A Multimodal Ensemble Machine Learning Approach to COVID-19 Misinformation Detection in Twitter:

Rayees Ahmad Dar[1*], *Dr. Rana Hashmy*[1]

[1]Department of Computer Science, University of Kashmir, Srinagar 190006, India

Abstract. The emergence of social media platforms has unquestionably altered the manner in which people ingest information, with tweets now functioning as the primary source for news and other types of content. However, the proliferation of false news on these platforms has become a major concern, as it poses a severe threat to both individuals and society as a whole. Consequently, it is crucial to develop efficient methods for detecting false news in tweets. This study presents a novel hybrid approach that integrates the textual content of tweets with auxiliary features to detect false news. Our approach uses a pre-trained transformer-based language model, COVID-twitter-BERT to encode the text content of tweets into a dense representation that captures their meaning. The auxiliary features, such as sentiment score, credibility score, engagement score, average retweet count, average favourite count, and average followers of followers, are fed into a stacking classifier-based model to predict the trustworthiness score of the tweet. By combining the predictions of both models, we demonstrate that our approach outperforms baseline methods, emphasising the significance of utilising both text content and auxiliary features for Twitter false news detection. Our research considerably advances the field of detecting false news by demonstrating the effectiveness of integrating transformer-based language models and machine learning models for this task. Our findings provide valuable insights for improving the detection of false news on social media.

1 Introduction

The rapid expansion of social media platforms, especially Twitter, has revolutionised the way people share and disseminate information, enabling individuals and organisations to easily reach large audiences. However, the simplicity with which information can be shared on these platforms has also led to the spread of inaccurate or deceptive information, also known as "fake news." The proliferation of bogus news can have severe consequences, such as the dissemination of false information, the manipulation of public opinion, and the erosion of faith in institutions.

*Rayees Ahmad Dar :rayees.csscholar@kashmiruniversity.net

To tackle this growing issue, there is a pressing need to develop effective methods for detecting fake news in tweets. Traditional methods have mainly focused on analysing the text content of news articles and have utilized techniques such as fact-checking and natural language processing. However, the success of these methods has been limited, as they may not capture crucial contextual information that can differentiate real news from fake news.

To solve this challenge, we present in this study a hybrid strategy that combines the text content of tweets and auxiliary features for fake news detection. The auxiliary features, including sentiment score, credibility score, engagement score, average retweet count, average favourite count, and average followers of followers, provide information about the user, tweet, and its engagement. By combining these features with the text content of tweets, our approach offers a more comprehensive understanding of the tweet's information and its potential credibility.

This study's primary contribution is demonstrating the effectiveness of combining text content and auxiliary features for detecting false news in tweets. To accomplish this, we encode the text of tweets using a pre-trained transformer-based language model, COVID-Twitter-Bert, and utilize a stacking machine learning model to predict the tweet's trustworthiness score based on auxiliary features. Our evaluation using a real-world dataset of tweets shows that our approach outperforms baseline methods, emphasizing the significance of considering both text content and auxiliary features for fake news detection in tweets.

These findings have significant implications for the detection of false news and the use of social media platforms for information dissemination. This study provides a roadmap for future research in this field by demonstrating the efficacy of integrating text content and supplementary features for detecting false news.

