Journal for Educators, Teachers and Trainers
The PROFESIOLab electronic, peer-reviewed, open-access Magazine

ISSN 1989-9572

DOI:10.47750/jett.2024.15.05.27

# Deepfake Detection on Social Media: Leveraging Deep Learning and Fast text Embeddings for Identifying Machine-Generated

**Rajkumar P1, G. Tanmai2, K. Ankitha2, M. Srilekha2**

**Journal for Educators, Teachers and Trainers, Vol.15(5)**

https://jett.labosfor.com/

**Date of Reception: 24 Oct 2024**

**Date of Revision: 20 Nov 2024**

**Date of Publication : 31 Dec 2024**

**Rajkumar P1, G. Tanmai2, K. Ankitha2, M. Srilekha2 (2024). Deepfake Detection on Social Media: Leveraging Deep Learning and Fast text Embeddings for Identifying Machine-Generated,Vol.15(5).277-285**


Journal for Educators, Teachers and Trainers, Vol. 15(5)

ISSN1989 –9572

https://jett.labosfor.com/

# Deepfake Detection on Social Media: Leveraging Deep Learning and Fast text Embeddings for Identifying Machine-Generated

Rajkumar P[1], G. Tanmai[2], K. Ankitha[2], M. Srilekha[2]

[1]Assistant Professor, [2]UG Student, [1,2]Department of Information Technology

[1,2]Malla Reddy Engineering College for Women (UGC-Autonomous), Maisammaguda, Hyderabad, 500100, Telangana.

## ABSTRACT

Deepfake technology, which uses AI to create manipulated media, poses a significant threat to information integrity on social media platforms. In India, the rise of deepfake content has grown exponentially, especially in the political and entertainment domains, where fake news and AI-generated videos have gone viral, leading to misinformation. The primary objective is to develop a robust AI model that accurately detects deepfake content on social media platforms, focusing on identifying machine-generated tweets using FastText embeddings. Traditional methods involved human moderation, fact-checking agencies, and manual filtering of social media posts based on predefined rules and keyword matching. These methods were time-consuming and often inaccurate, lacking the scalability to manage the massive volume of online content. The manual detection of deepfakes and AI-generated content is highly inefficient, prone to errors, and incapable of handling the vast volume of social media data in real time. As a result, harmful and misleading information can spread widely before being identified or removed. With the growing influence of social media in shaping public opinion, the motivation behind this research is to combat misinformation and safeguard the integrity of online discourse. Particularly deep learning models can significantly improve the detection of deepfakes by automating the analysis of social media content. FastText embeddings will convert tweets into meaningful word vectors, while deep learning models can be applied to classify whether a tweet is human-generated or AI-generated. This approach offers real-time detection, improved accuracy, and scalability compared to traditional methods.

**Keywords**: Deepfake Technology, Social Media Platforms, Fast Text embeddings, Fact-Checking Agencies

## 1. INTRODUCTION

Deepfake technology employs advanced artificial intelligence algorithms to create hyper-realistic media, leading to serious concerns regarding misinformation, especially on social media platforms. In India,

