2021 2nd Global Conference for Advancement in Technology (GCAT)
Bangalore, India. Oct 1-3, 2021

Microblogs Deception Detection using BERT and Multiscale CNNs

Chahat Raj
Department of Information Technology
Delhi Technological University
Delhi, India
0000-0003-0083-6812

Priyanka Meel
Department of Information Technology
Delhi Technological University
Delhi, India
0000-0002-1195-1712

2021 2nd Global Conference for Advancement in Technology (GCAT) | 978-1-6654-1836-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/GCAT52182.2021.9587698

***Abstract*—Online news consumption has rapidly increased, and so has the proliferation of false information. People worldwide have mainly become dependent on social media networks to intake news about the happenings around them. Also, the data is profoundly contaminated with wrong information that harms society in uncountable ways. It is of huge importance to be able to identify a false message. The research society is contributing to solving the problem by developing machine learning and deep learning algorithms. With misinformation spreading ubiquitously, various data modalities have emerged that become carriers of such false news. Research trend is advancing towards multi-modal fake news detection to authenticate text, images, and videos on the web. Existing studies have elaborated on the successful use of RNNs and CNNs. Being a new NLP technique, BERT has been used by a limited number of studies, while multiscale CNNs have not been explored yet to apply fake news detection. This research proposes a novel framework using BERT and multiscale CNNs to perform multi-modal fake news classification and achieve results higher than the existing state-of-the-art techniques.**

Keywords—Fake News, Misinformation, Multimodal, BERT, Convolutional Neural Networks.

I. INTRODUCTION

The problem of false information on the internet has grown into a global menace. With the overwhelming amount of data being created every instant, it has become difficult for internet users to distinguish between honest and deceptive content simply because of the volume at which its creators are discussing the deceptive content. Information created about real-time events is more prone to be engulfed by deceptiveness. Most users would not be able to verify it with legitimate sources and thereby acquire space to deploy the maleficent intent of its creators. The deception increases multi-fold with images or other forms of accompanying media along with the text, which is referred to as multi-modal information. False news content surfaces each day, spreading through text, images, videos, and speech. The high volume, velocity, and variety of such web information make real-time authentication difficult. Internet users fall prey to such ill content, which has multi-faceted effects on their lives. One such example of multimodal fake news is illustrated by Fig. 1 where a user tweets an image of a shark stating that it was spotted on the freeway in Houston during Hurricane Harvy. The global scenario of misinformation demands the tools to detect multi-modal content on the social platforms and efficiently prevent users from being affected by a stream of deceptive content, especially those with ill intent. Amongst the techniques used to create such multi-modal detectors, those based on deep learning algorithms have proved to be most accurate and have shown the ability to process large chunks of data at a time.

Social media platforms leverage users to interact and communicate freely. Twitter and Sina Weibo are two primarily used interactive platforms bringing together users' views through their posts. These online networks allow people to communicate through a fixed short-length text, images, and videos. These posts are termed microblogs. Such posts can at times be misinformative (unintentionally false) or disinformative (deliberately false). Machine learning algorithms have established baselines in the detection of fake textual information online. The research gap lies in authenticating multi- modal web content where information consists of multiple data modalities such as text, image, and video combinations. This has increased the complexity of verifying online news automatically. Recent research is advancing towards using deep neural networks to aid such verification tasks. This has resulted in the efficient detection of fake content. However, there is a lack of a robust algorithm that classifies real and fake content with high accuracy.

Fig. 1. Example of a Deceptive Microblog on Twitter

This paper presents a novel framework designed to perform microblogs classification using their textual and visual information. The framework is built on deep learning techniques, namely, Bidirectional Encoder Representations from Transformers (BERT) and Multiscale Convolutional Neural Networks for feature extraction. The proposed architecture proficiently categorizes fake and real news, with accuracies higher than the existing baselines.

The contributions of this paper are:

- We structure a novel framework that is built upon BERT and convolutional neural networks for fake news detection.
- We propose a novel architecture, namely Multiscale

CNN, consisting of two 2-D convolutional networks, each with a different parameter configuration for better feature extraction.

- The model incorporates a two-level early fusion technique of concatenation in which the first level fusion combines various features extracted from images from two convolutional neural networks. The second level fusion combines features from both textual and visual data streams to generate final classification.
- The performance of the proposed approach is established using two real-world datasets: Twitter (2016) and Sina Weibo (2016), and achieve 75% and 91% accuracies on each, respectively, outperforming the existing state-of-the-art frameworks.

