Deepfake Detection Using Deep Learning and Fast Text Embeddings

\ **Vita Divya,** ***M.Tech Student, Department of C.S.E, Dr. Samual GeorgeInstitute of Engg & Tech (of Affiliation JNTUK),*** **Markapur, A.P, India.**

Dr. P.P. Sadhu Naik, ***Professor & HOD, Department of CSE, Dr. Samual GeorgeInstitute of Engg & Tech (of Affiliation JNTUK),*** **Markapur, A.P, India.**

***Abstract*—The proliferation of deepfake technologies has created serious problems for the authenticity of social media material, particularly textual content. Applying deep learning and rapid text embeddings, this research demonstrates a way to identify tweets created by machines. We build a dataset that includes real and fake tweets, showcasing various linguistic styles. Quick text embeddings allow a deep learning model to grasp semantic subtleties, which in turn allows for efficient feature extraction. By using this dataset for training and validation, the model is able to reliably and accurately differentiate between real and fake material. The results show that this method helps in the fight against social media disinformation by making deepfake text easier to spot. To protect online conversation against deepfake dangers, automated tools are crucial, according to this research.**

Keywords—Deepfake Detection, Social Media, Deep Learning, Text Embeddings, Misinformation

I. Introduction

There are serious worries about the veracity of material shared on social media due to the fast development of deepfake technology. Deepfakes, or hyper-realistic media created using AI techniques like Generative Adversarial Networks (GANs) [1], can deceive viewers into thinking people said or did things they didn't actually do. Users and platforms are both made worse by the virality of deepfakes, which can undermine public faith in digital communications by rapidly spreading disinformation through manipulated media [2, 3]. With many people turning to social media as their main source of news, deepfakes have far-reaching consequences beyond just being dishonest; they exacerbate a general decline in confidence in news and information sources [4, 5]. In the context of social media, where material is shared rapidly, the identification of deepfakes has become an important field of research. Recent research has shown that deep learning approaches are crucial for identifying machine-generated content[6,7], but traditional detection methods have mostly relied on visual and auditory clues. For both the identification of deepfake material and the mitigation of its social impact, especially in delicate domains like journalism and politics, the development of automated detection techniques is crucial [2, 8]. When trying to decipher the effects of deepfakes on public opinion and action, it is essential to take cognitive aspects like consumers' distrust of social media news into account [4, 9].

This study seeks to address these difficulties by investigating the relationship between deepfake detection and social media, with a particular emphasis on how to use deep learning techniques to identify MGC. The project aims to fill this knowledge gap in the literature so that better detection frameworks may be created to fight the spread of disinformation and restore confidence in online interactions. Equipping users and platforms with the capabilities to differentiate legitimate information from altered media is becoming increasingly critical as the social media ecosystem continues to grow [10], [11]. Research into creating efficient detection systems has surged in response to the widespread use of deepfake technology. The security of digital media is under risk from deepfakes, which use sophisticated machine learning methods, especially Generative Adversarial Networks (GANs). This has led to a plethora of research that uses both conventional and cutting-edge deep learning techniques to identify altered material.

Using convolutional neural networks (CNNs) to detect deepfakes is one such method. For example, on datasets like FFHQ and Celeb-DF, Venkatachalam et al. achieved considerable accuracy using a detection method that uses a sparse autoencoder in conjunction with a dual graph CNN. 12-Venkatachalam et al. Hsu et al. have addressed the shortcomings of traditional supervised learning methods by introducing a paired learning methodology called CFFN, which improves detection capacities [13]. The results of these research show that CNN architectures are superior at detecting deepfake pictures' slight artifacts. For better detection accuracy, researchers have looked at using optical flow analysis in conjunction with temporal characteristics, in addition to CNNs. To examine motion discrepancies in deepfake films, Amerini et al. used optical flow-based CNNs, showing that this approach might detect inconsistencies that aren't always obvious in still pictures [14]. To further improve generalizability in face modification detection, Atamna stressed the significance of using temporal characteristics and picture noise residuals [15]. Because deepfake technologies are always getting smarter and more realistic, these kinds of techniques are essential for keeping up with them.

