated with the FC layer using the nonlinearity ReLU function and the FC layer along with the SoftMax function in the following equation:

$$A_m = \omega_{af_2}\psi\left(\omega_{af_1} h_m\right) \tag{16}$$

wherein h_m being a hidden state on the "m" time step, ω_{af_1} and ω_{af_2} stand as weights inside the two FC layers, and "ψ" to be the activation. A correlation between the mth word within the text and images is computed as given in the following equation:

$$a_m = \sum_{i=1}^{512} A_m(i) v_i \tag{17}$$

wherein $A_m(i)$ seems to be the attention value to the i th visual neuron. Based on this attention mechanism, the attention vector "A_m" formed by BiGRU in the combined feature learning of TSC could investigate which visual neurons are predicted to be more focused.

In the key layers and data features, the dynamic fusion of BERT and BiGRU proposes an intelligent classification model for metro onboard device failures. In the first 12 layers of the upstream model, BERT's transformer has been reassembled using the top five layers as the new layering scheme. Before being sent to BiGRU, the input features are fused, and their size is decreased. The forward and backward GRU sequences are combined to create a feature representation with more semantic data, and the bidirectional gating unit BiGRU ensures the sequence of feature vectors. In the last step, the generated feature vector is sent to the full connection (FC) layer, where it is connected to the SoftMax function, which gives the classification result. A final visual representation seems to be a set of values representing the affinity of each word: $A(P^{I'}) = [a_1, a_2, \ldots, a_n]$ (n = words count of sample text) (n = word count of assumed text). In contrast to object-level semantics on conventional optical recognition tasks, it is distinguished that high-level visual semantics might be particularly challenging to identify in RD tasks. Inside the attention model, no mechanism expressly ensures learning of this mapping correlation. However, training using this attention mechanism can reveal implicit relationships and enhance feature alignment. In the experiment section, its practical effects are assessed.

The deep multiview semantic document representation model is adopted to decide the semantic dependency of the matched text's semantic vector. The task of defining the semantic dependency of a text sequence is said in the following

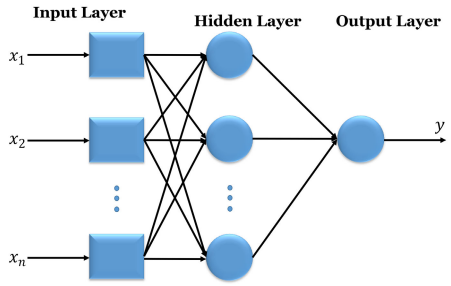

Fig. 3. MLP-based binary classifier.

TABLE I
MLP PARAMETERS

Layer	Neurons	Activation Function
Input	450	–
Hidden	450	ReLU (EUQ (2))
Output	1	Sigmoid (EQU (3))

equation for the case of a text sequence $X = \{x_0, x_1, \ldots, x_n\}$:

$$X(t) = \mathrm{SV} \int \frac{x(t)}{(1-t')'} \tag{18}$$

(24) where "t" denotes a segment of semantically similar text, SV denotes the text's semantic vector, and $[t']$ denotes the identical text with added interference data.

After the semantic dependency capture has been completed, feature extract and feature select high-dimensional abstract semantic data are extracted to achieve granularity at the word and sentence level. The structure of a multigranularity semantic FE model requires the implementation of a DNN.

Lastly, the textual feature representation $\boldsymbol{A}(\boldsymbol{O}_{\tau}^{\mathrm{TS}})$ and the visual feature representation $\boldsymbol{A}(\boldsymbol{P}^{I'})$ were combined to create the multimodal feature representation in the following equation:

$$A(O)_i^{\mathrm{TSI}} = \boldsymbol{A}(\boldsymbol{O}_{\tau}^{\mathrm{TS}}) \oplus \boldsymbol{A}(\boldsymbol{P}^{I'}). \tag{19}$$

G. Fake News Detector

A binary classification structure was selected using MLPs to train an ML model to derive the classes. The input and hidden layers of the MLPs' model have a single hidden layer with $n = 450$ neurons (see Fig. 3). ReLU and sigmoid, respectively, are the activation functions for the hidden and output layers (see Table I), and the network is optimized using the Adam optimizer applying binary cross entropy (19) and learning rate $l = 0.001$ in the following equations:

$$\mathrm{ReLU}(x) = \max(0, x) \tag{20}$$

$$\mathrm{sigmoid}(x) = \frac{1}{1+e^{-x}}. \tag{21}$$

An MLP uses the multimodal features representation (22) $\boldsymbol{A}(\boldsymbol{O}_{\tau}^{\mathrm{TS}})$ and $\boldsymbol{A}(\boldsymbol{P}^{I'})$ as an input to generate the predictor $\mathbb{P}(y_i)$

$$\mathbb{P}(y_i) = \mathrm{MLP}\boldsymbol{A}(\boldsymbol{O}_{\tau}^{\mathrm{TS}}) \oplus \boldsymbol{A}(\boldsymbol{P}^{I'}). \tag{22}$$