The organization of this paper is as follows: Section 2 navigates through the current research situation in the fake news domain and discusses existing techniques. Section 3 explains the proposed framework, its methodology, and architecture. Section 4 describes the datasets used, implementation procedure, results obtained, and baseline comparison. Section 5 concludes the paper by summarizing its contributions.

II. Related Works

With the rise in online information and a variety of data modalities, identifying false information has become dramatically complex. Traditional NLP approaches are only able to classify text based on its linguistic features. Only text-based classification is not enough for a piece of information to be correctly identified as fake or real. Fake news has to be characterized from varied perspectives to detect fake information robustly. Various studies in the past have highlighted the importance of multi-modal classification. This methodology views fake information on the internet as a combination of multiple data modalities. Studies have delved into analyzing news by using both text and images present in the piece of information. Though various algorithms for fake news detection have been designed, a curb to the problem has not yet been provided. Researchers have contributed by building multi- modal classifiers.

Wang et al. [1] used neural networks to create an Event Adversarial Neural Network (EANN). The model performs text classification using TextCNN and pre-trained VGG19 for image classification. The classification model is coupled with an event discriminator that discards event-specific features from the news items and keeps only shared characteristics for training. Another model by Khattar et al. [2], Multimodal Variational Autoencoder (MVAE), is built using bidirectional LSTM and VGG19 for text and image classification. The model is based on an encoder-decoder approach using a binary classifier to detect false information. The model, highly complex with two main components, takes longer than usual training time. Jin et al. [3] developed a multi-modal approach using Recurrent Neural Networks coined with an attention mechanism for classifying rumor and non-rumor microblogs. The model exploits textual, visual and contextual features to categorize a tweet. Proposing a novel approach, Qi et al. [4] designed Multi- domain Visual Neural Network (MVNN). Convolutional and Recurrent Neural Networks are used to retrieve pixel and frequency domain features. The patterns from both pipelines are fused using an attention mechanism. Liu and Wu [5] have also utilized a combination of CNN and RNN for the same. Yang et al. [6] combined explicit features of news items and implicit text and visual features for false news classification. CNNs are used for both text and image classification. Singhal et al. [7], by developing SpotFake, designed a method that uses Bidirectional Encoder Representations from Transformers (BERT) for textual feature extraction and VGG19 visual feature extraction, which is combined to give the final output. The model removes any need for an event discriminator and beats the existing state of the arts. Experimenting with text sentiments, Cui et al. [8] developed Sentiment Aware Multi-modal Embedding (SAME). This model combines user comments and their sentiments with the classification model that uses LSTM and CNN. Vishwakarma et al. [9] designed a veracity analysis model that analyses fake news images, also using the text present in them. Gupta and Meel [10] tackle this problem using supervised machine learning by designing a passive-aggressive classifier. Raj and Meel [11] proposed a Convnet approach by combining Text-CNN and Image-CNN, experimenting with AlexNet, ResNet50, MobileNetV2, VGG16, VGG19, InceptionV3, XceptionNet and DenseNet. They combine the multimodal streams of text and image using weightage average fusion mechanism. Meel and Vishwakarma [12] ensembled multiple approaches namely Hierarchical Attention Network (HAN), image captioning and image forensics techniques to analyse textual and visual manipulations in online news. In another work [13], they developed a semi-supervised convolutional approach exploiting linguistic and stylometric information in the textual content. Wu et al. [14] employed an Adversarial Network using the concept of Shared-Private model for credibility verification. Giachanou et al. [15] designed a neural network approach that fuses semantic details with textual and visual information making use of VGG16, VGG19, ResNet, Inception and Xception along with textual feature extractors. Song et al. [16] designed a novel architecture using residual networks for considering multimodal features. They employ Cross-modal Attention Residual and Multichannel convolutional neural Networks (CARMN) to extract multimodal features from multiple target domains. Jin et al. [17] combined statistical features with visual features and performed fake news classification using several machine learning algorithms like Support Vector Machine (SVM), Logistic Regression, Random Forest and K-Star algorithm.