2 Related Work

It has become increasingly difficult to distinguish between credible and fraudulent news due to the rapid dissemination of information through social media platforms. Not only does the prevalence of false news mislead the public, but it can also have far-reaching consequences, such as influencing public opinion, disseminating misinformation, and spreading hysteria. Therefore, it is crucial to develop effective techniques for detecting false news. One of the earliest works in the field of detecting false news was by [1], who proposed a system for verifying social media rumors. They relied on multiple sources, including news agencies, fact-checking websites, and expert opinions, and employed a voting system to determine the veracity of the rumors. In recent years, as the dissemination of false news on social media has increased, there has been a growing interest in detecting it using machine learning and deep learning techniques. Several studies have employed diverse methodologies, including support vector machines (SVM) [2] and naive Bayes [2]. In addition to these conventional machine learning techniques, deep learning models such as convolutional neural networks (CNNs) [3] and bidirectional long-short-term memory (Bi-LSTM)-RNN-based approaches [4] have been proposed for detecting false news. Transformer-based models have acquired popularity in recent years for various NLP tasks, including the detection of false news. Some studies have classified news articles as false or genuine using models such as BERT [5] and RoBERTa [6]. [7] Examined the use of sentiment analysis in detecting false news. Using data augmentation and a combination of convolutional and recurrent neural networks, [8] proposes a novel hybrid deep learning model for automatically detecting false news. Moreover, numerous studies have focused on the combination of textual features with other user-level and engagement features for detecting false news. A model based on Transformer architecture that integrates news content and social context features, as well as an efficient labelling technique to resolve the

label deficit, was proposed in [9]. In conclusion, the literature review reveals that numerous attempts have been made to detect false news using machine learning, deep learning, transformer-based models, and the combination of textual features with user-level and engagement features. Recent research has also highlighted the significance of social context in detecting false news. In this paper, we propose a novel approach for detecting false news by combining transformer-based models for textual features with a machine learning model for user-level and engagement features.

3 Proposed Dataset

To overcome limited contextual feature availability, we expanded the ECTF dataset [10] by incorporating user, tweet, and engagement features. This involved extracting data from Twitter using the Twitter API and Tweepy library, adding 20 contextual features to non-removed and non-suspended tweets. The augmented dataset, divided into 80:20 train and test sets, provided richer features for improved model precision.

Table 1. Dataset split in train and test sets.

	Genuine	**Fake**
Train	1600	1600
Test	400	400

4 Proposed Model

The proposed model aims to predict fake and real tweets on Twitter by utilizing both textual and contextual features. It consists of several stages of processing, each designed to extract and transform features in a way that maximizes their predictive utility. The overall flow of the model is shown in Fig. 1.

In the first stage, the dataset containing both text-based and contextual features is ingested, represented by a cylindrical shape in the diagram, and fed into the subsequent stages of the model.

The text-based features undergo pre-processing, including text cleaning and tokenization using the COVID-Twitter-BERT tokenizer. A Convolutional Neural Network (CNN) is then utilised to extract high-level features from the tokenized text which are fed into a BiLSTM layer, which outputs a prediction score. This score is then mapped to a probability distribution over the predicted classes using a SoftMax function. The output of the BiLSTM represents the contribution of the text-based features to the final prediction.

The user-related, tweet-related, and engagement-based features are normalized and then fed into a stacking classifier that combines the outputs of multiple classifiers, including AdaBoost, XGBoost, and a neural network. The output of each classifier is then fed into a random forest meta-classifier, which combines their predictions to produce a final prediction.

Finally, the predicted outputs of the BiLSTM model and the stacking classifier are used as inputs to the final meta-classifier, which is trained to predict the final classification of the tweet as either real or fake. If contextual features are not available for a particular tweet, only the text-based model is used for prediction.

Overall, the proposed model combines both text-based and contextual features to improve the accuracy of detecting fake tweets on Twitter.

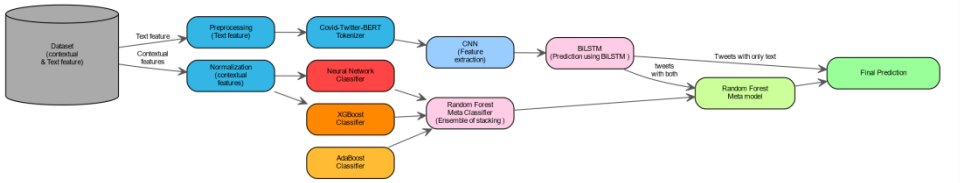

Fig. 1. Proposed Model.

