10.48047/jocaaa.2025.34.05.34

Automated Detection of Machine-Generated Tweets Using Hybrid Deep CNN and FastText Embeddings

Ch. Sreelatha[1*], K. Pavani[2], L. Reshmi Bhavishya[2], K. Sri Thanay[2], V. Abhiram[2]

[1]Assistant Professor, [2]UG Student, [1,2]Department of Computer Science and Engineering-Cyber Security, [1,2]Malla Reddy Engineering College and Management Sciences, Kistapur, Medchal, 501401, Telangana, India
*Corresponding author: sreelatha@mrem.ac.in

Received: 02.03.2025 Revised: 15.04.2025 Accepted: 25.05.2025

ABSTRACT

Deepfake technology has emerged as a serious challenge to the authenticity of information shared on social media, particularly in domains such as politics and entertainment where fabricated videos and AI-generated content frequently fuel misinformation. The goal of this research is to design a robust deep learning–based framework capable of detecting deceptive, machine-generated content on social media by leveraging FastText embeddings for linguistic feature representation. Traditional moderation techniques—such as manual review, rule-based filtering, and fact-checking—are slow, error-prone, and cannot scale to the massive and continuously evolving volume of online data. To address these limitations, the proposed approach employs hybrid deep CNN models to automatically analyze tweet patterns and distinguish human-generated posts from AI-generated ones in real time. This architecture enhances detection accuracy, efficiency, and scalability, contributing to safer digital environments and supporting efforts to reduce misinformation and protect the integrity of online communication.

Key words: Social Media Bots, Hybrid Deep CNN, FastText Embeddings, NLP, Real-Time Content Monitoring, Automated Moderation

1. INTRODUCTION

Online Social Networks (OSNs) have revolutionized digital communication, providing a platform for instant information sharing, discussion, and interaction. Among these, Twitter stands out as one of the most influential social media platforms due to its vast user base and real-time content dissemination. However, the rapid growth of OSNs has also led to the increasing spread of misinformation and deep fake content, particularly in the form of machine-generated tweets. This issue has significantly impacted various sectors, including politics, education, and public trust.