With the advances in machine learning, BERT models have shown tremendous efficiency for Natural Language Processing (NLP) tasks. Also, image classification tasks are widely built upon CNN models pre-trained on ImageNet [18] dataset. Multiscale convolutional architectures have not yet been explored for false news detection task. Thus, we create a detection mechanism using BERT and Multiscale CNNs and explore its efficacy.

III. Methodology

We propose a novel multimodal framework using BERT and Multiscale CNN to address the fake news issue. The idea behind the framework is to build a multi-modal algorithm that identifies false information based on linguistic and visual features present in a microblog post. Multi-modal data is processed through two pipelines for text and image inputs. The resulting feature vector sequences are concatenated to

provide the final prediction. In this section, we elaborate on the components of the proposed framework. The architecture of the proposed framework is provided in Fig. 2.

BERT$_{BASE}$: For textual representations, the proposed approach uses the Bidirectional Encoder Representations from Transformers (BERT) technique. Text sequences from the datasets are fed to the BERT$_{BASE}$ model, which consists of twelve encoders. These encoders are individual transformer blocks with twelve self-attention heads. We use the model pre- trained on plain text from vast resources of English Wikipedia and BookCorpus. The choice of this model is made due to its advantage over word2vec and GloVe embeddings. BERT can easily differentiate among the contexts of various occurrences of the exact words, in which other embeddings are incapable. The input sequences from the datasets move sequentially through each of the twelve encoder layers. Self-attention is applied at each encoder layer, and the results are forwarded to the next encoder in sequence. The output is sent through a fully connected layer, and resulting data are the required textual feature vectors, F_t.

Multiscale CNN: Existing literature in fake news detection has primarily used pre-trained convolutional neural networks. This study proposes the use of multiscale CNNs for this task. This usage is supported by the fact that a single sequential convolutional network caters to the specified input size. Whereas the advantage of multiscale CNNs lies in the allowance of running several convolutional models with different filter input sizes simultaneously. This lets the model analyse distinguished pixel regions. In practicality, the model can identify various objects in the image with increased precision than the single- scale convolutional model. The image sequences from the datasets are fed to the proposed multiscale model consisting of two sequential convolutional networks (Fig. 2). CNN 1 and CNN 2 consist of four and five 2D-convolutional layers, respectively. Each of these layers is followed by the ReLU activation function and two-dimensional max-pooling operational layers. We use kernel sizes of 3 and 5 for CNN 1 and CNN 2, thus allowing them to capture pixel information from different sized regions in the images and improve the model's efficiency.

Two-level Early Fusion: To combine various components of the proposed framework, we utilize concatenation which is an early fusion approach. This technique uses the concat function that fuses feature vectors from multiple streams. In our work, this early fusion technique is used at two levels: for adjoining CNN 1 and CNN 2 for image feature vector generation (F_v) and for further fusing this output with the textual feature vectors (F_t) received from the BERT pipeline.