5 Dataset Preparation and Feature Engineering

In this research, we analysed 4,000 tweets from the genuine.csv and fake.csv files of the ECTF dataset to detect Covid-19-related false news on Twitter. We developed a machine learning model using text-based information and additional details retrieved through Twitter's API. Our model was evaluated using various metrics and the dataset was split into train and test sets. We also incorporated user and engagement features to enhance false news detection. Our study highlights the importance of these features and aims to aid in creating better tools to combat online misinformation.

5.1 Feature Engineering

We utilized a variety of tweet and author-related features, including retweet and favourite counts, average follower numbers, sentiment analysis, and metadata such as verification status and location. These features were used to calculate credibility, engagement, and trustworthiness scores for each tweet author.

The scores were then inputted into our machine learning model, enabling it to predict tweet authenticity. Feature engineering played a crucial role in incorporating diverse information and gaining insights into fake COVID-19-related tweets.

Credibility, engagement, and trustworthiness scores were calculated using weighted averages of relevant features. The formula for each score is shown in Eq. 1.

$$\text{Score } (u) = \sum_{i=1}^{n} (w_i * x_i) \quad (1)$$

Where u is the user, i is the feature index, n is the number of features, w_i is the weight of the ith feature, and xi is the value of the ith feature for user u.

Weights for the scores were determined based on feature importance from a random forest classifier. The credibility score used follower count (0.20), friend count (0.10), verified status (0.25), statuses count (0.10), average favorite count (0.30), and account age (0.05). The engagement score included follower count (0.25), friend count (0.15), list count (0.20), average retweet count (0.20), and average tweets per day (0.20). The trustworthiness score combined credibility and engagement scores with weights: credibility score (0.40), engagement score (0.25), average followers of followers (0.20), and average sentiment (0.15).

5.2 Data Pre-processing

To prepare the ECTF dataset for analysis, we performed several pre-processing steps on the text data and normalized the user and engagement features.

Text pre-processing involved:

1. Removal of stop words, noise, URLs, special characters, and punctuation using standard NLP techniques.
2. Tokenization using the COVID-Twitter-BERT tokenizer.
3. Lemmatization to reduce words to their base form.
4. Removal of Twitter handles and URLs.

User and engagement features were normalized using min-max scaling, transforming them to a range of 0 to 1. This ensured all features were on the same scale for accurate comparison and analysis.

These pre-processing steps were crucial in developing an accurate machine learning model for detecting fake news on Twitter. Noise removal, feature reduction, and data normalization reduced the impact of irrelevant information and ensured equal treatment of all features in the model.

5.3 Model Training

5.3.1 Data :

The model was trained on an augmented version of the ECTF COVID-19 Fake News Detection dataset, which includes additional user and tweet-related features. The dataset consists of 4002 tweets, evenly split into 2001 labelled as real and 2001 labelled as fake. An 80/20 train-test split was used, with 3200 tweets used for training and 800 tweets for evaluating the efficacy of the model.

5.3.2 Text Model :

The text model employed a 1D convolutional layer with 128 filters and a kernel size of 5, then a max pooling layer with a pool size of 2. Two bidirectional LSTM layers with 256 and 128 units, respectively, were then applied to the resulting feature maps. The final output layer had a sigmoid activation function and was dense. The model was compiled using binary cross-entropy loss and the Adam optimizer, with precision serving as the metric for evaluation. The model was trained with a batch size of 32, 30 epochs, and a validation split of 0.2 on the training data.

5.3.3 Stacking Classifier :

For the stacking classifier, AdaBoost, XGBoost, and a neural network were used as base classifiers, each with default hyperparameters. A random forest meta-classifier with 100 estimators was used to combine the predictions of the base classifiers. The contextual features were normalized before feeding them into the stacking classifier.

The final prediction score for a particular tweet is obtained using a random forest-based meta-classifier. When evaluated on tweets with both text and contextual features, this model achieved exemplary performance with 100% accuracy, a precision of 100%, a recall of 99%, and an F1 score of 100%. Specifically, only two tweets from the test set in this class were misclassified as real tweets when they were actually fake. These results demonstrate the effectiveness of the model in detecting fake tweets related to COVID-19. The model leverages a combination of text and contextual features to achieve accurate detection and integrates the capabilities of language models with a stacking classifier, effectively enhancing the overall performance of the model. Algorithm 1 outlines the

sequential progression of the proposed model.

