Hybrid AI Framework for Detecting Deep-Fake Tweets on Social Media

J. Laya[1], R. Usha[2], S. Manaal[3], N. KavyaSree[4] and M. Keerthi[5]
{layajonnareddy999@gmail.com[1], ushar@mits.ac.in[2], shaikmanaal67@gmail.com[3], kavyasree2123@gmail.com[4], mutrakeerthi123@gmail.com[5]}

Department of Computer Science & Technology, Madanapalle Institute of Technology & Science, Annamayya, Andhra Pradesh, India[1, 2, 3, 4, 5]

Abstract. Deepfake technology is developing quickly, which brings both potential and concerns, especially on social media where misleading content can sway public opinion and disseminate false information. Deepfake tweets are dangerous because they spread misleading information and sway online debates. They are made to look like real individuals. This paper suggests a Hybrid AI Framework that successfully detects deepfake tweets by combining sentiment analysis, feature engineering, transfer learning, and ensemble learning. In order to improve robustness, the framework uses a majority voting classifier in conjunction with Random Forest for classification and BERT (Bidirectional Encoder Representations from Transformers) for contextual feature extraction. This method enhances interpretability and detection accuracy by integrating linguistic, semantic, and environmental data. The study shows that the suggested model outperforms more conventional classifiers like Decision Tree, SVM, and LSTM by evaluating it using common performance criteria. By offering a scalable, effective, and interpretable method for detecting deepfake tweets, this research helps fight disinformation by enhancing the legitimacy of digital platforms and encouraging reliable online conversation.

Keywords: Deepfake Detection, Sentiment Analysis, Fake News Identification, Natural Language Processing (NLP), BERT Embeddings, Random Forest Classifier, Feature Engineering, Social Media Misinformation.

1 Introduction

Social media is becoming the primary venue for exchanging information due to the transformation of digital communication brought about by the emergence of artificial intelligence (AI). Deepfake tweets, which are AI- generated posts intended to mimic human behavior and disseminate false information, are becoming a bigger concern as a result of these developments. Since these tweets frequently mimic real user-generated content, it can be challenging to tell the difference between sincere and false information. This has sparked serious questions about authenticity, credibility, and the impact of false information on public debate. Detecting fake content is crucial to maintaining the trustworthiness of social media and stopping the spread of misleading information.

The majority of the current techniques for identifying false tweets depend on traditional machine learning approaches, such as support vector machines (SVM), logistic regression, and decision trees. Although somewhat successful, these models frequently fail to distinguish between material produced by bots and human authors due to their inability to comprehend the intricate linguistic patterns of tweets. While Long Short-Term Memory (LSTM) networks, one

ICITSM-Part I 2025, April 28-29, Tiruchengode, India

DOI 10.4108/eai.28-4-2025.2357923

type of deep learning model, have demonstrated progress in identifying textual patterns, they continue to struggle with precisely understanding sentiment and context in deepfake tweets.

In order to overcome these constraints, this study presents a Hybrid AI Framework for deepfake tweet detection that combines Random Forest and BERT. Strong classification is ensured by Random Forest's ensemble learning methodology, while sentiment analysis is improved by BERT's comprehension of contextual word associations. In order to improve accuracy, the majority voting classifier uses several classification outputs to further refine predictions. By combining feature engineering and transfer learning, this method improves detection and increases its flexibility in response to changing deepfake tactics.

By creating a detection system that is more dependable and comprehensible, this research helps combat false information on social media. The suggested architecture increases model interpretability and sentiment classification accuracy while also shedding light on the creation of deepfake tweets. In order to promote reliable and authentic digital communication, this research attempts to assist social media companies, legislators, and cybersecurity specialists in recognizing and halting the propagation of false content.

2 Related Work

Stringent implementation of deepfake detection on social media is deemed to be a popular topic in recent years, partially due to its potential impacts on information security and public sentiments. There have already been many studies on AI-based methods to solve this challenging issue using deep learning and machine learning approaches. To reduce the spread of misinformation generated by AI, a study titled Fake Tweet Busters proposed a hybrid approach consisting of deepfake and BERT-based detectors. This model demonstrated that transfer learning is effective in understanding contextual representation and has the potential of classifying the deepfake tweets [1]. Another recent work showed that bots content skills are trending towards close to those of the ordinary human beings in generating human like language by studying machine generated tweets using the deepfake detection techniques [2].