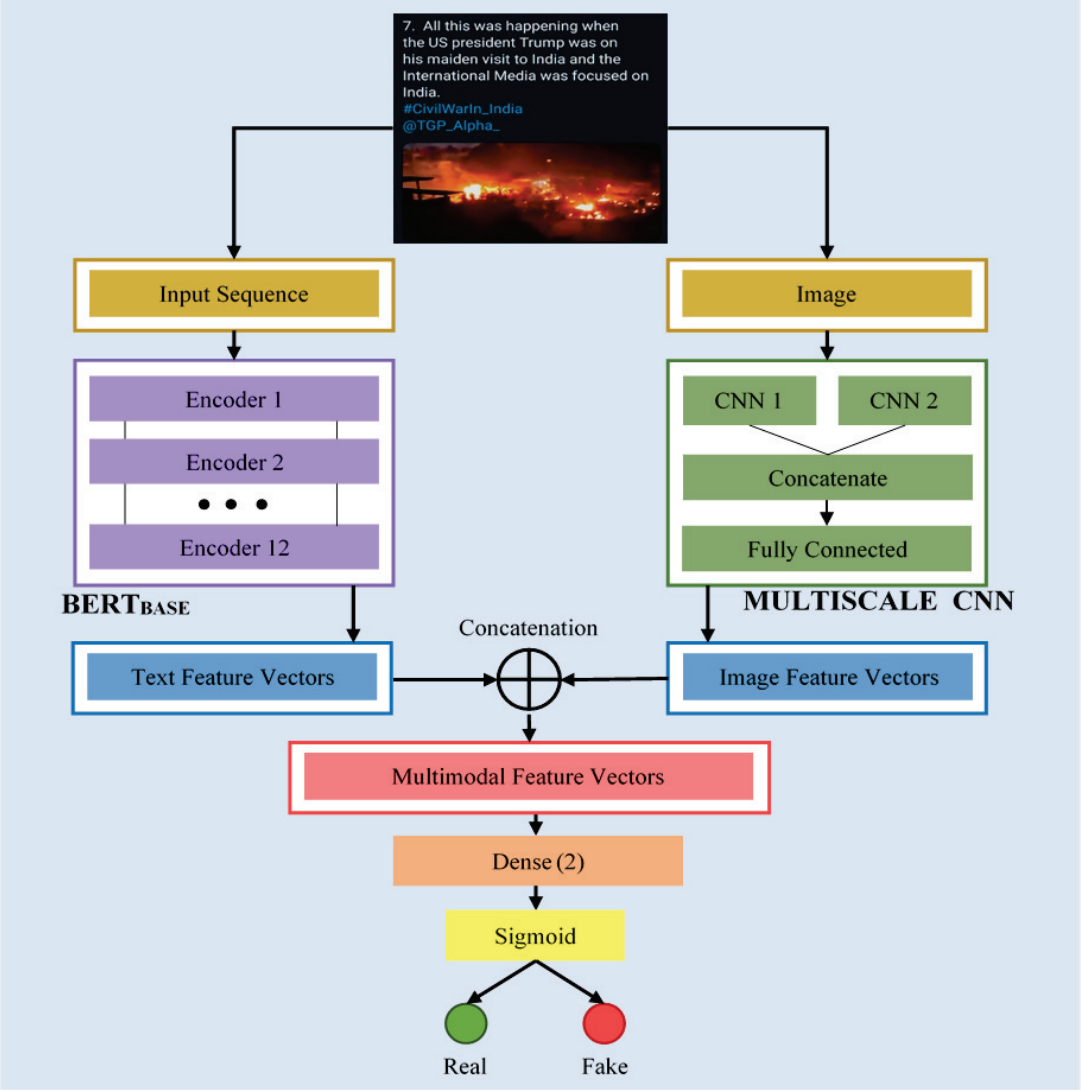

Fig. 2. Proposed framework for multi-modal fake news detection using BERT and Multiscale CNN