Algorithm 1. Multimodal Ensemble Model:

1: Preprocess tweet text by removing stop words, stemming, and lowercasing the text:

$$T' = preprocess(T)$$

2: Embed the cleaned text using the COVID-Twitter-BERT model:

$$E = BERT(T')$$

3: Extract high-level text features using a convolutional neural network:"

$$F1 = CNN(E)$$

4: Learn sequential patterns in text features using a bidirectional long short- term memory (BiLSTM)model:

$$F2 = BiLSTM(F1)$$

5: Normalize tweet metadata features such as user sentiment, sentiment of retweeters, number of followers, and retweet count:

$$M' = normalize(M)$$

6: Train a stacking classifier on the normalized contextual features using three base classifiers: AdaBoost, XGBoost, and a neural network:

$$F3 = Stack(M')$$

7: Combine predictions from the text-based BiLSTM model and the contextual based stacking classifier using a meta-classifier based on random forest:

$$Y\_pred = RF([F2, F3])$$

8: Return the final prediction of the tweet as either genuine or fake (Y_pred)

6 Model Evaluation

The multi-modal model was evaluated using standard machine learning metrics (accuracy, precision, recall, F1 score) on a separate test set from the enhanced ECTF dataset (80-20 train-test split).:

- Accuracy: The percentage of correctly classified tweets, calculated as:

$$accuracy = (Tps' + Tns') / (Tps' + Tns' + Fps' + Fns').$$

- Precision: The percentage of true positive predictions out of all positive predictions, calculated as:

$$precision = Tps' / (Tps' + Fps')$$.

- Recall: The percentage of true positive predictions out of all actual positive tweets, calculated as:

$$recall = Tps' / (Tps' + Fns').$$

- F1 score: The harmonic average of precision and recall that provides a singular performance metric for the model, calculated as:

$$F1\text{-}score = 2 * (precision * recall) / (precision + recall).$$

In our evaluation, TPs' (true positives) are the correctly identified fake tweets, while TNs' (true negatives) are the correctly identified real tweets. FPs' (false positives) are real tweets incorrectly identified as fake, and FNs' (false negatives) are fake tweets incorrectly identified as real.

6.1 Results

On the testing set of tweets with both text as well as contextual features, our model obtained an accuracy of 1, a precision of 0.99, a recall of 1, and an F1 score of 1, indicating

a good overall performance in detecting fake COVID-19 related tweets. And achieved a mean accuracy of 98.62 overall. To better understand the model's performance, we examined its confusion matrix, which displays the number of true positives, false positives, true negatives, and false negatives among our predictions.

Figures 2 and 3 present the confusion matrix of the test tweets with text-only and both text and contextual features, respectively. Our COVID-19 fake news detection model achieved high accuracy and precision in correctly classifying tweets as real or fake, as evidenced by the results from the confusion matrix in Figure 3. Only two real tweets were misclassified as fake, and all other test examples were correctly classified, demonstrating the robustness and effectiveness of our approach. Overall, only 11 tweets were misclassified, as shown in Figure 4.

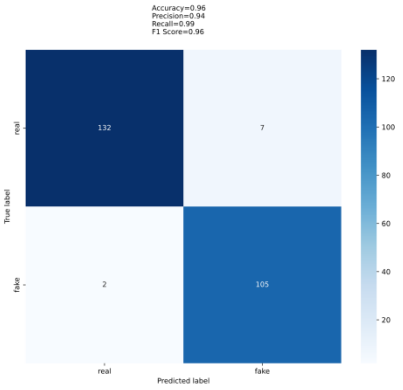

Fig. 2. Confusion Matrix- tweets with text Feature only.