For determining the classification loss, binary cross-entropy (23) is used, wherein "n" stands for the number of news, and "Y_i" is a ground-truth label for the ith news

$$H_p(q) = -\frac{1}{n}\sum_{i=1}^{n} y_i \cdot \log(\mathbb{P}(y_i)) + (1-y_i)\cdot \log(1-\mathbb{P}(y_i)). \tag{23}$$

IV. EXPERIMENTAL SETUP

A. Dataset for Experiment

In order to assess the performance, this work creates many competing baselines. This research examines many state-of-the-art language-vision models and the conventional feature-based single-modal techniques, which involve simple and direct fusion techniques for RD. In order to evaluate this model for FND, different study experiments are accepted using real-world datasets, including Weibo, Fakeddit, and PHEME [21]. Including submissions and comments, a Fakeddit dataset is to be a unique FN dataset. From an original Fakeddit dataset, this work chose seven multimodal subreddits for this article and left out each single-modal subreddits. After an original Fakeddit splitting method, submissions of 13 396 were selected as the training dataset, with 9060 identified as TN and 4336 as FN.

One thousand and six of these 1413 submissions in the validation set are TN, while 447 are FN, totaling 1413 in the validation set. There are 1453 submissions in the test set, of which 939 are determined as actual news and 474 as FN. Twitter tweets containing five breaking news, each tweet responding to them, is included in the PHEME dataset. There are 5802 annotated tweets, 3830 TN, and 1972 FN, with a 5802 mark. Weibo Dataset, from where this article gets the dataset with the objective ground-truth labels, is used to evaluate this proposed approach fairly. This article crawls explicitly every verified false rumor post on Weibo's official rumor-debunking system from May 2012 to January 2016 to determine its authenticity. The technique promotes the reporting of suspicious tweets on Weibo by regular users. Afterward, a group of reputable users would review the cases and determine whether they were real. This method is a reliable resource for compiling rumor tweets for literature. Verified tweets from the legitimate Xinhua News Agency in China determined the nonrumor tweets.

B. Experimental Settings

The distributed representation of words is what this article used for the textual feature. An HCAN model is pretrained for these three datasets using the entire dataset within an unsupervised manner using default parameter settings following conventional preprocession text. This research obtains an embedding feature in 32 dimensions for each word inside the datasets. Because vocabulary size is prominent in a one-hot representation method, a short text may have inadequate word features, which is why word embedding representation is better. This article considers the most socially critical literary qualities as the social context feature. This work obtains 16 social features from a Weibo dataset and 18 from a PHEME dataset from available data and 21 from Fakeddit. A few texts'

semantic features are included because of their significance in the RD task, like expressive division and the number of first-order pronouns.

This research work uses the output of the inception-ResNet's second-to-last layer, which was trained on the ImageNet dataset, to be the visual feature. It has a 4096-feature size. This work proposed a graphical subnetwork that might remain improved along with more data; however, this article allows for future studies since there is not currently any image set directly relevant to this task. This research uses Bi-GRU to conduct the RNN's combined learning for TSC. This work uses the nonlinear activation function and sets the hidden layer dimension to 16. It also deploys a batch size of 128 instances to train the complete network. Every NN model has trained about 100 epochs within the experiments, followed by the early stopping point for reporting the results.

C. Baselines

The baselines described in the following are used in the experiments on those three datasets.

1) **Decision Tree Classifier [22]:** DT is a classifier for modeling Twitter information credibility using manual features.
2) **Gated Recurrent Unit [23]:** Efficient for the early RD, GRU models a microblog like the variable-length time series using a multilayer generic GRU network.
3) **CNN [24]:** After modeling the relevant postings as fixed-length sequences, CNN uses a convolution network to learn rumor representations.
4) **Event Adversarial Neural Networks [25]:** By concatenating the post's graphical and textual features and employing the adversarial method for obliterating event-specific features through postrepresentation, this method is a post-level FND model that seeks to identify whether the post in question is authentic.
5) **Multimodal Knowledge-Aware Event Memory Network [26]:** The method suggested multimodal and knowledge data to learn the postrepresentation. EMN does this to extract the event-invariant features for RD.