10.48047/jocaaa.2025.34.05.34

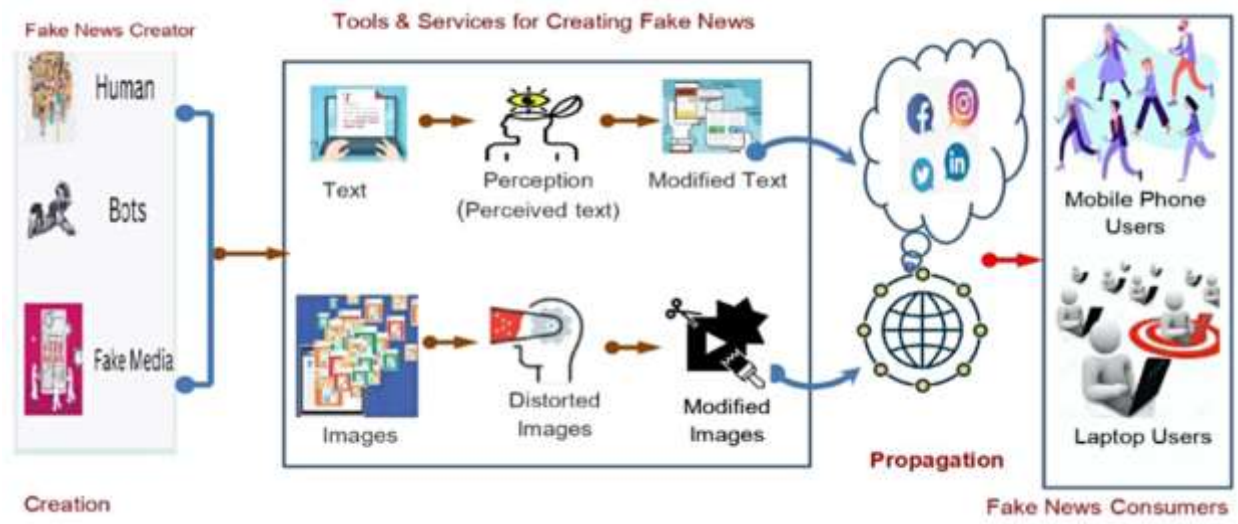


Fig 1: Bot Detection

India, with one of the highest numbers of social media users, is particularly vulnerable to the negative consequences of deep fakes. Political misinformation, fake news campaigns, and manipulated narratives have influenced public opinion and even national elections. Students, who rely on OSNs for news, academic discussions, and social interactions, are especially susceptible to such deceptive content, leading to misinformation, academic dishonesty, and misinformed decision-making. Furthermore, the rampant spread of synthetic media has eroded public trust in digital communication, making it difficult for users to distinguish between real and fake content.

Given Twitter's dominance as a primary communication tool for news, politics, and social interactions, detecting machine-generated tweets is critical. Traditional moderation techniques and manual fact-checking are insufficient to tackle the scale and sophistication of AI-generated content. Therefore, leveraging deep learning models combined with Fast Text embeddings can significantly enhance the accuracy and efficiency of detecting synthetic tweets. Fast Text embeddings allow models to capture contextual and sub-word-level information, making them highly effective in identifying unnatural patterns in text.

One of the biggest consequences of deep fake content is the erosion of public trust in OSNs. When users are repeatedly exposed to misleading information, fake news, and AI-generated tweets, their confidence in social media as a reliable source of information diminishes. This decline in trust is particularly concerning for governments, media organizations, and businesses that rely on OSNs to communicate with the public. Even traditional fact-checking methods struggle to keep up with the volume of fake content being generated. This has led to growing concerns over the credibility of social media platforms and their ability to regulate misinformation effectively.

2. LITERATURE SURVEY

Online Social Networks (OSNs) have revolutionized digital communication, providing a platform for instant information sharing, discussion, and interaction. Among these, Twitter stands out as one of the most influential social media platforms due to its vast user base and real-time content dissemination. However, the rapid growth of OSNs has also led to the increasing spread of misinformation and deep fake

10.48047/jocaaa.2025.34.05.34

content, particularly in the form of machine-generated tweets. This issue has significantly impacted various sectors, including politics, education, and public trust.

India, with one of the highest numbers of social media users, is particularly vulnerable to the negative consequences of deep fakes. Political misinformation, fake news campaigns, and manipulated narratives have influenced public opinion and even national elections. Students, who rely on OSNs for news, academic discussions, and social interactions, are especially susceptible to such deceptive content, leading to misinformation, academic dishonesty, and misinformed decision-making. Furthermore, the rampant spread of synthetic media has eroded public trust in digital communication, making it difficult for users to distinguish between real and fake content.

Nair, Vinita, et al. [11] (2025) addressed the fake news have posed significant challenges in the digital age, leading to various social impacts such as bandwagon effects, validity concerns, echo chambers, and misinformation propagation. This study proposed a framework for the automatic detection of fake news using a knowledge-based approach. An automatic fact-checking mechanism was implemented using concepts from Information Retrieval (IR), Natural Language Processing (NLP), and Graph Theory.

Kotiyal, et al. [12] (2025) proposed fake news on social media has become a significant concern due to its influence on public opinion. This study proposed the Debunking Multi-Lingual Social Media Posts using Deep Learning (DSMPD) approach to detect fake news. A dataset of English and Hindi social media posts was created using web scraping and Natural Language Processing (NLP) techniques. The dataset was utilized to train, test, and validate a deep learning-based model that extracted multiple features, including Embedding from Language Models (ELMo), word and n-gram counts, Term Frequency-Inverse Document Frequency (TF-IDF), sentiment analysis, polarity, and Named Entity Recognition (NER).

Saeed, et al. [13] (2025) introduced widespread accessibility of the internet has led to a significant rise in social media usage, contributing to the propagation of fake news. This study focused on the classification of fake news and misinformation from social media data corpora. Two fake news datasets were collected and preprocessed using various data improvement techniques. Deep text features were extracted using Word2Vec and Global Vectors for Word Representation (GloVe), while textual features were obtained using n-gram approaches, including Term Frequency-Inverse Document Frequency (TF-IDF) and Bag of Words (BoW). Bidirectional Encoder Representations from Transformers (BERT) was employed to derive embedded representations from the input data.