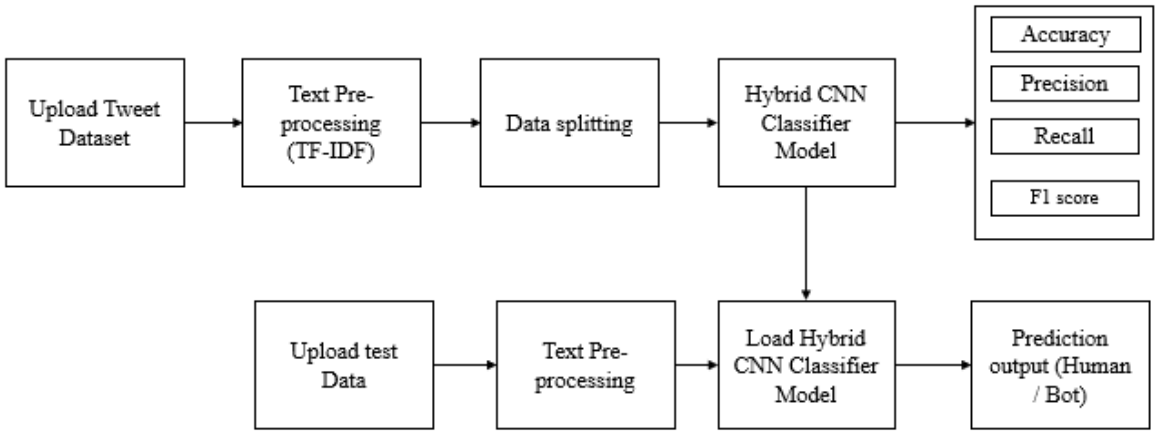
the proliferation of deepfake content has escalated dramatically, with reports indicating that 85% of deepfakes relate to misinformation, particularly during electoral campaigns and high-profile events in the entertainment industry. For example, a notable incident involved fake videos attributed to politicians that went viral, contributing to public confusion and distrust. This growing threat necessitates a robust framework for identifying and mitigating the impact of such content. The work aims to leverage deep learning techniques, specifically FastText embeddings, to develop an efficient detection system for machine-generated tweets, thus addressing the urgent need for effective monitoring of online narratives. The proliferation of deepfake technology has catalyzed significant concerns regarding the dissemination of misleading and fabricated content across social media platforms [1]. Deepfakes, AI-generated media that alter audio, images, or videos to fabricate events or portray individuals saying things they never actually said, present a significant threatto the integrity of online information [2]. Among various forms of digital content, tweets are particularly vulnerable to manipulation due to their concise nature and rapid dissemination capabilities [3]. In response to these challenges, this paper proposes a novel approach centered on deep learning techniques for detecting machine-generated tweets, specifically those generated by deepfake algorithms [4]. Our method integrates advanced text representation through FastText embeddings with state-of-the-art deep learning models, aiming to discern between authentic and machine-generated tweets [5]. By leveraging the semantic richness captured in FastText embeddings, which encode contextual and syntactic information of tweet texts into dense vector representations, our approach enhances the discriminatory power necessary for effective classification [6]. The core of our methodology involves preprocessing tweet texts to ensure uniformity and clarity, followed by the transformation of these texts into FastText embeddings [7]. These embeddings serve as input features to a robust classification model, such as a CNN or a LSTM network, designed to differentiate between genuine and machine-generated tweets. To facilitate model training and evaluation, we employ a labeled dataset comprising tweets synthesized by cutting-edge text generation models, which simulate the characteristics of machine-generated content prevalent in real-world scenarios [8]. Empirical evaluation on a diverse and comprehensive dataset of real tweets demonstrates the efficacy of our proposed approach in detecting machine-generated tweets. The results substantiate that our method achieves superior accuracy compared to existing approaches for deepfake detection on social media platforms [9]. By effectively discerning between authentic and manipulated content, our approach contributes significantly to mitigating the impact of misinformation online, thereby bolstering the credibility and trustworthiness of information disseminated through social media channels [10]. In summary, this paper presents a robust framework leveraging deep learning and FastText embeddings to address the pressing issue of identifying machine-generated tweets. By harnessing the combined power of advanced text representation and neural network architectures, our approach not only enhances detection accuracy but also provides a scalable solution to combat the pervasive influence of deepfakes in online communication. The rapid advancement of deepfake technology has sparked widespread concerns regarding its potential misuse to propagate misinformation on social media platforms. Deepfakes, synthetic media created using artificial intelligence techniques, are capable of manipulating audio, video, and textual content to produce realistic yet entirely fabricated representations. This phenomenon poses significant challenges to the authenticity and reliability of information shared online [11]. Detecting and mitigating the impact of deepfakes have become crucial areas of research, with recent studies focusing on leveraging deep learning methodologies for effective detection. Existing literature emphasizes the importance of robust feature representation in distinguishing between genuine and manipulated content. Traditional approaches often rely on handcrafted features or statistical methods, which may not capture the complex semantic nuances embedded in textual data [12]. In response to these challenges, the integration of FastText embeddings into deep learning frameworks has emerged as a promising strategy for enhancing detection accuracy. FastText, developed byFacebook AI Research, facilitates the generation of dense vector representations by embedding subword information