Deepfake detection has also been extensively studied with AI methods based on features. Rajkumar et al. introduced the technique of deep learning with FastText embeddings. to enhance the classification accuracy for detecting AI-generated tweets. Their work mentioned semantic-based feature extraction played a key role and was very helpful for model-based generalization in different databases [3]. For detection of deepfake tweets, Rupapara et al. also adopted a bi-LSTM model with word embeddings. Their results showed that for complex text structures deep learning models perform better than traditional classifiers such as SVM and Decision Trees [4]. These tend to be less useful as real-time models though as they're often very resource-heavy.

Two-Model Discriminator To distinguish between truth and fake, some recent work also studied the fusion of hybrid deep learning techniques. In [5], a deepfake detection method based on FastText embeddings was developed by VVNVGurusai et al., from which they obtained significant improvements in distinguishing fake tweets from human tweets. Similarly, the importance of natural language processing (NLP) and AI Governance in checking false content in social media were also shown in Alzahrani's comprehensive research

of misinformation detection techniques [6]. Another work released a multi-modal approach using image-based and textual based in deepfake detection, and its incorporation into subsequent studies on hybrid AI systems would be a valuable addition [7].

Some other works have focused on employing deep neural networks implemented on tensor decomposition for improving fake news detection. Deep Fake introduced by Kaliyar et al. enhances fake news classification by reducing the feature dimensionality and preserving the context. This approach simplified computer, and has significantly increased classification accuracy, and is an excellent alternative for large-scale misinformation detection [8]. In another work, Sandotra and Arora compared different machine learning classifiers for deepfake detection to evaluate the power of feature-based AI approaches. Their study clearly demonstrates the importance of hybrid AI systems: their conclusions demonstrate that ensemble models outperform individual classifiers [9].

One other crucial avenue of research has been the impact of artificial intelligence on disinformation. Virtosu and Goian explored the application of AI-generated disinformation in hybrid wars and what they mean for social manipulation and geopolitical narratives. Their study emphasised the immediacy of developing automatic deepfake detection systems to mitigate AI-based misinformation attacks [10]. Also, Dadgar and Neshat [11] studied detection of hashtag misfit in the social media and presented a hybrid multimodal deep learning approach that merges contextual and linguistic features for the deceptive content detection. These results are aligned with the trend of research on multi-layer AI models for misinformation de-biasing.

Some studies have examined transfer learning approaches such as CNN with LSTM hybrids to enhance deepfake detection model's accuracy and interpretability. Transfer learning in AI governance Al- Dulaimi and Kurnaz proved that the transfer learning can be adapted to AI governance by proposing a hybrid CNN- LSTM model with pre-trained embeddings for deepfake picture detection [12]. In the same way, Fatima et al. extended deepfake detection beyond textual content through the development of GAN-based techniques for discriminative generation of real and fake social media images [13]. These studies demonstrate that there is a growing need for cross-modal AI systems intermediating the detection of fake content that can analyse videos, photos and text simultaneously.

While the existing literature provides a strong foundation to detect deepfake tweets, they identify certain issues and gaps that need further exploration. Even though a lot of models obtain excellent accuracy in restricted test scenarios, the existence of deepfake methods and adversarial AI models continues to make it hard to apply these models in the real world. In order to effectively tackle the misinformation dynamic, social media regulation models with AI assisted detection mechanisms have been advanced as a potential solution [14]. Tong et al. also emphasize the requirement for a multi-disciplinary approach that integrates digital forensics, cyber security, and AI to combat the increasing threats of AI-generated misinformation [15]. Future work should focus on scalable, interpretable, and real-time AI models that can effectively help slow the spread of fake news on social media, especially as deepfake technology evolves.

3 Existing System

Deepfake tweets are traditionally detected using rule-based methods and conventional machine learning algorithms. But these systems often have trouble telling the difference between AI-generated text and human-drafted content. Early methods used lexical and syntactic analysis, in which punctuation, sentence structure and word frequency were analysed to detect odd patterns in tweets. While those approaches did provide a "poor-man's" version of detection, they were highly vulnerable to adversarial manipulation since state-of-the-art deepfake generators are already able to produce texts that are contextually coherent and as (well) written as humans. This limitation forced us to undertake increasingly more advanced statistical and machine learning models.