V. Performance Analysis

A. Evaluation Metrics

By proving four evaluation criteria, true positive (TP), true negative (TN), false positive (FP), and false negative (FN)—depending upon the relationship among predicted and TN classifications—this research work assessed the model using four performance criteria based on such parameters.

1) Accuracy: The proportion of correct predictions to all predictions is represented by the ratio in the following equation:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{24}$$

2) Precision: A ratio of correct optimistic predictions for all positive predictions is used to compute a positive predicted value in the following equation:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{25}$$

TABLE II
FAKEFIND VERSUS BASELINE MODELS FOR THREE DATASETS

Dataset	Methods	Accuracy	Fake			Real		
			Precision	Recall	F1	Precision	Recall	F1
WEIBO	DTC	0.6312	0.6329	0.6301	0.6309	0.665	0.700	0.682
	GRU	0.7927	0.8139	0.7927	0.7891	0.814	0.826	0.820
	CNN	0.7112	0.713	0.7112	0.711	0.7206	0.7213	0.7209
	EANN	0.7212	0.7353	0.7228	0.7160	0.772	0.803	0.787
	MKEMN	0.8863	0.8645	0.8782	0.8392	0.8707	0.8803	0.8612
	Proposed	0.9306	0.9077	0.9221	0.8812	0.9142	0.9243	0.9043
FAKEDDIT	DTC	0.548	0.581	0.536	0.558	0.577	0.523	0.548
	GRU	0.736	0.729	0.739	0.734	0.745	0.714	0.729
	CNN	0.651	0.655	0.631	0.643	0.663	0.644	0.653
	EANN	0.7177	0.7382	0.7179	0.7104	0.7213	0.7436	0.7224
	MKEMN	0.816	0.809	0.819	0.814	0.816	0.801	0.809
	Proposed	0.8486	0.8414	0.8518	0.8466	0.8486	0.8330	0.8414
PHEME	DTC	0.6399	0.6391	0.6211	0.6395	0.6519	0.6335	0.6523
	GRU	0.8374	0.8382	0.8374	0.8312	0.8550	0.8541	0.8478
	CNN	0.7007	0.7413	0.7074	0.6896	0.7561	0.7215	0.7034
	EANN	0.7177	0.7382	0.7179	0.7104	0.7530	0.7323	0.7246
	MKEMN	0.8580	0.8586	0.8589	0.8588	0.8758	0.8761	0.8760
	Proposed	0.9009	0.9015	0.9018	0.9017	0.9196	0.9199	0.9198

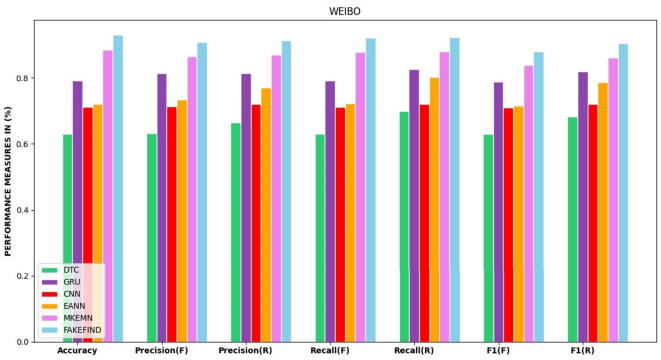

Fig. 4. Fakefind against baselines for the WEIBO.

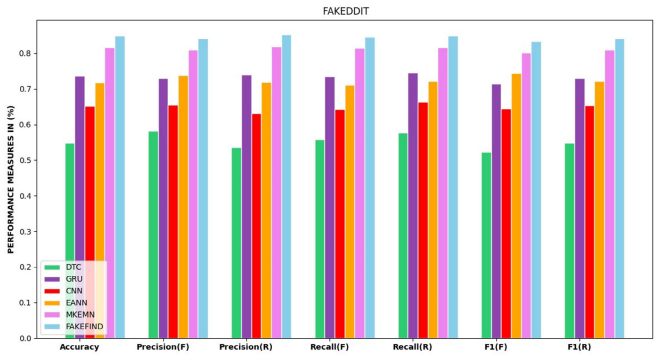

Fig. 5. Fakefind against baselines for the FAKEDDIT.