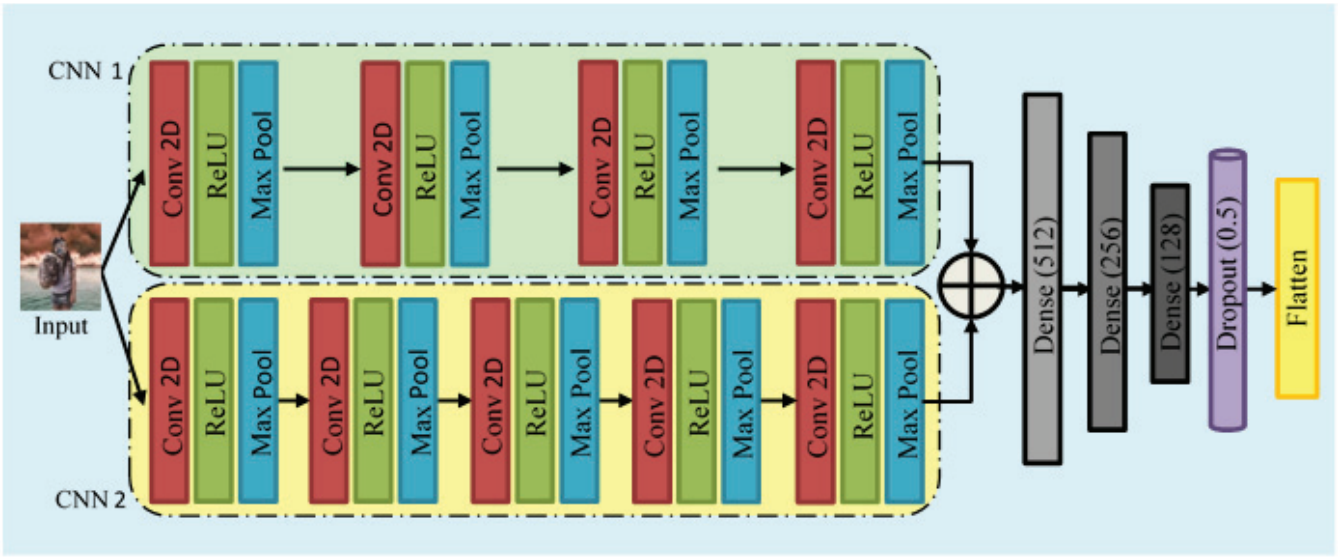

Fig 3. Proposed architecture of Multiscale Convolutional Neural Network

Mathematically, the visual feature vector F_v, is obtained by the operation,

$$F_v = F_{v1} \oplus F_{v2} \quad (1)$$

Where $Fv1$ and $Fv2$ are the outputs from CNN 1 and CNN 2 and $\oplus$ is the concatenation operation. To generate the multi- modal feature vector F, another concatenation operation is performed on textual and visual feature vectors (2).

$$F = F_t \oplus F_v \quad (2)$$

After receiving the final multi-modal features, dense layers are added with shapes reducing consequently from 512 to 256 to 128 and then 2. To reduce overfitting, a Dropout of value 0.5 is used. A sigmoid layer is added thereafter to generate the final classification result that predicts the microblog into the real or fake category. The parameters mentioned above have been decided upon by performing iterative experiments and thus selecting the values resulting in the best performance.

IV. Experiments

This section describes the datasets used for assessing the proposed framework. We then discuss the results achieved during the testing phase and their comparison with established state-of-the-art algorithms.

A. Datasets

Experiments are performed on two standard publicly available real-world datasets: Twitter and Weibo. These datasets are collections of microblogs from social media networks. With the availability of multi-modal information in the microblogs, these datasets are appropriate for our work.

Twitter: This benchmark dataset was released as a part of the MediaEval 2016 [19] workshop for the 'Verifying Multimedia Use' task. The task aimed at detecting fake news on Twitter. The dataset consists of various information like text, image/video, and social context. The dataset consists of two sets: the development set and the testing set. We filter tweets containing both textual information and images. The development set is used for training and validation, and the algorithm is verified on the testing set.

Weibo: This dataset is a collection of microblogs from the authoritative news agency of China, Xinhua News Agency, and Weibo, a Chinese microblogging service. The dataset is a collection of microblogs collected between May 2012 to January 2016. The collection is verified by the official rumor debunking system of China. For the experiments, tweets along with their images are used. The dataset is split in a 4:1 ratio in accordance with previous literature and their investigations.

B. Implementation

Google Colab is used for performing all experiments using python 3. It allocated 12 GB NVIDIA TESLA K80 GPU and 13.53 GB of RAM. For NLP processing, the NLTK library is used. Tokenization is performed using Regex. Stemming and Porter Stemmer and WordNet Lemmatizer aid lemmatization. For image classification, images are pre-processed to a fixed input size of 224*224 to feed to 2D CNNs. The sigmoid function is used for supporting the classification into binary categories: real and fake. Adam optimizer is used for the visual model. Images are trained with a batch size of 64, the maximum which Google Colab could allocate in a runtime. We train the neural network model for 200 epochs using early stopping. The final result is calculated using four necessary metrics: accuracy, precision, recall, and f1- score.

C. Results

In this section, we describe the achieved results on two datasets, Twitter and Weibo, and compare them with the results of existing state-of-the-art methods, Att-RNN [3], EANN [1], and TI-CNN [6], in terms of accuracy, precision, recall, and f1- score. We re-implemented these state-of-the-art algorithms by setting up the environment and parameters used in their research to validate the performances. A baseline comparison of our approach with these methods is provided in Table I.

On the Twitter dataset, we achieve a real/fake classification accuracy of 75%. The precision, recall, and f1 scores are reported to be 85.1%, 72.9%, and 75.8%, respectively. For the Weibo dataset, the classification

accuracy obtained is 91.4%. precision, recall, and f1 scores are 87%, 89.1%, and 85.8%, respectively.