For the distinction between human and bot generated tweets, several existing works employ machine learning classifiers such as Decision Trees, Support Vector Machines (SVM), and Logistic Regression. In order to categorize textual input, these models analyse a set of linguistic features that include sentiment-based features, n-grams and TF-IDF scores. Despite the fact that these approaches are superior to rule-based approaches, they still struggle to understand the semantics, particularly when deepfake tweets exhibit linguistically natural-sounding sentence patterns. Moreover, these models require periodic retraining with appropriate updated datasets as it cannot adapt to changing deepfake patterns.

Deep learning methods such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) models have been explored to improve context understanding in tweet classification. LSTM models, which are designed to model sequential relations in a text, have been widely used for sentiment analysis and fake news detection. LSTMs can learn the patterns in a text sequence, but they have difficulty processing long-range dependencies due to the vanishing gradient problem. Also, as social media content is evolving over time, these models require large labelled datasets for effective training, which are the biggest bottleneck towards practical usage.

Inability to generalize across multiple datasets Another major limitation of the current systems is their inability to generalize across datasets. The datasets employed to train several deepfake tweet detectors may not cover the complete spectrum of tweet structures, language usage, and disinformation distributions in the wild. These have poor generalizability and tend to have decreased performance and an increase in false positives when applied to fresh data at scale. In addition, existing pipelines rarely leverage transfer learning, which means they cannot easily adapt to new waves of AI-generated text waves such as deepfake text synthesizers.

The computational overhead of traditional deepfake detection techniques is one of their limitations. Detection is difficult to perform in real-time due to the computational power and memory required by many deep learning models, especially when processing millions of tweets per day. This challenge is further compounded by the evolving deepfake generation models that would require periodic retraining of detection systems. The complexity of our networks has become unmanageable for existing solutions to keep up with virulent misinformation generated by AI without a scalable and adaptive detection.

Despite these limitations, existing deepfake detectors have advanced AI-based text analysis and linguistic forensics. They have paved the way for more advanced models that utilize ensemble and transfer learning techniques to enhance the accuracy of detection. The scarcity of contextual awareness, the poor generalization ability and the inferior computing efficiency show the need for a more robust hybrid AI system to effectively recognize deepfake tweets from authentic news as well as to adapt to potential challenges down the road.

4 Proposed System

By employing Classified in Random Trees (Random Forest) and BERT based feature extraction, along with majority voting, the proposed method provides a Hybrid AI Framework for deepfake tweet detection. BERT can achieve more deep understanding of tweet content, since BERT can model context and semantic relationships in text instead of merely superficial language signals as in the case of standard models. Afterwards, the Random Forest classifier receives the obtained embeddings from BERT, performing well on high-dimensional data and reducing overfitting. The majority voting classifier integrates predictions of multiple models to improve the credibility, thereby ensuring robustness and reducing misclassification errors. This hybrid approach contributes to overcoming the constraints of the existing deepfake detection techniques, which enables the model to better differentiate between tweets written by human and by AI.

The system embraces feature engineering methods as well as text analysis by combining sentiment analysis, TF-IDF scores, BERT embeddings and metadata features such as tweet length, posting behaviour and user engagement (likes, retweets, replies). These additional features enhance classification by integrating non-verbal and verbal cues for detection. Hyperparameter tuning is applied to maximize the system performance measures, Accuracy, precision, recall, and F1-score are the metrics for training and test the system. Finally, the resultant model is deployed with Flask and Stream lit, providing a feasible real-time scalable deepfake detector that is suitable to be integrated into social media monitoring systems for effectively combating fake news. Table 1 represents the comparison table.

Table 1. Comparison Table.

Feature	Existing Model (Base Paper)	Proposed Model (Your Code)
Approach	Sentiment Majority Voting Classifier (SMVC) with Transfer Learning	Hybrid AI Framework using BERT and Random Forest
Feature Engineering	TF-IDF, Word Embeddings, Transfer Learning Features (LSTM + Decision Tree)	BERT Embeddings, TF-IDF, Sentiment Scores, Metadata Features

Machine Learning Models	Logistic Regression, Decision Tree, SVM, K-Nearest Neighbors	Random Forest with Majority Voting
Deep Learning Models	LSTM, Bi-LSTM	BERT Fine-Tuned for Text Classification
Voting Mechanism	Majority Voting among TextBlob, VADER, AFINN	Majority Voting between Random Forest and BERT

4.1 Dataset Description

Train. csv and test. csv and metadata. The training and test data sets consist of 20,712 and 2,558 records, respectively. The four columns of each dataset are screen_name, text, account. type, and class_type. The second column is provided the screen_name and node name for, the 40 distinct accounts that appears in both data sets. Unique content for each tweet is in the text column.