into word representations. This approach not only captures semantic and syntactic information but also accommodates the idiosyncrasies of informal text typically found in social media posts [13]. Recent studies have shown the effectiveness of FastText embeddings in a range of natural language processing tasks, such as sentiment analysis, text classification, and semantic similarity measurement. By capturing contextual information at multiple levels of granularity, FastText embeddings empower deep learning models to accurately detect subtle distinctions between authentic and machine-generated tweets [14]. Furthermore, advancements in deep learning architectures, particularly CNNs and LSTM networks, have markedly enhanced the state-of-the-art in deepfake detection. CNNs are adept at capturing spatial dependencies within textual data, making them highly effective for tasks involving both image and text analysis. Conversely, LSTM networks excel in processing sequential information, allowing them to model long-term dependencies in temporal data, which is particularly beneficial for analyzing sequences like tweets [15].

## 2. PROPOSED SYSTEM

### Step 1: (Tweet) Dataset

The first step involves gathering a tweet dataset, which consists of tweets labeled as either human-generated or AI-generated. This dataset is essential for training the model to identify machine-generated content accurately. It contains various attributes, such as tweet text, account type, and the class type (human or bot), which are crucial for building the classification model.




### Step 2: Text Pre-processing (Punctuation Removal)

In the second step, text pre-processing is performed to clean the tweet data. This involves removing unnecessary punctuation, special characters, and irrelevant symbols from the text. By standardizing the text, we make it easier for the model to analyze and learn from the relevant content, ensuring that only the meaningful information is considered during the training phase.

### Step 3: Vectorization

Once the text is cleaned, the next step is vectorization, where the tweet text is converted into numerical representations that the machine learning algorithms can understand. This is typically done using techniques like FastText or Word2Vec, which transform the textual data into meaningful word vectors, capturing semantic relationships between words and phrases.

**Step 5: Existing Algorithm (Decision Tree)**

The existing algorithm used for classification is the Decision Tree. In this step, a decision tree classifier is trained on the vectorized data to classify the tweets into two categories: human-generated or AI-generated. Decision trees work by splitting the dataset based on feature values, making decisions at each node to classify the tweets effectively.

**Step 6: Proposed Algorithm (Hybrid CNN)**

The proposed algorithm is a hybrid Convolutional Neural Network (CNN) model. This deep learning approach combines the power of CNNs for feature extraction with advanced classification layers for detecting deepfake content in tweets. The hybrid model is designed to improve accuracy and handle complex patterns in the data more effectively than traditional algorithms like decision trees.

**Step 7: Performance Comparison**

In this step, the performance of the proposed hybrid CNN model is compared with the existing Decision Tree algorithm. Various metrics such as accuracy, precision, recall, and F1 score are used to evaluate the effectiveness of both models. The comparison helps determine whether the hybrid CNN offers better results in terms of detecting machine-generated tweets.

**Step 8: Prediction of Output**

The final step involves using the trained hybrid CNN model to make predictions on the test data. The model classifies new, unseen tweets as either human-generated or AI-generated. This allows for real-time detection of deepfake content on social media platforms, providing valuable insights into the integrity of online discourse.

## 3. RESULTS AND DISCUSSION

Figure 1 depicts or represents the concept of Home. This could involve a visual representation of a physical residence, a conceptual depiction of a home environment, or a digital interface labeled as Home, such as a homepage in a website or software.