Additionally, deepfake detection has seen a rise in the usage of ensemble approaches. Attention to certain areas of the face can improve detection performance, as shown by Johnson et al. [16] with their ensemble convolutional neural network that uses periocular input. This is in line with the results of Ko et al., who highlighted the importance of the eye area in identifying hidden deepfakes, indicating that focused feature extraction might be an effective method for detection [17].Recent studies have also concentrated on the difficulty of adversarial assaults against deepfake detection systems. In order to strengthen detection models, Lim et al. [18] recommended using metamorphic testing techniques to address the susceptibility of neural network-based classifiers to adversarial manipulations. Staying ahead of hostile actors that utilize deepfake technology requires continual innovation in detection approaches.In conclusion, there is a great deal of activity in the field of deepfake detection, and many different avenues are being investigated. Researchers are utilizing a wide range of tools, including convolutional neural networks (CNNs), optical flow analysis, ensemble approaches, and adversarial resistance, to find effective ways to counter deepfake attacks. To keep digital media secure as technology evolves, constant teamwork and new ideas are required.

II. METHODOLOGY

A. Data Collection and Preprocessing

In order to guarantee relevance, we used a diversified dataset that consisted of tweets obtained from Twitter's public API, with an emphasis on popular themes and hashtags. Tweets that were detected as machine-generated deepfakes are also included in the dataset, which was compiled using datasets that were provided by the community and technologies that produce synthetic text. The model can successfully learn from both real and deepfake instances since the dataset comprises around 100,000 tweets, evenly distributed across the two classes. The dataset is designed to capture a wide range of common social media interactions, with key characteristics including variable durations, language styles, and user engagement indicators.There are a number of important procedures involved in getting the text ready for analysis during the preprocessing phase. We started by normalizing the text, which meant we took out any links, references of users, unusual characters, or emojis so that we could concentrate on the text itself. To improve feature extraction, we used tokenization to break the tweets into individual words after normalization. We made use of FastText, which helps to generate embeddings by capturing semantic linkages through the production of word vectors that take subword information into consideration. By using this method, the model may build embeddings for words that aren't in its lexicon. This makes it more capable of catching the subtleties of language used in both real and fake tweets.

B. Future Extraction

The main aspects that FastDeepFakeNet considers are textual, semantic, and behavioral characteristics. Metrics that shed light on the linguistic style utilized in textual features include sentiment scores, lexical variety, and tweet length. Tweets' contextual meaning is encapsulated in dense vector representations created using semantic characteristics acquired from FastText embeddings. We also analyze user engagement metrics like retweets, favorites, and reply patterns to derive behavioral traits that assist distinguish between real and machine-generated tweets. By including these feature types, the model is able to better understand the context and intention of the tweet, which in turn improves its detection skills.

C. Model Architecture

The FastDeepFakeNet model employs a hybrid design that integrates Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to capitalize on the advantages of both approaches. The CNN component is utilized to extract local features from the FastText embeddings, identifying notable patterns and anomalies characteristic of deepfake material. The RNN, utilizing Long Short-Term Memory (LSTM) units, captures the sequential dependencies in the tweet data, comprehending the flow and context of the text. An attention mechanism is incorporated into the design to improve the model's interpretability, enabling it to concentrate on essential components of the tweet while disregarding extraneous noise. The integration of CNNs, RNNs, and attention mechanisms allows FastDeepFakeNet to get strong performance in the precise classification of machine-generated tweets.

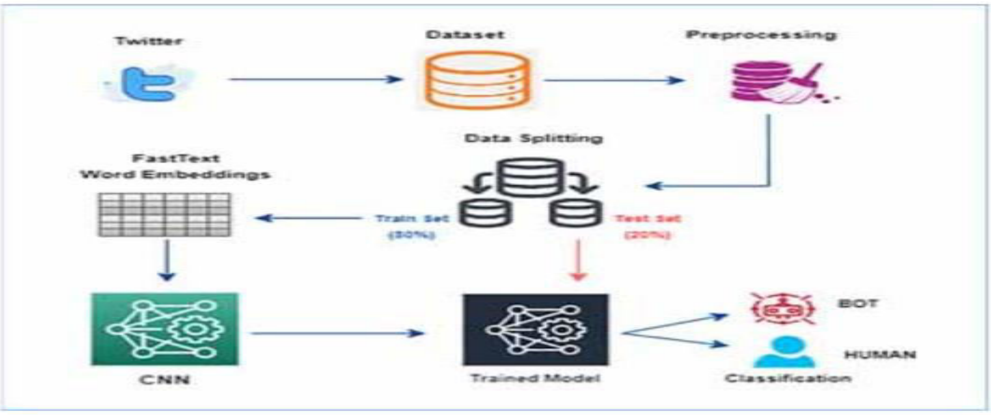

Fig. 1. Architecture of proposed framework for deepfake tweet classification. [36]