3) Recall: A ratio of correct optimistic predictions for all correctly predicted results, R being the model's sensitivity, is given in the following equation:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{26}$$

4) F1-score: The F1-score is the harmonic mean of the precision and recall determining the model's testing accuracy in the following equation:

$$\text{F1score} = \frac{2}{\text{Recall}^{-1} + \text{Precision}^{-1}}. \tag{27}$$

B. Quantitative Evaluation

Hence, experimental findings of this recommended model and other baseline techniques are shown in Table II and Figs. 4–6. This research article observes the following results from Table II.

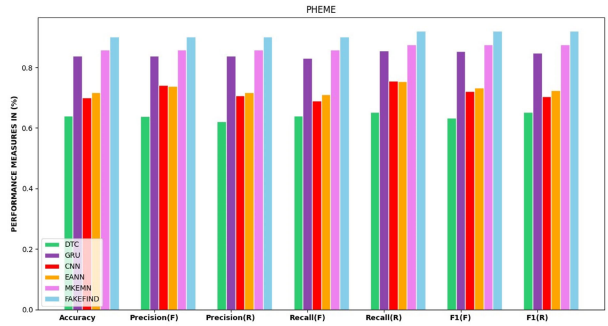

Fig. 6. Fakefind against baselines for the PHEME.

The decision tree classifier (DTC) model performs the worst out of all the approaches, proposing that the manual features may not be sufficiently solid for FND. CNN is a supervised technique that uses NN training to capture local features between different words. Their performance is only superior to DTC on three datasets due to CNN ignoring the long-range semantic relationships between terms and the insufficiency of local features for postassessment. Event adversarial neural networks (EANNs) outperforms CNN. It is because ANN uses VGG-19 to extract graphical elements and text CNN to extract textual features, providing complimentary data to enhance FND. Due to RNN's natural volume to deal with variable-length sequences of postings, GRU outperforms CNN upon three datasets, particularly a PHEME dataset, while CNN requires more data to reach a decision. All baseline models perform significantly worse than the multimodal knowledge-aware event memory network (MKEMN) model, except for the proposed model. By capturing and storing event-invariant attributes in external memory, the event memory network can improve efficiency by retrieving the stored data when the system meets newly emerging events. Comparing all the baselines, the performance analysis of the proposed model is the best. Two characteristics are responsible for its superiority: 1) the proposed model combines textual, visual information, and social context concepts to improve the post text's semantic information and 2) the model involves an opinion-based knowledge network that extracts the clear stance of the posts. The proposed model performs significantly better on the three datasets than any benchmark model. This model successfully learned the standard features from multiple modalities, as shown by its complete accuracy of 78.8% on the Weibo set and 68.2% on the Twitter set.

C. Ablation Study

This article created several comparison baselines that are condensed variations to the proposed model by leaving specific components for further study of the effects of every modality inside the proposed model, and the output is exposed in Tables III–V and Figs. 7–9.

1) **Fakefind Without Social Context (F Without S):** The social context feature is removed from the proposed model and has relied upon KE text and image.
2) **Fakefind Without Image (F Without I):** The visual subnetwork is removed. The binary classifier receives the KE by combining TSC features bought through Bi-GRU.

TABLE III
MODIFIED FAKEFIND PERFORMANCE FOR WEIBO

Method	WEIBO						
	Accuracy	Fake			Real		
		Precision	Recall	F1	Precision	Recall	F1
Fakefind	0.9306	0.9077	0.9221	0.8812	0.9142	0.9243	0.9043
F w/o S	0.902697	0.880493	0.894447	0.854725	0.886808	0.896586	0.877132
F w/o I	0.865472	0.844184	0.857562	0.819479	0.868523	0.878099	0.859047
F w/o S+I	0.828247	0.807875	0.820678	0.784232	0.850239	0.859613	0.840962
F w/o K	0.915725	0.893201	0.907356	0.867061	0.899607	0.909526	0.889792
F w/o S+K	0.888253	0.866405	0.880136	0.84105	0.872619	0.88224	0.863098
F w/o K+I	0.836046	0.815482	0.828405	0.791617	0.838993	0.848244	0.829839
F w/o S+I+K	0.816652	0.796565	0.809188	0.773253	0.838335	0.847578	0.829188