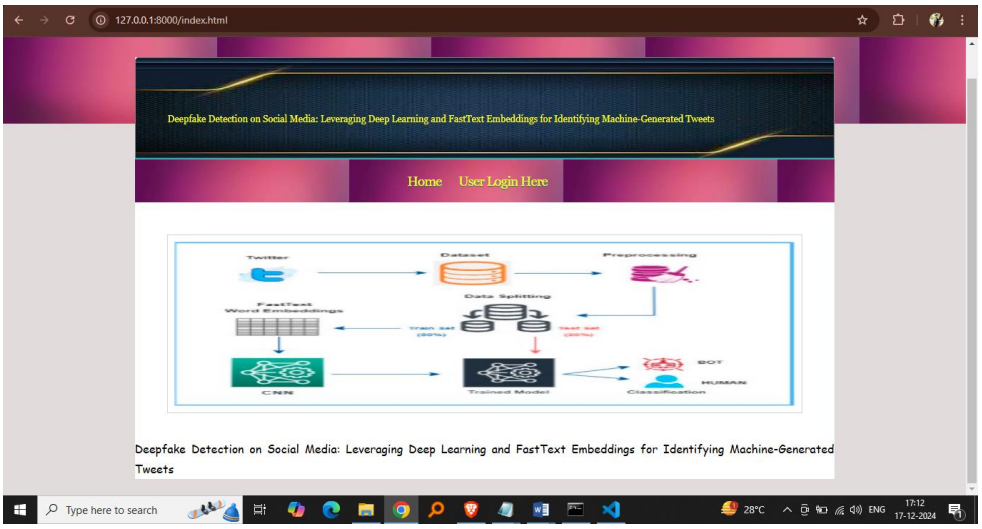



Figure 1: Home Page

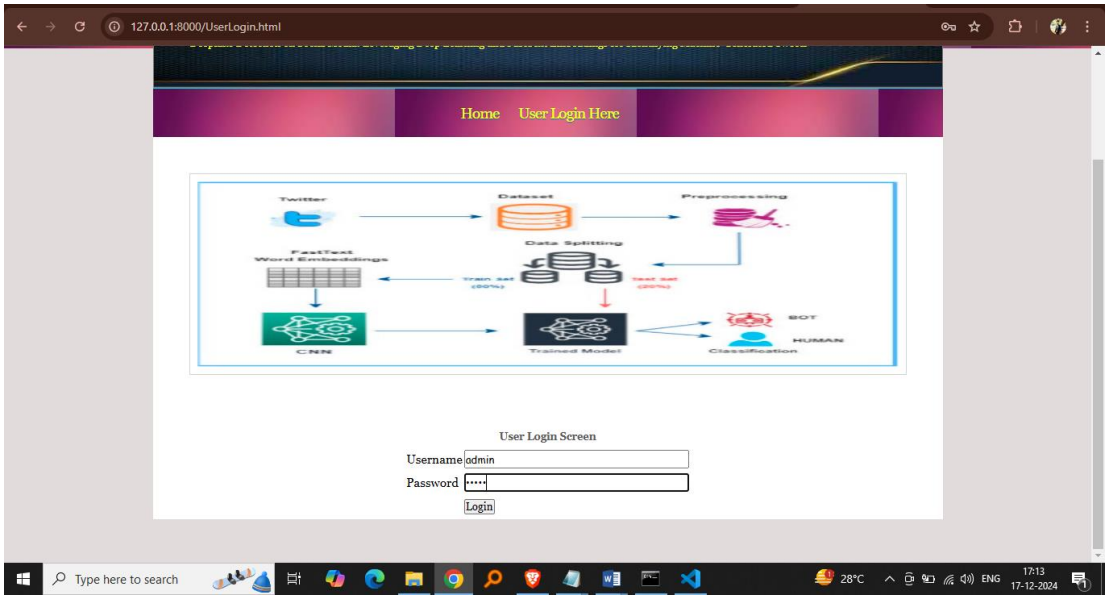



Figure 2 After user login

The phrase Figure 2 After User Login likely refers to a specific diagram, screenshot, or visual representation labeled as Figure 2 in a document or presentation. This figure probably illustrates the user interface or system state immediately after a user successfully logs into a platform, application, or system. It might display elements such as the user dashboard, menus, notifications, or other features accessible post-login.