TABLE I. RESULTS AND BASELINE COMPARISON ON TWITTER AND WEIBO DATASETS

Dataset	Method	Accuracy	Precision	Recall	F1-Score
Twitter	Att-RNN [3]	0.664	0.749	0.615	0.676
	EANN [1]	0.715	0.822	0.638	0.719
	TI-CNN [6]	0.732	0.840	0.712	0.745
	Our Approach	**0.750**	**0.851**	**0.729**	**0.758**
Sina Weibo	Att-RNN [3]	0.779	0.778	0.799	0.789
	EANN [1]	0.827	0.847	0.812	0.829
	TI-CNN [6]	0.831	0.839	0.826	0.844
	Our Approach	**0.914**	**0.870**	**0.891**	**0.878**

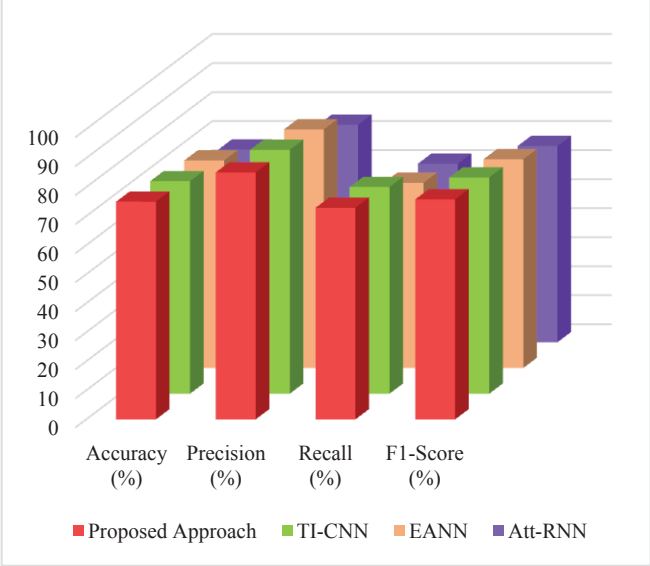

Fig 4. Performance comparison on Twitter Dataset

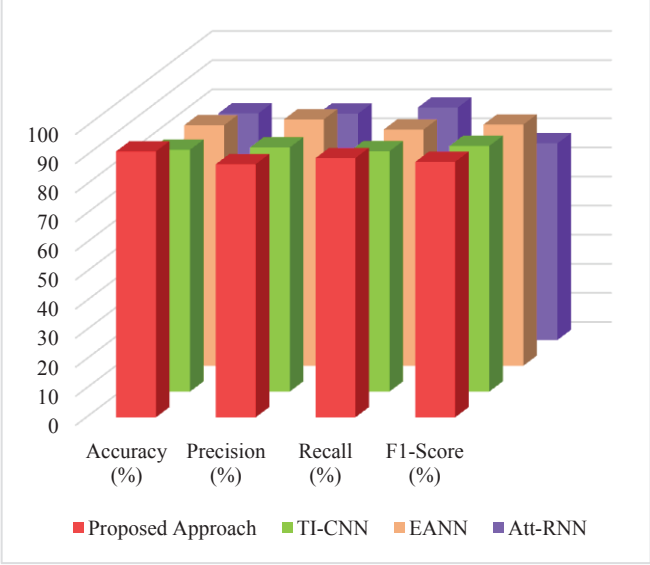

Fig 5. Performance Comparision on Sina Weibo Dataset

As observable from the table, our approach surpasses the scores obtained by existing state-of-the-art techniques and provides better classification results. The approach, robust and quick to train, can perform efficient classification of fake and real news.

Visual comparison of results on Twitter and Weibo datasets are illustrated in Fig. 4 and Fig. 5. The performance of our model is regarded to the proposed combination of BERT and Multiscale CNNs. The use of the pre- trained BERT model enhances text-based fake news detection as it is capable of understanding the contextual meanings of words for every usage. This is the first research that uses a Multiscale CNN approach for fake news detection to the fullest of our knowledge. The novel architecture built on two different CNNs allows the visual feature extraction from different pixel regions with varying kernel sizes in an image, thus improving overall image classification accuracy. In addition to this, the early fusion approach combines feature vectors at an early stage prior to training the model. The model is trained on a single combined feature vector set F, eliminating the need to train each network individually. This results in lower training time and making the proposed framework highly robust.