TABLE IV
MODIFIED FAKEFIND PERFORMANCE FOR FAKEDDIT

Method	FAKEDDIT						
	Accuracy	Fake			Real		
		Precision	Recall	F1	Precision	Recall	F1
Fakefind	0.8486	0.8414	0.8518	0.8466	0.8486	0.8330	0.8414
F w/o S	0.8231	0.816119	0.826207	0.821163	0.823181	0.808049	0.816119
F w/o I	0.780749	0.774051	0.783619	0.778835	0.765473	0.751402	0.758907
F w/o S+I	0.644966	0.639434	0.647338	0.643386	0.775657	0.761399	0.769003
F w/o K	0.835062	0.827898	0.838132	0.833015	0.835062	0.819711	0.827898
F w/o S+K	0.81001	0.803061	0.812988	0.808025	0.81001	0.79512	0.803061
F w/o K+I	0.754203	0.747733	0.756976	0.752355	0.739447	0.725854	0.733104
F w/o S+I+K	0.635937	0.630482	0.638275	0.634378	0.764798	0.750739	0.758237

TABLE V
MODIFIED FAKEFIND PERFORMANCE FOR PHEME

Method	PHEME						
	Accuracy	Fake			Real		
		Precision	Recall	F1	Precision	Recall	F1
Fakefind	0.9009	0.9015	0.9018	0.9017	0.9196	0.9199	0.9198
F w/o S	0.873873	0.874484	0.87479	0.874688	0.891974	0.892285	0.892182
F w/o I	0.857657	0.858257	0.858556	0.858456	0.85703	0.85733	0.85723
F w/o S+I	0.809008	0.809574	0.809857	0.809763	0.85887	0.85917	0.85907
F w/o K	0.886486	0.887106	0.887415	0.887312	0.904848	0.905164	0.905058
F w/o S+K	0.859891	0.860492	0.860793	0.860693	0.877702	0.878009	0.877907
F w/o K+I	0.828496	0.829076	0.829366	0.829269	0.827891	0.828181	0.828084
F w/o S+I+K	0.797682	0.79824	0.798519	0.798426	0.846845	0.847141	0.847043

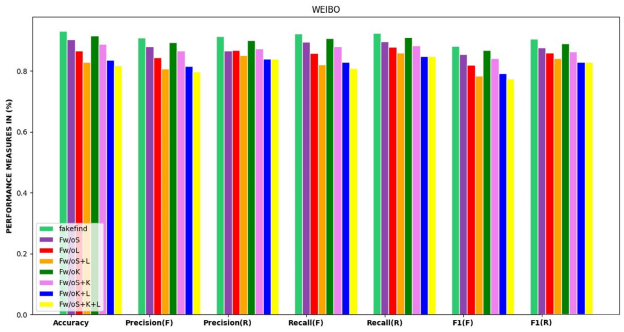

Fig. 7. Modified Fakefind versus WEIBO.

3) **Fakefind Without Image + Social (F Without S + I):** The model is trained without image and social context. The knowledge is used as an input to train the model.
4) **Fakefind Without KE (F Without K):** The proposed model is trained without a KE.
5) **Fakefind Without social Context + KE(F Without S + K):** The proposed model is trained without KE, and social context takes only image and text input.
6) **Fakefind Without Image + KE(F Without K + I):** The proposed model is trained without a KE and image.

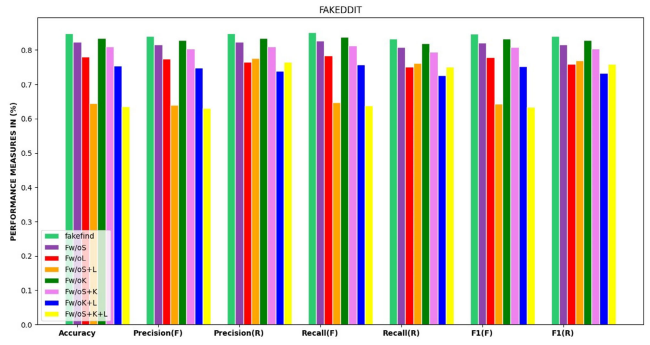

Fig. 8. Modified Fakefind versus FAKEDDIT.

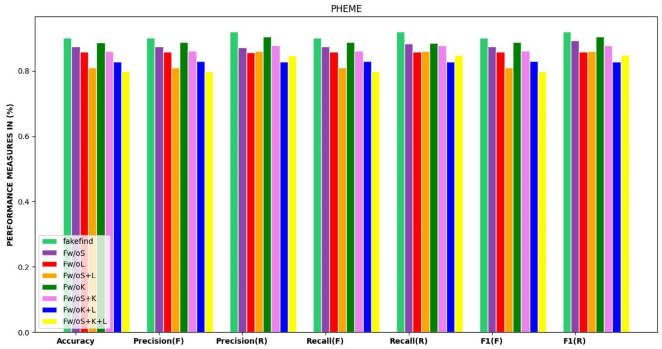

Fig. 9. Modified Fakefind versus PHEME.