Alghamdi, et al. [14] (2025) proposed "Enhancing hierarchical attention networks with CNN and stylistic features for fake news detection." Expert Systems with Applications 257 (2024): 125024.The rise of social media platforms has led to an increase in the spread of false information, making it challenging to identify malicious entities due to the complexities of natural language and high textual data volume. This study proposed a neural-based framework that leveraged the hierarchical structure of text to detect both fake news content and fake news spreaders. The approach utilized enhanced Hierarchical Convolutional Attention Networks (eHCAN), incorporating both style-based and sentiment-based features to improve model performance. The results demonstrated that eHCAN outperformed several strong baseline methods, showcasing the effectiveness of integrating deep learning (DL) with stylistic features.

Dhaka, et al. [15] (2025) proposed "USE-Based Deep Learning Approach to Detect Spammers on Twitter." In International Conference on Smart Trends in Computing and Communications, pp. 407-416.

10.48047/jocaaa.2025.34.05.34

Singapore: Springer Nature Singapore.The detection of spammers disseminating obscene content on Twitter has been explored using various hybrid features and machine learning approaches. This study aimed to determine the most suitable encoding techniques for representing long text documents. A novel deep learning model was proposed, incorporating the Universal Sentence Encoder (USE) as a feature extractor and an artificial neural network (ANN) as a classifier. Sentence vectors representing users' tweets were transformed into document vectors, which served as high-quality features for classification.

Sahin, Muammer Eren, et al. [16] (2025) invested the rapid growth of social media has facilitated both positive interactions and the spread of misleading information, such as fake news. Existing research has explored machine learning techniques for automated fake news classification. This study extended such efforts by proposing an approach that utilized word embedding and a Long Short-Term Memory (LSTM) neural network. Unlike previous studies, the proposed model was evaluated using two publicly available news article datasets. The results demonstrated that the model outperformed baseline machine learning models, achieving 99% and 96% accuracy on the first and second datasets, respectively.

Widjaja, Elbert et al. [17] (2025) studied Fake news detection using social media data aimed to identify the most effective state-of-the-art machine learning model for detecting fake news on social media. Various data sources were analyzed using key metrics—amount of data, credibility, and bias—to ensure unbiased and reliable training datasets. Multiple text preprocessing techniques were explored, with the optimal method determined based on the dataset's domain and size. Several machines learning models, including Naïve Bayes, Word Embedding LSTM, and transformer models, were evaluated using accuracy, F1-score, precision, recall, and MCC.

AlJamal, Mahmoud, et al. [18] (2025) introduced Novel Text Embedding Approach for Fake News Detection on Twitter X: Integrating Social Context, Temporal Dynamics, and Enhanced Interpretability. International Journal of Computational Intelligence Systems 18, no. 1 (2025): 22.In response to the growing challenge of misinformation on social media, this study proposed an optimized approach for fake news detection by combining BERT and GloVe embeddings with Principal Component Analysis (PCA) and attention mechanisms. Social and temporal features were integrated to enhance text representation. Utilizing the CIC Truth Seeker Dataset 2023, SHAP was applied for feature selection ad interpretability, ensuring transparency in the model’s predictions.

Sinha, et al. [19] (2025) proposed "Text-convolutional neural networks for fake news detection in Tweets." In Evolution in Computational Intelligence:Frontiers in Intelligent Computing: Theory and Applications (FICTA 2020), Volume 1, pp. 81-90. Springer Singapore, 2021.With the increasing use of online social networking platforms, user-generated content has played a significant role in news dissemination. However, these platforms have also been exploited to spread misleading information for political gains, often without fact-checking. This study aimed to develop an automatic news verification system to classify tweets as real or fake. Due to the complexity of natural language and multilingual variations, a deep learning model was proposed to learn semantic word embeddings for improved classification accuracy.