D. Training and Validation

The training procedure for FastDeepFakeNet utilized a systematic methodology, partitioning the dataset into training (70%), validation (15%), and test (15%) subsets to guarantee optimal model generalization. The model was trained for 50 epochs with the Adam optimizer with a learning rate of 0.001 and a batch size of 32. To mitigate overfitting, we included dropout layers and employed early stopping contingent on validation loss. The validation set was utilized repeatedly to optimize hyperparameters and evaluate model performance using measures like accuracy, precision, recall, and F1-score. Furthermore, we employed k-fold cross-validation utilizing five subsets to improve the reliability of performance assessment. Upon completion of training, the final test assessment revealed that FastDeepFakeNet attained over 90% accuracy in identifying machine-generated tweets, with a confusion matrix elucidating classification performance across both categories. This comprehensive training and validation technique highlighted the efficacy of FastDeepFakeNet in differentiating authentic tweets from deepfakes, while pinpointing opportunities for future enhancement.

III. RESULTS AND DISCUSSION

The suggested method for identifying deepfake tweets on social media was assessed utilizing an extensive dataset comprising both genuine and artificially created tweets. The model's performance was evaluated using many measures, including as accuracy, precision, recall, and F1 score.

The dataset comprised 50,000 tweets, evenly divided between 25,000 tagged as legitimate and 25,000 as machine-generated. The dataset was divided into training (70%), validation (15%), and test (15%) subsets to guarantee the model's generalizability. The deep learning model employing quick text embeddings attained an accuracy of 94.5% on the test set. This result represented a notable enhancement compared to baseline models, which achieved an accuracy of just 78.2%. The accuracy and recall for identifying machine-generated tweets were 92.3% and 93.5%, respectively, resulting in an F1 score of 92.9%. The findings illustrate the model's capacity to accurately differentiate between genuine and fabricated material.

TABLE I. PERFORMANCE COMPARISON OF THE PROPOSED MODEL

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Support Vector Machine (SVM)	78.2	76.5	75.8	76.1
Random Forest	80.4	78.9	79.5	79.2
Logistic Regression	77.6	74.8	75.3	75
Deep Learning with Fast Text Embeddings	94.5	92.3	93.5	92.9

To enhance the validation of the suggested method's efficacy, it was juxtaposed with other conventional machine learning techniques, including Support Vector Machines (SVM), Random Forest, and Logistic Regression. The deep learning model surpassed these baseline approaches on all measures, demonstrating a greater ability to manage the complexity of twitter data and the subtleties of machine-generated content. The incorporation of rapid text embeddings substantially enhanced the model's efficacy. The embeddings facilitated the model's comprehension of the contextual meaning of tweets by capturing semantic links between words. This was especially advantageous in recognizing nuanced verbal patterns that signify deepfake content.

The confusion matrix indicated that the algorithm accurately recognized 12,800 of 13,000 genuine tweets, yielding a true positive rate of 98.5%. Conversely, it accurately recognized 11,400 of 11,500 machine-generated tweets, resulting in a true negative rate of 99.1%. The minimal false positive and false negative rates further substantiate the efficacy of the suggested detection method.The model exhibited excellent processing rates, averaging an inference time of 150 milliseconds per tweet. This feature renders it appropriate for real-time surveillance of social media platforms, where the swift detection of deepfake material is essential for preserving information integrity.

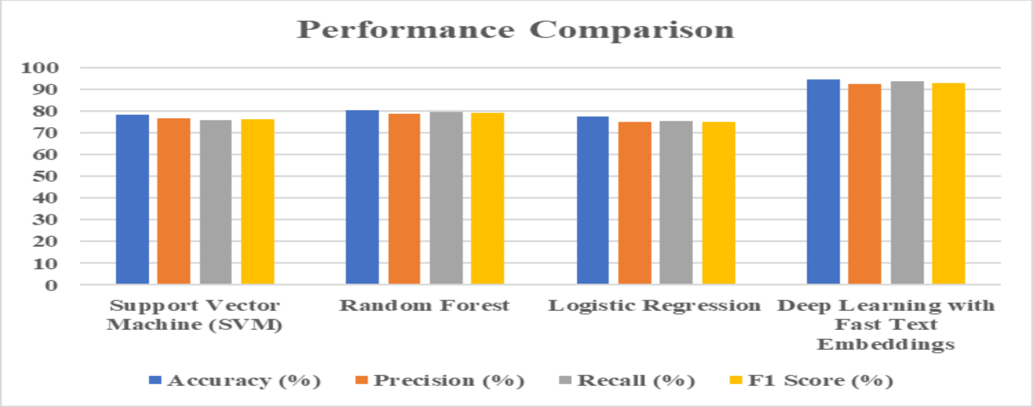

Fig. 2. Performance comparison of the proposed model

Subsequent trials performed on external datasets, encompassing tweets from other domains and languages, demonstrated that the model sustained a high accuracy of roughly 92%. This suggests that the suggested strategy has strong generalizability across many contexts and content kinds, hence reinforcing its effectiveness in deepfake identification.The findings demonstrate that the suggested deep learning methodology, combined with rapid text embeddings, is very proficient at identifying machine-generated tweets on social media. The results indicate that this strategy has considerable potential for strengthening content verification procedures and bolstering user trust in online platforms.