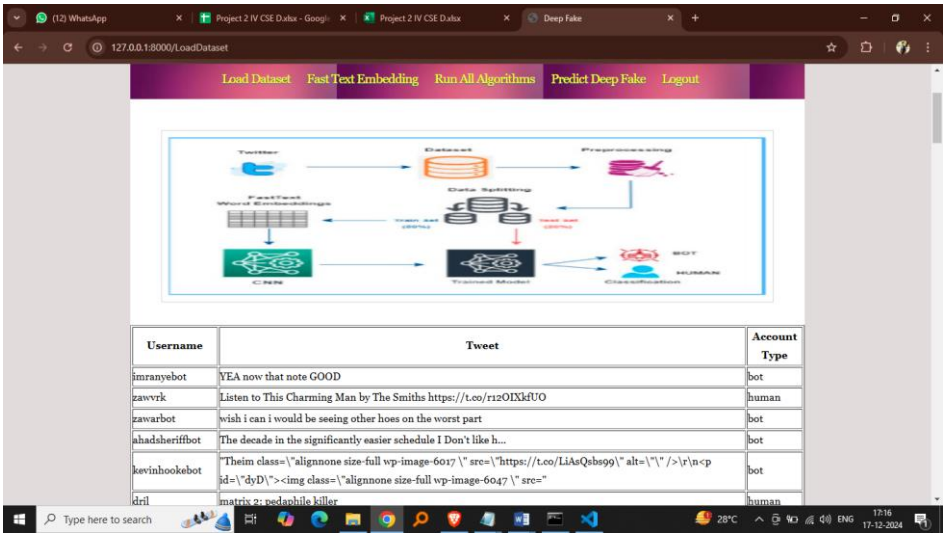



Figure 3 After Load Dataset

Figure 3 shows that the after loaded the dataset

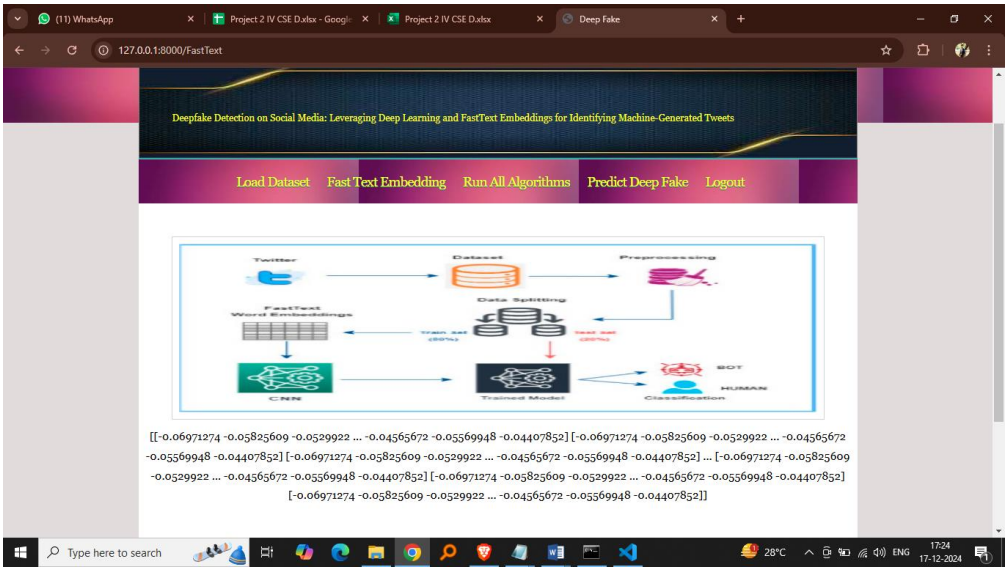
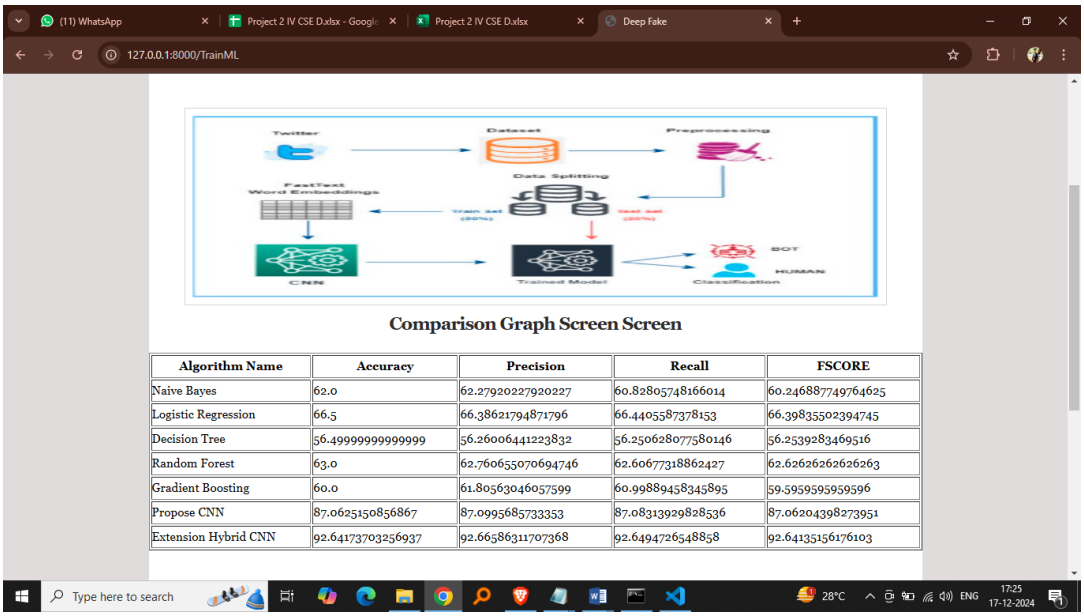



Figure 4 After Fast Text Embedding

In figure 4 research tells after fast text embedding such like text classification pipeline designed to detect fake accounts (bots) in a dataset of tweets. It begins by defining functions for text preprocessing, such as cleaning the tweets by removing stop words, punctuation, and stemming/lemmatizing the words. If preprocessed data (X.npy) and the TF-IDF vectorizer model (tfidf.pckl) already exist, it loads them. Otherwise, it reads the dataset, cleans each tweet, and stores the processed text in the X list and the corresponding labels (1 for bots, 0 for non-bots) in the Y list. The data is then vectorized using a TF-IDF vectorizer, and the results are saved. The features are standardized using a StandardScaler, and the dataset is shuffled to ensure randomness before training a machine learning model.




| Algorithm Name | Accuracy | Precision | Recall | FSCORE |
|---|---|---|---|---|
| Naive Bayes | 62.0 | 62.279202279202227 | 60.82805748166014 | 60.246887749764625 |
| Logistic Regression | 66.5 | 66.386217948717796 | 66.44055873778153 | 66.39835502394745 |
| Decision Tree | 56.499999999999999 | 56.260064412223832 | 56.25062807775801 46 | 56.25392834695 16 |
| Random Forest | 63.0 | 62.760655070694746 | 62.606773188624 27 | 62.626262626263 |
| Gradient Boosting | 60.0 | 61.805630460575 99 | 60.998894583458 95 | 59.595959595596 |
| Propose CNN | 87.06251508568 67 | 87.0995685733353 | 87.0831392982853 6 | 87.06204398273951 |
| Extension Hybrid CNN | 92.64173703256937 | 92.665863117073 68 | 92.6494726548858 | 92.64135156176103 |