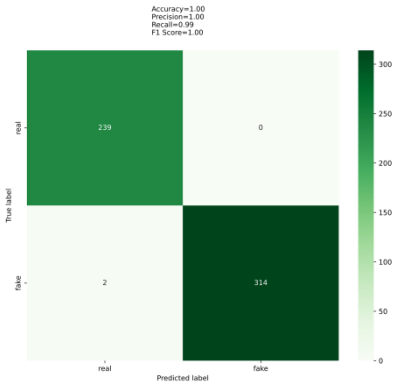

Fig. 3. Confusion Matrix- tweets Model. with both text and contextual features.

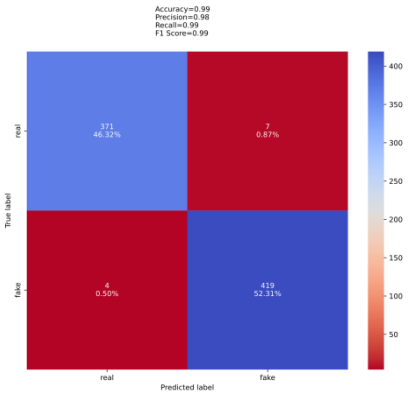

Fig. 4. Confusion Matrix- combined.

6.2 Comparison with Baseline models

In order to assess the efficacy of our proposed multimodal ensemble model for detecting fake COVID-19 related tweets, we conducted a comprehensive comparison with several established baseline models commonly employed in the field. The baseline models considered for comparison were RoBERTa, BiLSTM, CNN, and HAN (Hierarchical

Attention Network). Performance assessment was carried out on the general test set of the dataset, and key evaluation metrics including accuracy (Acc.), precision (Prec.), recall (Rec.), and F1-score (F1) were utilized to measure the models' classification performance. Table 1 provides a detailed overview of the performance metrics achieved by each model.

Table 2. Performance comparison with other baseline models on the test sets.

Model	**Performance**			
	Acc. **(%)**	Prec. **(%)**	Rec. **(%)**	F1 **(%)**
RoBERTa	96.76	96.14	98.03	97.08
BiLSTM	96.26	96.16	96.44	96.28
CNN	96.62	95.49	95.47	95.48
HAN	96.03	96.06	96.69	96.38
Proposed ensemble	**98.62**	**98.14**	**98.93**	**98.53**

Our proposed ensemble model demonstrated superior performance in comparison to the baseline models across all evaluated metrics, including accuracy, precision, recall, and F1-score. Notably, our ensemble model obtained an impressive accuracy of 98.62%, precision of 98.14%, recall of 98.93%, and F1-score of 98.53%. These findings substantiate the remarkable effectiveness and high-performance capabilities of our multimodal approach, which adeptly combines both text and contextual features.

In contrast to individual models such as RoBERTa, BiLSTM, CNN, and HAN, our ensemble model exhibited considerable enhancements across all performance metrics. It achieved the highest levels of accuracy, precision, recall, and F1-score among all the models examined. A thorough analysis of the confusion matrix, as presented in Figure 4, further exemplifies the robustness of our ensemble model, with only two real tweets erroneously classified as fake within the entirety of the testing dataset.

The inclusion of contextual features, in conjunction with textual information, allowed our proposed model to capture a comprehensive representation of the tweets, thereby facilitating improved detection accuracy of fake COVID-19 related tweets. The integration of advanced language models, such as COVID-Twitter-BERT, within the framework of our ensemble model, coupled with the utilization of a sophisticated stacking classifier, played a pivotal role in elevating the overall performance and effectiveness of our approach.