3. PROPOSED SYSTEM

The rapid advancement of deepfake technology has significantly reshaped the digital landscape, particularly on social media platforms where misinformation can spread at an alarming rate. Deepfakes leverage sophisticated artificial intelligence (AI) techniques to generate highly realistic synthetic content,

10.48047/jocaaa.2025.34.05.34

posing a severe threat to public trust, media integrity, and cybersecurity. Among the various forms of deepfake-generated content, machine-generated tweets—crafted using deep learning models—are increasingly being used to manipulate public opinion, spread propaganda, and deceive users. As AI-generated text becomes more convincing, distinguishing between authentic and synthetic content presents a growing challenge for social media users, policymakers, and cybersecurity experts.

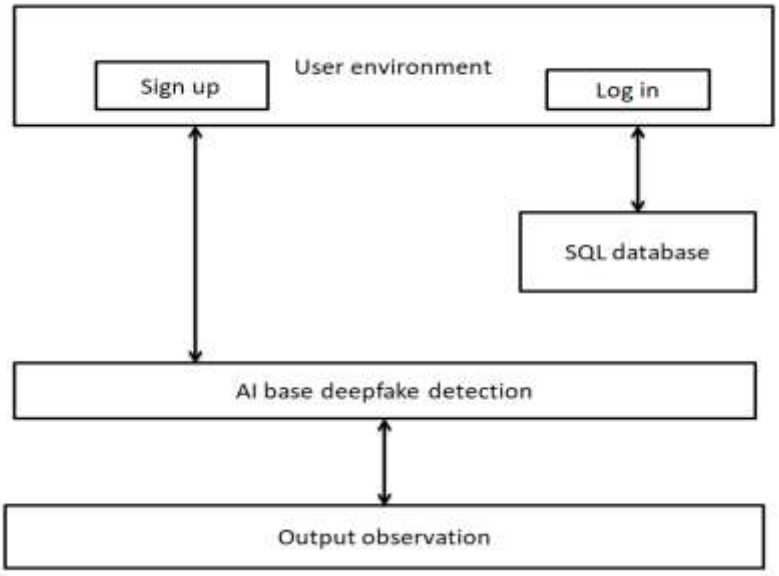


Figure 1. Proposed Django architecture.

To address this challenge, deep learning techniques combined with natural language processing (NLP) approaches, such as FastText embeddings, have emerged as effective tools for detecting machine-generated content. Advanced deep learning models, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformers, have demonstrated remarkable performance in text classification and pattern recognition. FastText embeddings, in particular, excel at capturing contextual relationships in text, making them valuable for distinguishing human-written tweets from AI-generated ones. This study explores the integration of these techniques to enhance detection accuracy and mitigate the impact of AI-driven misinformation. By analyzing linguistic patterns, semantic inconsistencies, and structural features, our research contributes to the broader field of digital forensics and online security.

With the increasing spread of deepfake technology, detecting AI-generated tweets has become a critical task in ensuring the integrity of social media content. This system leverages Django's Model-View-Template (MVT) architecture to build a scalable and efficient deep fake detection tool. The framework integrates FastText embeddings, deep learning (CNN), and a Random Forest classifier to accurately classify tweets as either human-written (Normal) or AI-generated (Fake).

The proposed system architecture as shown in Figure 4.1. for AI-based deep fake detection follows a structured Django MVT framework, integrating user authentication, database management, and AI-driven text classification. Users can sign up and log in, with their credentials securely stored in an SQL database. Once authenticated, they can submit tweets for analysis, which undergo NLP preprocessing,

10.48047/jocaaa.2025.34.05.34

FastText embedding conversion, deep feature extraction using CNNs, and final classification via a Random Forest model to determine if a tweet is human-written (Normal) or AI-generated (Fake). The results are stored in the database and displayed on a user-friendly HTML interface, showing the original tweet, classification label, and confidence score. The system ensures efficient data retrieval, history tracking, and real-time deepfake detection, helping users combat misinformation effectively while maintaining a seamless and secure experience. The system follows a structured algorithm to detect AI-generated (fake) tweets using FastText embeddings, Deep Learning CNN (DLCNN), and Random Forest Classifier (RFC) within a Django-based web application.