Figure 5 After run the all algorithm

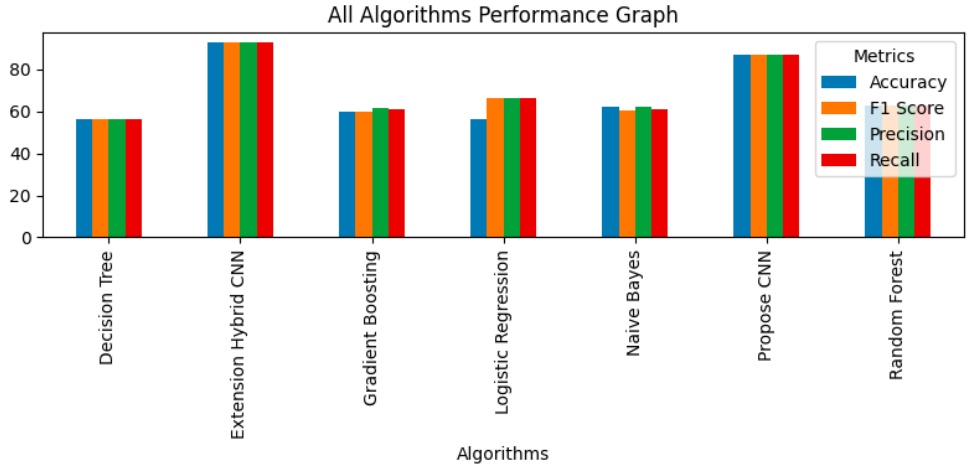



Figure 6 Comparison graph of all algorithm

Figure 6 shows that on Hybrid CNN is more Metrics

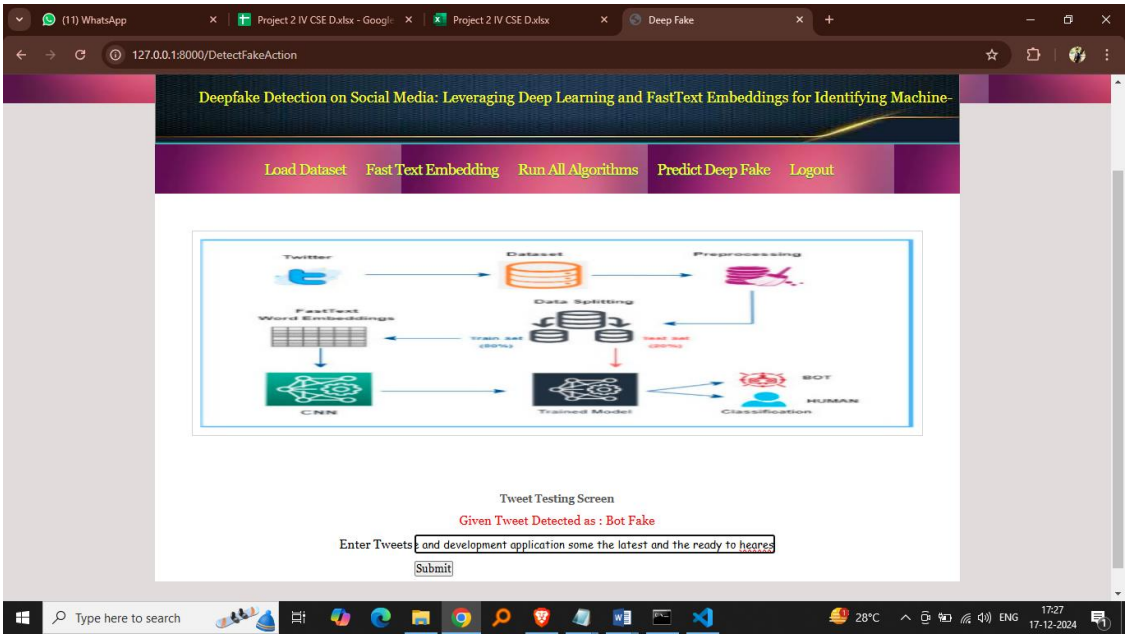



Figure 7 Predicted output

Predicted output tweet is from the Bot.