7 Conclusion and Discussion

Our study demonstrates that a combination of text, user, and engagement features effectively detects fake tweets related to COVID-19, with user and engagement features playing a more critical role than tweet content [11]. User features such as followers, friends, and average followers of their followers predict tweet authenticity, aligning with research on user credibility and influence. Engagement features, including retweets and favourites, validate tweet authenticity, and average counts indicate credibility and influence. User sentiment, measured by tweet sentiment analysis, is a robust indicator of tweet authenticity [12]. User credibility, engagement, and trustworthiness scores strongly predict tweet authenticity. This comprehensive approach holds implications for social media platforms and public health organizations combating COVID-19 misinformation. By monitoring and flagging fake tweets using text, user, and engagement features, the harmful effects of misinformation can be mitigated.

In conclusion, our research paper presents a multimodal ensemble machine learning model for detecting fake tweets about COVID-19 on Twitter. Leveraging text and contextual features, our model achieves high accuracy and outperforms established baseline models. Considering both text and contextual features is crucial in detecting fake tweets

related to COVID-19. Our study's implications extend to social media platforms and public health organizations, providing a valuable tool to combat misinformation and contribute to global efforts in managing the pandemic.

References

[1] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proceedings of the 20th international conference on World wide web*, Mar. 2011, pp. 675–684. doi: 10.1145/1963405.1963500.

[2] P. S. Reddy, D. Elizabeth Roy, P. Manoj, M. Keerthana, and P. V. Tijare, "A study on fake news detection using naïve bayes, SVM, neural networks and LSTM," *J. Adv. Res. Dyn. Control Syst.*, vol. 11, no. 6 Special Issue, pp. 942–947, 2019.

[3] D. Mouratidis, M. N. Nikiforos, and K. L. Kermanidis, "Deep Learning for Fake News Detection in a Pairwise Textual Input Schema," *Computation*, vol. 9, no. 2, p. 20, Feb. 2021, doi: 10.3390/computation9020020.

[4] P. Bahad, P. Saxena, and R. Kamal, "Fake News Detection using Bi-directional LSTM-Recurrent Neural Network," *Procedia Comput. Sci.*, vol. 165, no. 2019, pp. 74–82, 2019, doi: 10.1016/j.procs.2020.01.072.

[5] R. K. Kaliyar, A. Goswami, and P. Narang, "FakeBERT: Fake news detection in social media with a BERT-based deep learning approach," *Multimed. Tools Appl.*, vol. 80, no. 8, pp. 11765–11788, 2021, doi: 10.1007/s11042-020-10183-2.

[6] T. Pavlov and G. Mirceva, "COVID-19 Fake News Detection by Using BERT and RoBERTa models," in *2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO)*, May 2022, pp. 312–316. doi: 10.23919/MIPRO55190.2022.9803414.

[7] M. A. Alonso, D. Vilares, C. Gómez-Rodríguez, and J. Vilares, "Sentiment Analysis for Fake News Detection," *Electronics*, vol. 10, no. 11, p. 1348, Jun. 2021, doi: 10.3390/electronics10111348.

[8] O. A. Hanshal, O. N. Ucan, and Y. K. Sanjalawe, "Hybrid deep learning model for automatic fake news detection," *Appl. Nanosci.*, vol. 13, no. 4, pp. 2957–2967, Apr. 2023, doi: 10.1007/s13204-021-02330-4.

[9] S. Raza and C. Ding, "Fake news detection based on news content and social contexts: a transformer-based approach," *Int. J. Data Sci. Anal.*, vol. 13, no. 4, pp. 335–362, 2022, doi: 10.1007/s41060-021-00302-z.

[10] W. S. Paka, "Combining exogenous and endogenous signals with a semi-supervised co-attention network for early detection of COVID-19 fake tweets".

[11] J. George, S. M. Skariah, and T. A. Xavier, "Role of Contextual Features in Fake News Detection: A Review," in *2020 International Conference on Innovative Trends in Information Technology (ICITIIT)*, Feb. 2020, vol. 10, no. 2, pp. 1–6. doi: 10.1109/ICITIIT49094.2020.9071524.

[12] K. Sharma, F. Qian, H. Jiang, N. Ruchansky, M. Zhang, and Y. Liu, "Combating Fake News," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 3, pp. 1–42, May 2019, doi: 10.1145/3305260.