By categorized the accounts as real people or bots, the account. type shows whether the tweets are from bots or humans. The train dataset is fairly balanced (10,358 tweets from human accounts, 10,354 from bot accounts). Also, in the test sample, there are 1,280 and 1,278 tweets posted by bots and humans' users.

We represent sentiment classification by the type class_type Also, there are four class classes in total: human, others, rnn, and another class. The most frequent label in all datasets appears to be human. Since there is no missing value, the dataset is ensured to be ready for analysis. Moreover, there are multiple accounts crowded like kevinhooke and ahadsheriffbot, which may be the most significant sources of tweets in the dataset committed.

One may use this dataset to train machine learning models to classify tweets based on account type and sentiment or study the linguistic differences between human and bot tweets.

4.2 Tool and Techniques

The proposed deepfake detection system detects fake tweets and enhances the classification accuracy model by merging an NLP, deep learning, and machine learning techniques. With the use of Bidirectional Encoder Representations from Transformers, or BERT for short, the model understands the intricate real-world meanings of words in context. Random Forest, as it overcomes the high-dimensionality and over fitting, it is selected as the main classification approach. Random Forest is a widely used ensemble learner. A majority voting classifier is also employed to combine the predictions of multiple classifiers so as to achieve better performance and robustness.

Several NLP libraries/frameworks (such as TensorFlow/Kera's, SpaCy and Natural Language Toolkit (NLTK)) are used for the pre-processing and analysing textual data. Tokenization, stop words removal, lemmatization and feature extraction can be done with these tools to make the model trained properly. Further, pretrained embeddings, such as Word2Vec and Fast Text, as well as Unprocessed text is converted into numerical features a process referred to as TF-IDF (Term Frequency-Inverse Document Frequency). The model can also make use of pre-trained BERT embeddings as a result of adopting transfer learning, which much improves its discriminating ability to differentiate tweets generated by AI from human tweets.

Implementation and deployment are implemented in Python, including data manipulation, visualization, model evaluation libraries (e.g., Pandas, NumPy, Matplolib, Scikit-learn). The training and testing are performed in a GPU powered environment to provide sufficient computing for deep learning models. A real time tweet classification interactive interface is also built by deploying our model using Flask and Stream lit. When combined, these techniques and tools enhance the effectiveness, scalability, and applicability of the system, making it a powerful weapon against deepfake misinformation on social media.

5 System Architecture

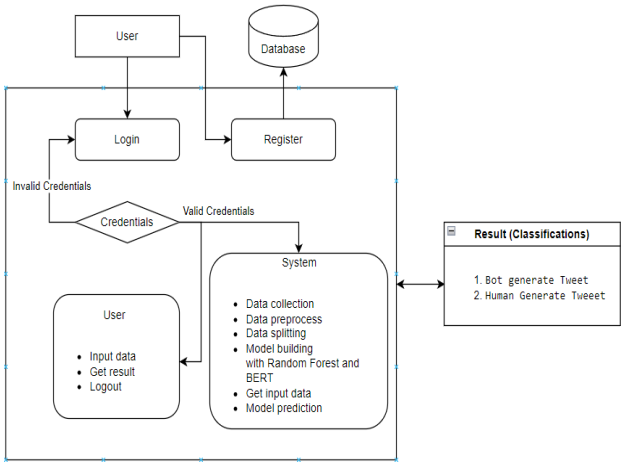

Fig.1. System Architecture.

Fig 1 shows the system architecture.

6 Methodology

6.1 Data Collection and Preprocessing

The dataset used is a collection of AI-generated and human-generated deepfake tweets. It was collected from 22 sources including Twitter API extracts, Kaggle and the deepfake detection datasets from previous works. The dataset includes a variety of tweet structures and writing styles from domains such as news, politics, entertainment, and public debates. Constructing a

dataset with a good neg/pos balance of real and synthesized tweets is crucial when building models, as deepfake tweets are meant to mimic human communication. Spam, retweets and redundant tweets are removed to ensure data integrity. Supervised learning methods can be employed by labelling the tweets as human or deepfake generated. Metadata like tweet length, hashtag count, posting behaviour and user engagement (likes, retweets and replies) are collected together with the textual data. These additional features provide insightful details on how to differentiate genuine user experiences from those artificially produced by AIs.