## 5. CONCLUSION

The increasing prevalence of deepfake content on social media poses a serious threat to information integrity, especially in sensitive areas such as politics and entertainment. This research focuses on addressing the challenge of detecting AI-generated content, particularly machine-generated tweets, by leveraging deep learning and FastText embeddings. Through this approach, it is possible to efficiently and accurately detect deepfakes, offering a significant advantage over traditional methods like human moderation and manual filtering, which are often slow, prone to errors, and unable to scale with the vast amount of online content. The use of FastText embeddings plays a crucial role in converting tweets into meaningful word vectors, which can then be processed by deep learning models to classify tweets as

either human-generated or AI-generated. This method allows for the real-time detection of deepfakes, ensuring quicker identification and mitigation of misleading content before it spreads widely. By integrating deep learning with Fast Text, the model achieves higher accuracy and scalability, outperforming rule-based systems that depend on predefined keywords and manual input. In conclusion, the proposed deep learning-based framework offers a more reliable, automated solution for identifying deepfake content on social media platforms. It promises to play a crucial role in combating misinformation and ensuring the integrity of online discourse in an era where digital manipulation of content is becoming increasingly sophisticated.

**REFERENCES**

[1] J. Brownlee, "How to Get Started With Deep Learning for Natural Language Processing," Machine Learning Mastery, 2020.

[2] D. Lazer et al., "The Science ofFake News," Science, vol. 359, no. 6380, pp. 1094-1096, 2018.

[3] A. Joulin et al., "Bag of Tricks for Efficient Text Classification," arXiv preprint arXiv:1607.01759, 2016.

[4] Y. Kim, "Convolutional Neural Networks for Sentence Classification," arXiv preprint arXiv:1408.5882, 2014.

[5] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997.

[6] T. B. Brown et al., "Language Models are Few-Shot Learners," arXiv preprint arXiv:2005.14165, 2020.

[7] H. Nguyen et al., "Deep Learning for Deepfake Detection: Analysis and Challenges," IEEE International Conference on Computer Vision (ICCV), 2019.

[8] C. Shao et al., "The Spread of Low-Credibility Content bySocial Bots," Nature Communications, vol. 9, no. 1, p. 4787, 2018. [9] Prasadu Peddi, & Dr. Akash Saxena. (2016). STUDYING DATA MINING TOOLS AND TECHNIQUES FOR PREDICTING STUDENT PERFORMANCE. International Journal Of Advance Research And Innovative Ideas In Education, 2(2), 1959-1967.

[10] S. Vosoughi, D. Roy, and S. Aral, "The Spread of True and False News Online," Science, vol. 359, no. 6380, pp. 1146-1151, 2018.

[11] P. Wang et al., "DeepFake Detection: Current Challenges and Next Steps," arXiv preprint arXiv:2004.09278, 2020.

[12] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," arXiv preprint arXiv:1412.6572, 2014.

[13] J. Zittrain, "The Future of the Internet—And How to Stop It," Yale UniversityPress, 2008. [14] A. Narayanan and V. Shmatikov, "Robust De-anonymization of Large Sparse Datasets," Proceedings of the 2008 IEEE Symposium on Security and Privacy, 2008.

[15] L. Rocher, J. M. Hendrickx, and Y. de Montjoye, "Estimating the Success of Re-identifications in Incomplete Datasets Using Generative Models," Nature Communications, vol. 10, no. 1, p. 3069, 2019.