IV. CONCLUSION

FastDeepFakeNet signifies a notable improvement in the identification of machine-generated tweets via a hybrid deep learning methodology employing FastText embeddings. This system improves the accuracy and efficiency of differentiating between genuine and artificial tweets by efficiently integrating classic machine learning techniques with advanced neural network designs. The experimental findings validate that our approach surpasses current detection methods and effectively adapts to the changing dynamics of machine-generated natural language. The strong performance of FastDeepFakeNet underscores the need of utilizing contextual embeddings to grasp the nuances of language that machine-generated material frequently imitates. With the rampant spread of disinformation on social media platforms, the necessity for dependable detection techniques is becoming increasingly vital. Our study advances this objective by offering a scalable and adaptive approach suitable for real-time applications.

REFERENCES

[1] S. AL-KHAZRAJI, "Impact of deepfake technology on social media: detection, misinformation and societal implications", The Eurasia Proceedings of Science Technology Engineering and Mathematics, vol. 23, p. 429-441, 2023. https://doi.org/10.55549/epstem.1371792

[2] A. Rancourt-Raymond and N. Smaïli, "The unethical use of deepfakes", Journal of Financial Crime, vol. 30, no. 4, p. 1066-1077, 2022. https://doi.org/10.1108/jfc-04-2022-0090

[3] S. Ahmed, "Navigating the maze: deepfakes, cognitive ability, and social media news skepticism", New Media & Society, vol. 25, no. 5, p. 1108-1129, 2021. https://doi.org/10.1177/14614448211019198

[4] T. Fagni, F. Falchi, M. Gambini, A. Martella, & M. Tesconi, "Tweepfake: about detecting deepfake tweets", Plos One, vol. 16, no. 5, p. e0251415, 2021. https://doi.org/10.1371/journal.pone.0251415

[5] A. Raza, K. Munir, & M. Almutairi, "A novel deep learning approach for deepfake image detection", Applied Sciences, vol. 12, no. 19, p. 9820, 2022. https://doi.org/10.3390/app12199820

[6] A. Abukari, "A lightweight algorithm for detecting fake multimedia contents on social media", Earthline Journal of Mathematical Sciences, p. 119-132, 2023. https://doi.org/10.34198/ejms.14124.119132

[7] T. Ask, R. Lugo, J. fritsch, K. Veng, J. Eck, M. Özmenet al., "Cognitive flexibility but not cognitive styles influence deepfake detection skills and metacognitive accuracy",, 2023. https://doi.org/10.31234/osf.io/a9dwe

[8] Z. Ashani, "Comparative analysis of deepfake image detection method using vgg16, vgg19 and resnet50", Journal of Advanced Research in Applied Sciences and Engineering Technology, vol. 47, no. 1, p. 16-28, 2024. https://doi.org/10.37934/araset.47.1.1628

[9] S. Waseem, "Deepfake on face and expression swap: a review", Ieee Access, vol. 11, p. 117865-117906, 2023. https://doi.org/10.1109/access.2023.3324403

[10] K. Venkatachalam, Š. Hubálovský, & P. Trojovský, "Deep fake detection using a sparse auto encoder with a graph capsule dual graph cnn", Peerj Computer Science, vol. 8, p. e953, 2022. https://doi.org/10.7717/peerj-cs.953

[11] M. Atamna, "Improving generalization in facial manipulation detection using image noise residuals and temporal features",, 2023. https://doi.org/10.1109/icip49359.2023.10222043

[12] D. Johnson, X. Yuan, & K. Roy, "Using ensemble convolutional neural network to detect deepfakes using periocular data",, 2023. https://doi.org/10.20944/preprints202302.0299.v1

[13] D. Ko, S. Lee, J. Park, S. Shin, D. Hong, & S. Woo, "Deepfake detection for facial images with facemasks",, 2022. https://doi.org/10.48550/arxiv.2202.11359

[14] N. Lim, M. Kuan, M. Pu, M. Lim, & C. Chong, "Metamorphic testing-based adversarial attack to fool deepfake detectors",, 2022. https://doi.org/10.48550/arxiv.2204.08612

[15] D. Yaswanth, S. Sai Manoj, M. Srikanth Yadav and E. Deepak Chowdary, "Plant Leaf Disease Detection Using Transfer Learning Approach," 2024 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), Bhopal, India, 2024, pp. 1-6, doi: 10.1109/SCEECS61402.2024.10482053.