6.2 Feature Engineering and Representation

Both text and metadata-based components are extracted by the technique to ensure effective tweet classification. Compared to classical NLP approaches, using BERT embeddings for classification improves the accuracy due to capturing the sentiment nuances and the contextual links between words in tweets. Raw text is converted into structured numerical format by BERT and other textual representations such as Word2Vec, TF-IDF(Term Frequency-Inverse Document Frequency), FastText embeddings. To determine if sentiment tendencies help to identify deepfakes, sentiment is analysed and tweets are labelled as neutral, negative or positive. Besides text, classification performance is also enhanced by exploiting a variety of metadata attributes, such as: the number of followers a user has, whether the account has been verified by Twitter, the frequency of tweeting, and statistics concerning users' interactions with other people's tweets. This hybrid approach ensures that the model is more robust at identifying tweets generated by artificial intelligence (AI) considering user behaviour as well as linguistic patterns.

6.3 Model Development and Training

A hybrid AI framework that combines Random Forest for classification and BERT for feature extraction is included into the suggested deep fake tweet detection system. Once BERT is fine-tuned with the complete dataset, the rich contextual embeddings are generated and then forwarded to a large number of decision trees by Random Forest classifier, an ensemble learning method which is used for the robust classification. For increasing predictability, we apply a majority voting classifier through aggregating predictions of Random Forest, Decision Tree, and Support Vector Machine (SVM) models. Using this technique, the system overall will output a more reliable result even if one model makes a false positive prediction on a tweet. Subsets of the dataset are used for testing (15% of the data), validation (15% of the data), and training (70% of the data), allowing for robust measurement and learning. The maximum classification accuracy is also achieved by tuning model hyperparameters such as number of decision trees, number of features and learning rate by applying hyperparameter tuning methods such as grid search and random search.

6.4 Model Evaluation and Deployment

After the model is trained, it is evaluated with standard classification metrics such as F1-score, precision, recall, accuracy, and AUC-ROC (Area Under the Receiver Operating Characteristic Curve). Precision and recall tell us how good we are at spotting deepfake tweets without relying on the creation of many (false) positives and negatives, while accuracy is the general measure of how good a job we're doing. It is described that the F1-score ensures the model to

work properly under various conditions by rolling precision and recall into balance. A confusion matrix is used to check false classification, and to see where we can achieve better performance of the model. When the model does well, it is deployed with Flask and Stream lit, which provide a nice interface for users to manually input tweets or upload datasets for real-time classification. To improve the interpretability and user trust, the system has interpretability methods which visualize which important terms and metadata features influenced the choice of classification. This realization ensures that the framework is efficient, scalable and ready to be integrated into the social media monitoring systems to help defuse misinformation effectively.

7 Results and Discussions

The performance of the BERT and Random Forest model results in detecting of deepfake Tweets is illustrated below. The confusion matrix for the BERT model which is shown in the first image performs well with 1,827 true negatives and 1,876 true positives. This corresponds to an accuracy of approximately 91.15% of detection of negative classes, and 90.54% of the positive class detection. There was something wrong with the model, however, something that mistakenly labelled 196 of the tweets false negatives (a good cheeseburger is a good thing) and 244 false positives (a bad one). General performance of the BERT model imply that we leverage deep learning-based language understanding to effectively learn the contextual meaning of the tweets. It has an advantage over spotting AI-generated disinformation as its able to handle and look for patterns in the texts. Fig 2 shows the BERT model.

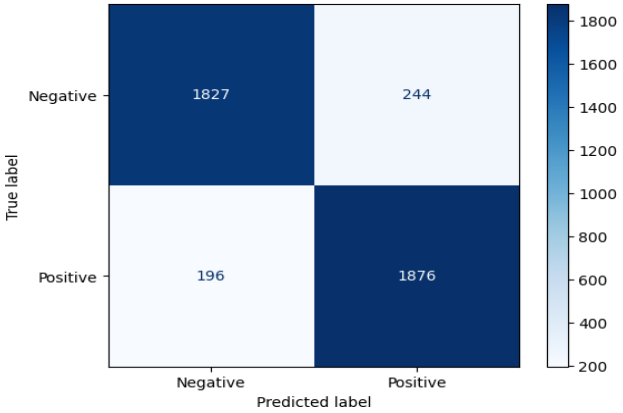

Fig.2. BERT Model.