V. CONCLUSION

This work proposes a novel approach for the veracity analysis of microblogs online. The proposed framework performs the classification of social media information into fake or real. The framework components include a 12-encoder BERT for extracting textual features and a multiscale convolutional model with two CNNs for extracting visual features. The final prediction is made by concatenating the outputs obtained from BERT and Multiscale CNN model, using two-level early fusion. The experimentation is performed on two publicly available real-world datasets, Twitter and Weibo. Results achieved via experimentation demonstrate the competence of the proposed framework. Our architecture has surpassed state-of-the-art methods in fake news detection by 2% and 8% on Twitter and Weibo, respectively. We aim to build more robust algorithms that could be useful for real-world applications to conquer the misinformation scenario for future work.

REFERENCES

[1] Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., ... & Gao, J. (2018, July). Eann: Event adversarial neural networks for multi-modal fake news detection. *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining* (pp. 849-857).

[2] Khattar, D., Goud, J. S., Gupta, M., & Varma, V. (2019, May). Mvae: Multi-modal variational autoencoder for fake news detection. In *The World Wide Web Conference* (pp. 2915-2921).

[3] Jin, Z., Cao, J., Guo, H., Zhang, Y., & Luo, J. (2017, October). Multi- modal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on Multimedia* (pp. 795-816).

[4] Qi, P., Cao, J., Yang, T., Guo, J., & Li, J. (2019, November). Exploiting multi-domain visual information for fake news detection. In *2019 IEEE International Conference on Data Mining (ICDM)* (pp. 518-527). IEEE.

[5] Liu, Y., & Wu, Y. F. (2018, April). Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 32, No. 1).

[6] Yang, Y., Zheng, L., Zhang, J., Cui, Q., Li, Z., & Yu, P. S. (2018). TI- CNN: Convolutional neural networks for fake news detection.

arXiv preprint arXiv:1806.00749.

[7] Singhal, S., Shah, R. R., Chakraborty, T., Kumaraguru, P., & Satoh, S. I. (2019, September). Spotfake: A multi-modal framework for fake news detection. In *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)* (pp. 39-47). IEEE.

[8] Cui, L., Wang, S., & Lee, D. (2019, August). Same: sentiment-aware multi-modal embedding for detecting fake news. In Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining (pp. 41-48).

[9] Vishwakarma, D. K., Varshney, D., & Yadav, A. (2019). Detection and veracity analysis of fake news via scrapping and authenticating the web search. *Cognitive Systems Research*, *58*, 217-229.

[10] Gupta, S., & Meel, P. (2021). Fake News Detection Using Passive-Aggressive Classifier. In *Inventive Communication and Computational Technologies* (pp. 155-164). Springer, Singapore.

[11] Raj, C., & Meel, P. (2021). ConvNet frameworks for multi-modal fake news detection. *Applied Intelligence*, 1-17.

[12] Meel, P., & Vishwakarma, D. K. (2021). HAN, image captioning, and forensics ensemble multimodal fake news detection. *Information Sciences*, *567*, 23-41.

[13] Meel, P., & Vishwakarma, D. K. (2021). A temporal ensembling based semi-supervised ConvNet for the detection of fake news articles. *Expert Systems with Applications*, *177*, 115002.

[14] Wu, L., Rao, Y., Nazir, A., & Jin, H. (2020). Discovering differential features: Adversarial learning for information credibility evaluation. *Information Sciences*, *516*, 453-473.

[15] Giachanou, A., Zhang, G., & Rosso, P. (2020, September). Multimodal fake news detection with textual, visual and semantic information. In *International Conference on Text, Speech, and Dialogue* (pp. 30-38). Springer, Cham.

[16] Song, C., Ning, N., Zhang, Y., & Wu, B. (2021). Knowledge Augmented Transformer for Adversarial Multidomain Multiclassification Multimodal Fake News Detection. *Neurocomputing*.

[17] Jin, Z., Cao, J., Zhang, Y., Zhou, J., & Tian, Q. (2016). Novel visual and statistical image features for microblogs news verification. IEEE transactions on multimedia, 19(3), 598-608.

[18] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*, 1097-1105.

[19] Larson, M., Soleymani, M., Gravier, G., Ionescu, B., & Jones, G. J. (2017). The benchmarking initiative for multimedia evaluation: MediaEval 2016. *IEEE MultiMedia*, *24*(1), 93-96.