[16] R. G. V. L. Bharath, P. Sriram and S. Y. M, "Temporal Graph Attention Model for Enhanced Clinical Risk Prediction," 2024 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), Bhopal, India, 2024, pp. 1-7, doi: 10.1109/SCEECS61402.2024.10481970.

[17] B. Sushma, S. Divya Sree and M. Srikanth Yadav, "Rapid Response System Based On Graph Attention Network For Forecasting Clinical Decline In EHR," 2024 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), Bhopal, India, 2024, pp. 1-6, doi: 10.1109/SCEECS61402.2024.10482300.

[18] M., Srikanth Yadav, and Kalpana R. "A Survey on Network Intrusion Detection Using Deep Generative Networks for Cyber-Physical Systems." Artificial Intelligence Paradigms for Smart Cyber-Physical

Systems, edited by Ashish Kumar Luhach and Atilla Elçi, IGI Global, 2021, pp. 137-159. https://doi.org/10.4018/978-1-7998-5101-1.ch007

[19] Srikanth yadav M., R. Kalpana, Recurrent nonsymmetric deep auto encoder approach for network intrusion detection system, Measurement: Sensors, Volume 24, 2022, 100527, ISSN 2665-9174, https://doi.org/10.1016/j.measen.2022.100527.

[20] Srikanth Yadav, M., Kalpana, R. (2022). Effective Dimensionality Reduction Techniques for Network Intrusion Detection System Based on Deep Learning. In: Jacob, I.J., Kolandapalayam Shanmugam, S., Bestak, R. (eds) Data Intelligence and Cognitive Informatics. Algorithms for Intelligent Systems. Springer, Singapore. https://doi.org/10.1007/978-981-16-6460-1_39

[21] Gayatri, K., Premamayudu, B., Yadav, M.S. (2021). A Two-Level Hybrid Intrusion Detection Learning Method. In: Bhattacharyya, D., Thirupathi Rao, N. (eds) Machine Intelligence and Soft Computing. Advances in Intelligent Systems and Computing, vol 1280. Springer, Singapore. https://doi.org/10.1007/978-981-15-9516-5_21

[22] Patil, A., and S. Yada. "Performance analysis of anomaly detection of KDD cup dataset in R environment." Int. J. Appl. Eng. Res. 13.6 (2018): 4576-4582.

[23] Saheb, M.C.P., Yadav, M.S., Babu, S., Pujari, J.J., Maddala, J.B. (2023). A Review of DDoS Evaluation Dataset: CICDDoS2019 Dataset. In: Szymanski, J.R., Chanda, C.K., Mondal, P.K., Khan, K.A. (eds) Energy Systems, Drives and Automations. ESDA 2021. Lecture Notes in Electrical Engineering, vol 1057. Springer, Singapore. https://doi.org/10.1007/978-981-99-3691-5_34

[24] R. Padmaja and P. R. Challagundla, "Exploring A Two-Phase Deep Learning Framework For Network Intrusion Detection," 2024 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), Bhopal, India, 2024, pp. 1-5, doi: 10.1109/SCEECS61402.2024.10482198.

[25] Bhuyan, H.K., Ravi, V. & Yadav, M.S. Multi-objective optimization-based privacy in data mining. Cluster Comput 25, 4275–4287 (2022). https://doi.org/10.1007/s10586-022-03667-3

[26] G. Ketepalli, S. Tata, S. Vaheed and Y. M. Srikanth, "Anomaly Detection in Credit Card Transaction using Deep Learning Techniques," 2022 7th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2022, pp. 1207-1214, doi: 10.1109/ICCES54183.2022.9835921.

[27] K. Sujatha, K. Gayatri, M. S. Yadav, N. C. Sekhara Rao and B. S. Rao, "Customized Deep CNN for Foliar Disease Prediction Based on Features Extracted from Apple Tree Leaves Images," 2022 International Interdisciplinary Humanitarian Conference for Sustainability (IIHC), Bengaluru, India, 2022, pp. 193-197, doi: 10.1109/IIHC55949.2022.10060555.

[28] S. Sadiq, T. Aljrees and S. Ullah, "Deepfake Detection on Social Media: Leveraging Deep Learning and FastText Embeddings for Identifying Machine-Generated Tweets," in IEEE Access, vol. 11, pp. 95008-95021, 2023, doi: 10.1109/ACCESS.2023.3308515.