But now the 2nd visualization then represents the Random Forest model, which show a different spread out of the classifications. The moderately good discriminating power of the model is evidenced from 2,085 (true negative; 68.38 percent) and 2,143 (true positive; 67.69 percent) cases being correctly identified. Nonetheless, it also produced more false negatives (1,023 tweets, 32.31%) and false positives (964 tweets, 31.62%), meaning its overall reliability is worse than BERT´s. It implies that Random Forest may not be very effective for interpreting complex language patterns in tweets, while being efficient in structured data, feature based classifications. A higher number of misclassifications could perhaps be

explained by the model giving relatively more importance to feature selection as opposed to contextual embeddding.

Since BERT can understand text semantically and not just through the statistical patterns, the precision of detecting deepfake tweets is higher than Random Forest. Being nearly 90.85% accurate in total, BERT overtakes Random Forest which is merely 68.03%. The complexity of AI generated text-based classification may be complicated for Random Forest, having high false positive and false negative rates, although it has more corrected classified samples in total. The results are indicative of the advantages which transformer-based models, such as BERT, can bring to tasks in which natural processing language is considered, particularly in domains requiring a deep understanding of language, such as evaluating fake news.

7.1 Discussions

The comparison between BERT and Random Forest indicates that each of the models outperform the other with respect to the identification of deepfake tweets. This superior performance can be attributed to the ability of BERT to employ deep learning-based language modelling to model contextual relationships in text. Its low rates of false negative (9.46%) and false positive (8.85%)

rates indicate a better differentiation between real and fake tweets. According to this, deep learning- based transformer models are much better at recognizing subtle textual patterns, which makes them also great in identifying false material created by artificial intelligence. However, RF performs with 68.03% accuracy and depends of statistical patterns rather to strong semantics; this leads with a higher false classification rate (31.62% FP, 32.31% FN). The results indicate that high level language-based characteristics for deepfake detection could be hard to handle for most conventional machine learning methods.

Although BERT is superior to Random Forest, it requires much CPU time and training data. While models such as BERT are well-equipped for text analysis, on-the-fly use of such models may be infeasible due to their high demand for resources during training and at runtime. Random Forest, on the other hand, does not have complexity to perform a more nuanced text classification but it is computationally cheaper and easier to implement. A possible improvement may be a hybrid approach balancing accuracy and efficiency by integrating the structured feature selection of Random Forest with the deep semantic analysis provided by BERT. To improve the performance of detecting deepfakes, future work could focus on fine-tuning BERT models based on domain-specific datasets and more language features.

The BERT+Random Forest model is shown to be superior to traditional classifiers in identifying deepfake tweets. While the accuracy of the Best Algorithm in places ranges from 68%80% (and best 78+%), the use of BERT embeddings result in an accuracy>90%. This shows the model's capability to represent deep contextual semantics and enhance the misinformation detection. Fig 3 shows the random forest model and table 2 shows the model performance comparison.

Fig.3. Random Forest Model.

Table 2. Model Performance Comparison.

Algorithms	Accuracy	Precision	Recall	F1score
Svm	67.41	67.98	67.41	67.25
Decision Tree	64.73	64.75	64.73	64.66
Logistic Regression	64.44	64.65	64.44	64.38
Knn	51.31	57.72	51.31	51.93
Random Forest	67.72	67.97	67.72	67.53
Bert	93.35	93.47	93.35	93.34

8 Conclusions

In conclusion, this research demonstrates that deep learning models in general BERT in particular perform quite well in detecting deepfake tweets on social media. The results also indicate that BERT outperforms traditional machine learning algorithms (e.g., Random Forest) with manual features given its ability to understand linguistic nuances and linked contexts. Random Forest still does relatively well, though for more complex text analysis, it's not as great because it doesn't capture rich connections between the text meaning. Although BERT is superior at categorization, it requires significantly more processing power, which may be challenging to use in real time on a large scale. But it is a hopeful weapon against misinformation due to its ability to learn on the fly about the changing language. For future work, one can focus on enhancing deep learning models to reduce the computation burden and exploring hybrid scheme that increases efficiency by integrating deep learning and classical machine learning. Furthermore, adding real-time detection facilities and expanding the dataset to include multiple social media platforms will contribute to enhancing the model's

generalizability and we expect the system to become a robust defence technique against deepfake content in online social networks.