7) **Fakefind Without Image + Social +KE(F Without S + I + K):** The model is trained without image, social context, and knowledge. The text network is used as an input to train the model, and the averaged outputs are fed into a binary classifier.

Tables III–V show the performance ratings of the modified baselines on the three datasets. This article refers that all the components, including social context features, image features, and knowledge, are critical for our Fakefind to achieve the best RD performance. If this work removes one or more, performance will suffer to some extent. Image features are essential since accuracy on all datasets decreases by 5%–7% without them. Knowledge and characteristics of the social context are also crucial. Without social context and image information, the accuracy is 10%–20% worse across all datasets than Fakefind. Whereas, without knowledge features alone, the accuracy versus all three datasets decreases by 2%–3%. Without social and knowledge traits, accuracy can fall by up to 5%, while omitting knowledge and visual elements cause accuracy to fall by about 15%. The model's accuracy is reduced to a greater extent of 25% when image, social, and knowledge components are removed.

VI. Conclusion

This article proposes Fakefind, a hybrid convolution neural networks (CNN) + RNNs that uses a stance generated from a CNN-based KE to merge characteristics from text, image, and social context for spotting rumors. The related text and social environment for a given feed and response is fused individually and combined with bidirectional recurrent neural network (Bi-GRU). The combined representation is then input into a KE, a convolution-based network that uses a TS-O encoder, and an OCN layer to extract stance-indicative features. These are then fused with visuals developed by inception-ResNet-v2 networks, which have already been trained. The Bi-GRU output at each time step is used as neuron-level attention to coordinate visual aspects throughout the fusion. The proposed method's performance is compared to multimodal fusion methods based on neural networks and different feature-based frameworks in RD based on multimedia content, according to multiple research works using WEIBO, FAKEDDIT, and PHEME datasets.

References

[1] A. D. Holan, *2016 Lie of the Year: Fake News*. Washington, DC, USA: Politifact, 2016.

[2] D. M. J. Lazer et al., "The science of fake news," *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018.

[3] S. Kogan, T. J. Moskowitz, and M. Niessner, "Fake news: Evidence from financial markets," *SSRN Electron. J.*, 2018, doi: 10.2139/ssrn.3237763.

[4] A. Robb, "Anatomy of a fake news scandal," *Rolling Stone*, vol. 1301, pp. 28–33, Jan. 2017.

[5] B. Edkins. *Americans Believe They Can Detect Fake News, Studies Show They Can't*. Accessed: Dec. 21, 2022. [Online]. Available: www.forbes.com/sites/breedkins/2016/12/20/Americans-believe-they-can-detect-fake-news-studies-show-they-cant/

[6] P. Kar, Z. Xue, S. P. Ardakani, and C. F. Kwong, "Are fake images bothering you on social network? Let us detect them using recurrent neural network," *IEEE Trans. Computat. Social Syst.*, vol. 10, no. 2, pp. 783–794, Apr. 2023.

[7] Y. Chen, Y. Lv, X. Wang, L. Li, and F. Wang, "Detecting traffic information from social media texts with deep learning approaches," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 8, pp. 3049–3058, Aug. 2019, doi: 10.1109/TITS.2018.2871269.

[8] S. Lin, F. Zhou, G. Zheng, K. Li, and M. Zhou, "Multi-facet information processing algorithms for news video based on event combing," in *Proc. 8th Int. Conf. Digit. Home (ICDH)*, Dalian, China, Sep. 2020, pp. 266–270, doi: 10.1109/ICDH51081.2020.00052.

[9] K. A. Qureshi, R. A. S. Malick, M. Sabih, and H. Cherifi, "Complex network and source inspired COVID-19 fake news classification on Twitter," *IEEE Access*, vol. 9, pp. 139636–139656, 2021.

[10] K. Soga, S. Yoshida, and M. Muneyasu, "Propagation-based fake news detection using a combination of different content features," in *Proc. IEEE 11th Global Conf. Consum. Electron. (GCCE)*, Oct. 2022, pp. 400–401, doi: 10.1109/GCCE56475.2022.10014073.

[11] J. Golbeck et al., "Fake news vs satire: A dataset and analysis," in *Proc. 10th ACM Conf. Web Sci.*, New York, NY, USA, 2018, pp. 17–21.