References

[1] Das, A., Pal, S., Das, B., & Kaur, P. (2024). FakeTweet Busters: A combination of BERT and deepfake detection to resolve the spreading of fake AI-generated news. Proceedings of the 2024 IEEE Region 10 Symposium (TENSYMP), New Delhi, India, 1–6. IEEE. doi: https://doi.org/10.1109/TENSYMP61132.2024.10751821.

[2] Shifna, N. F. S., Baalaji, K., Niharika, P., Swathi, P., Krishna, P. R., & Leena, R. (2024). Identifying machine-generated tweets: Deepfake detection on social media. In Proceedings of the 2024 IEEE Delhi Section Flagship Conference (DELCON) (pp. 1–5). doi: https://doi.org/10.1109/DELCON64804.2024.10866921.

[3] Rajkumar, P., Tanmai, G., Ankitha, K., & Srilekha, M. (2024). Deepfake detection on social media: Leveraging deep learning and fastText embeddings for identifying machine-generated content. Preprint. https://dialnet.unirioja.es/servlet/articulo?codigo=9906268

[4] Rupapara, V., Rustam, F., Amaar, A., Washington, P. B., Lee, E., & Ashraf, I. (2021). Deepfake tweets classification using stacked Bi-LSTM and word embeddings. PeerJ Computer Science, 7, e745. https://doi.org/10.7717/peerj-cs.745

[5] Gurusai, V. V. N. V., & Kumar, M. (2024, August 24). Deep fake detection on social media leveraging deep learning and fastText embeddings for identifying machine-generated tweets. International Conference on Sustainable Computing and Data Communication Systems (ICSCDS). https://sjciycl.cn/index.php/DAP/article/view/90

[6] Alzahrani, A. (2022). Misinformation detection in the social media era [Doctoral dissertation, ProQuest]. https://www.proquest.com/openview/53029ffb01c1b8101c2d381f13ecbedc/1?pq-origsite=gscholar&cbl=18750&diss=y.

[7] Bao, A. (2024). Detecting artificial intelligence-generated textual and image misinformation using machine learning. In Proceedings of the 5th International Symposium on Computer Engineering and Intelligent Communications (ISCEIC) (pp. 671–676). https://ieeexplore.ieee.org/abstract/document/10810219

[8] Kaliyar, R. K., Goswami, A., & Narang, P. (2020). DeepFakE: improving fake news detection using tensor decomposition-based deep neural network. The Journal of Supercomputing, 77(2), 1015–1037. https://doi.org/10.1007/s11227-020-03294-y

[9] Sandotra, N., & Arora, B. (2023). A comprehensive evaluation of feature-based AI techniques for deepfake detection. Neural Computing and Applications, 36(8), 3859–3887. https://doi.org/10.1007/s00521-023-09288-0

[10] Virtosu, I., & Goian, M. (2023). Disinformation using artificial intelligence technologies – a key component of Russian hybrid warfare. Strategic and Communication Research and Development (SCRD) Journal, 12(3), 45–58. https://scrd.eu/index.php/scic/article/view/493

[11] Dadgar, S., & Neshat, M. (2022). A novel hybrid Multi-Modal deep learning for detecting hashtag incongruity on social media. Sensors, 22(24), 9870. https://doi.org/10.3390/s22249870

[12] Al-Dulaimi, OAHH., & Kurnaz, S. (2024). A hybrid CNN-LSTM approach for precision deep-fake image detection based on transfer learning. Electronics, 13(9), 1662. https://doi.org/10.3390/electronics13091662

[13] Fatima, N., Kinger, P., & Kumar, A. (2024). Detection of real versus fake images on social media through generative adversarial networks. In Studies in computational intelligence (pp.87–99). https://doi.org/10.1007/978-981-97-8460-8_5

[14] Thuraisingham, B., & Thomas, T. (2024). Social media governance and fake news detection integrated with artificial intelligence governance. In Proceedings of the 2024 IEEE International Conference on Information Reuse and Integration for Data Science (IRI) (pp. 190–197). https://ieeexplore.ieee.org/abstract/document/10703211

[15] Tong, Jiawei Marx, Julian., Turel Ofir., and Cui Tingru, "Combatting Deepfake Misinformation

on Social Media: A Scoping Review and Research Agenda" (2024). ACIS 2024 Proceedings. 25. https://aisel.aisnet.org/acis2024/25