[12] N. J. Conroy, V. L. Rubin, and Y. Chen, "Automatic deception detection: Methods for finding fake news," *Proc. Assoc. Inf. Sci. Technol.*, vol. 52, no. 1, pp. 1–4, 2015.

[13] H. Ahmed et al., "Detection of online fake news using *n*-gram analysis and machine learning techniques," in *Proc. Int. Conf. Intell., Secure, Dependable Syst. Distrib. Cloud Environ.*, 2017, pp. 127–138.

[14] H. Hichem, M. Elkamel, M. Rafik, M. T. Mesaaoud, and C. Ouahiba, "A new binary grasshopper optimization algorithm for feature selection problem," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 34, no. 2, pp. 316–328, Feb. 2022, doi: 10.1016/j.jksuci.2019.11.007.

[15] Q. Li and Q. Zhang, "A unified model for financial event classification, detection and summarization," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, 2021, pp. 4668–4674.

[16] M. L. Newman, J. W. Pennebaker, D. S. Berry, and J. M. Richards, "Lying words: Predicting deception from linguistic styles," *Personality Social Psycho. Bull.*, vol. 29, no. 5, pp. 665–675, May 2003.

[17] *Google at ICLR 2023*. Accessed: May 1, 2023. [Online]. Available: https://ai.googleblog.com/2016/09/a-neural-network-for-machine.html

[18] M. R. Morris, S. Counts, A. Roseway, A. Hoff, and J. Schwarz, "Tweeting is believing? Understanding microblog credibility perceptions," in *Proc. ACM Conf. Comput. Supported Cooperat. Work*, Feb. 2012, pp. 441–450.

[19] S. Hangloo and B. Arora, "Content-based fake news detection using deep learning techniques: Analysis, challenges and possible solutions," in *Proc. 5th Int. Conf. Comput. Intell. Commun. Technol. (CCICT)*, Sonepat, India, Jul. 2022, pp. 411–417.

[20] A. Choudhry, I. Khatri, M. Jain, and D. K. Vishwakarma, "An emotion-aware multitask approach to fake news and rumor detection using transfer learning," *IEEE Trans. Computat. Social Syst.*, early access, Dec. 19, 2022, doi: 10.1109/TCSS.2022.3228312.

[21] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata, and R. Procter, "Detection and resolution of rumours in social media: A survey," *ACM Comput. Surv.*, vol. 51, no. 2, pp. 1–36, 2018.
[22] E. Canhasi, R. Shijaku, and E. Berisha, "Albanian fake news detection," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 21, no. 5, pp. 1–24, Sep. 2022, doi: 10.1145/3487288.
[23] Y.-W. Lu, C.-Y. Hsu, and K.-C. Huang, "An autoencoder gated recurrent unit for remaining useful life prediction," *Processes*, vol. 8, no. 9, p. 1155, Sep. 2020.
[24] S. Feng, R. Banerjee, and Y. Choi, "Syntactic stylometry for deception detection," in *Proc. 50th Annu. Meeting Assoc. Comput. Linguistics*, vol. 2, 2012, pp. 171–175.
[25] Y. Wang et al., "EANN: Event adversarial neural networks for multi-modal fake news detection," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, New York, NY, USA, Jul. 2018, pp. 849–857, doi: 10.1145/3219819.3219903.
[26] H. Zhang, Q. Fang, S. Qian, and C. Xu, "Multi-modal knowledge-aware event memory network for social media rumor detection," in *Proc. 27th ACM Int. Conf. Multimedia*, New York, NY, USA, Oct. 2019, pp. 1942–1951.

Sudhakar Sengan (Member, IEEE) received the M.E. degree from the Faculty of Computer Science and Engineering, Anna University, Chennai, Tamil Nadu, India, in 2007, and the Ph.D. degree in information and communication engineering from Anna University, in 2014.

He is currently working as a Professor and the Director of international relations with the Department of Computer Science and Engineering, PSN College of Engineering and Technology, Tirunelveli, Tamil Nadu, India. He is the Recognized Research Supervisor at Anna University, under the faculty of Information and Communication Engineering (ICE). He has 20 years of experience in teaching/research/industry. He has contributed more than 200 articles and chapters for many high-quality Scopus and Science Citation Index/Science Citation Index Expanded (SCI/SCIE)-indexed journals and books. He has filled 20 Indian and three international patents in various fields of interest. His research interests include security, MANET, the Internet of Things (IoT), cloud computing, and machine learning (ML)/deep learning (DL).

Dr. Sengan is a member of professional bodies like MISTE, Member in Institute of Electrical and Electronics Engineers (MIEEE), Member in International Association of Engineers (MIAENG), Member in International Association of Computer Science and Information Technology (MIACSIT), Member in Computer Society of India (MICSI), Member in Institution of Engineers (MIE), and Member of International Economics Development and Research Center (MIEDRC). He received the Award of Honorary Doctorate (Doctor of Letters-D.Litt.) from International Economics University, SAARC Countries in Education and Students Empowerment, in April 2017.

Subramaniyaswamy Vairavasundaram received the Ph.D. degree from Anna University, Chennai, Tamil Nadu, India, in 2013.

He continued the extension work with the Department of Science and Technology support as a Young Scientist Award Holder. He is currently working as a Professor with the School of Computing, SASTRA Deemed University, Thanjavur, India. He has 18 years of experience in an Academician and a Researcher. He has contributed more than 175 articles and chapters for many high-quality technology journals and books that are being edited by internationally acclaimed professors and professionals. He has received funded and consultancy projects from DST-SERB, ICSSR–IMPRESS, MHRD, MHI, TVS MOTORS, and SERB–MATRICS.

Dr. Vairavasundaram is on the reviewer board of several international journals and has been a program committee member for several international/national conferences and workshops. He serves as a Guest Editor for various special issues of reputed international journals. He is serving as a Research Supervisor and successfully guided five research scholars, and he is also a visiting expert to various universities in India and Abroad. His interests include recommender systems, the Internet of Things, artificial intelligence, machine learning, and big data analytics.

Logesh Ravi received the B.Tech. degree in computer science and engineering, the M.Tech. degree in networking from Pondicherry University, Puducherry, India, in 2012 and 2014, respectively, and the Ph.D. degree in artificial intelligence and recommender systems from the SASTRA Deemed University, Thanjavur, India, in 2019.

He is currently a Senior Assistant Professor with Centre for Advanced Data Science, Vellore Institute of Technology, Chennai, Tamil Nadu, India. His research was funded and sponsored by Science and Engineering Research Board, Department of Science and Technology, New Delhi, India. He serves as an Academic Coordinator and is an Active Participant in Academic Administration and Research in Computing Sciences. He has published more than 100 papers in reputable international journals and conferences. His research interests include artificial intelligence, recommender systems, big data, machine learning, social computing, information retrieval, and human–computer interaction.

Dr. Ravi also serves as a Guest Editor for many reputed international journals.

Ahmad Qasim Mohammad AlHamad received the B.Sc. degree in computer engineering from the Jordan University of Science and Technology, Ar-Ramtha, Jordan, in 2000, the M.Sc. degree from Queen Margaret University, Musselburgh, Edinburgh, U.K., in 2002, and the Ph.D. degree in information systems from Coventry University, Coventry, U.K., in 2011.

He has 20 years of teaching experience, including being the Dean of the IT College, University of Fujairah, Fujairah, United Arab Emirates, for three years. He is currently working as a Visiting Professor with the University of Sharjah, Sharjah, United Arab Emirates. His research interests include e-commerce, e-learning, machine learning, and metaverse.

Hamzah Ali Alkhazaleh (Member, IEEE) received the B.Sc. degree in management information systems from the Al Albyte University, Mafraq, Jordan, in 2007, the M.Sc. degree in information technology from the Northern University of Malaysia, Changlun, Malaysia, in 2010, and the Ph.D. degree in computer science (AI) from the National University of Malaysia, in 2016.

He is an Assistant Professor with the College of Engineering and Information Technology, the University of Dubai, Dubai, United Arab Emirates. He has more than ten years of teaching and research experience at the university level. His research interests include AI and OR, particularly computational optimization algorithms, which involve different real-world applications for single and multiaim continuous and combinatorial optimization problems, such as timetabling, scheduling, routing, nurse rostering, dynamic optimization, data mining problems, and intrusion detection. He has published many articles in international journals and peer-reviewed international conferences. He has served as a program committee for various international conferences and reviewer for journals.

Meshal Alharbi received the M.Sc. degree in computer science from Wayne State University, Detroit, MI, USA, in 2014, and the Ph.D. degree in computer science from Durham University, Durham, U.K., in 2020.

He has ten years of experience in teaching/research/industry. He is an Assistant Professor of artificial intelligence (AI) with the Department of Computer Science, Prince Sattam Bin Abdulaziz University, Al-Kharj, Saudi Arabia. His research interests include AI and algorithms, agent-based modeling and simulation applications, disaster/emergency management and resilience, optimization applications, and machine learning.