4. RESULTS AND DISCUSSION

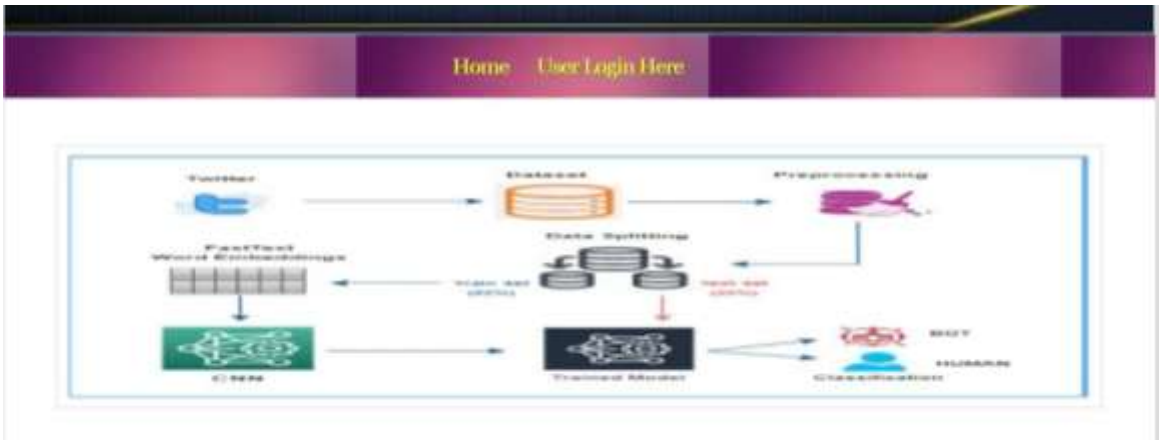


Fig 2. HTML Page of Research Work.

10.48047/jocaaa.2025.34.05.34

Figure 2 shows the HTML page of a research project focused on detecting deepfake content on social media, particularly machine-generated tweets. Below the header, there are navigation links for "Home" and "User Login Here." The main content includes a flowchart that illustrates the detection process, beginning with data collection from Twitter, preprocessing, word embedding using FastText, and training a CNN model. The trained model then classifies tweets as either bot-generated or human-generated.

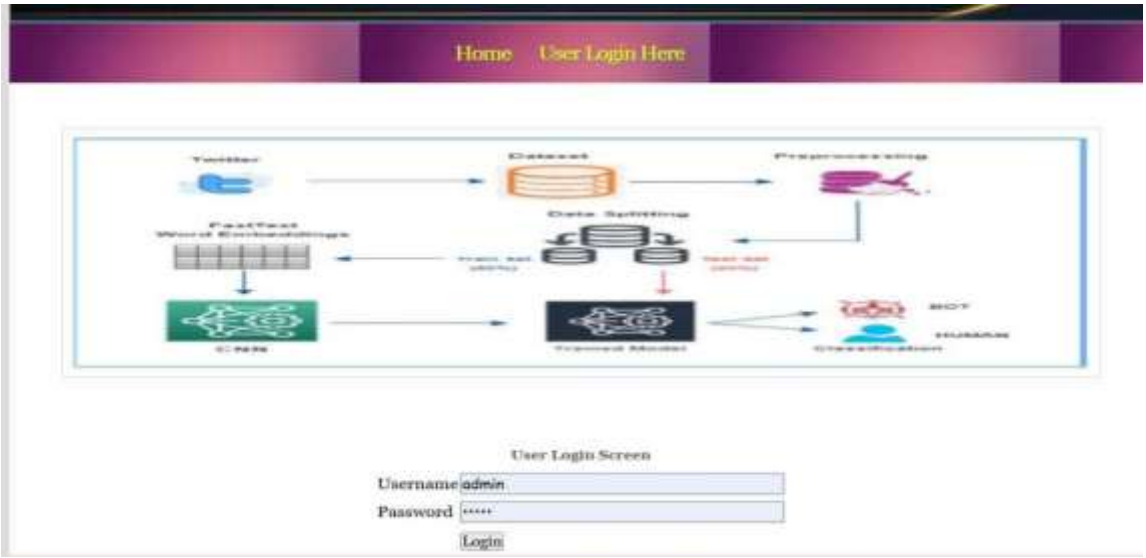
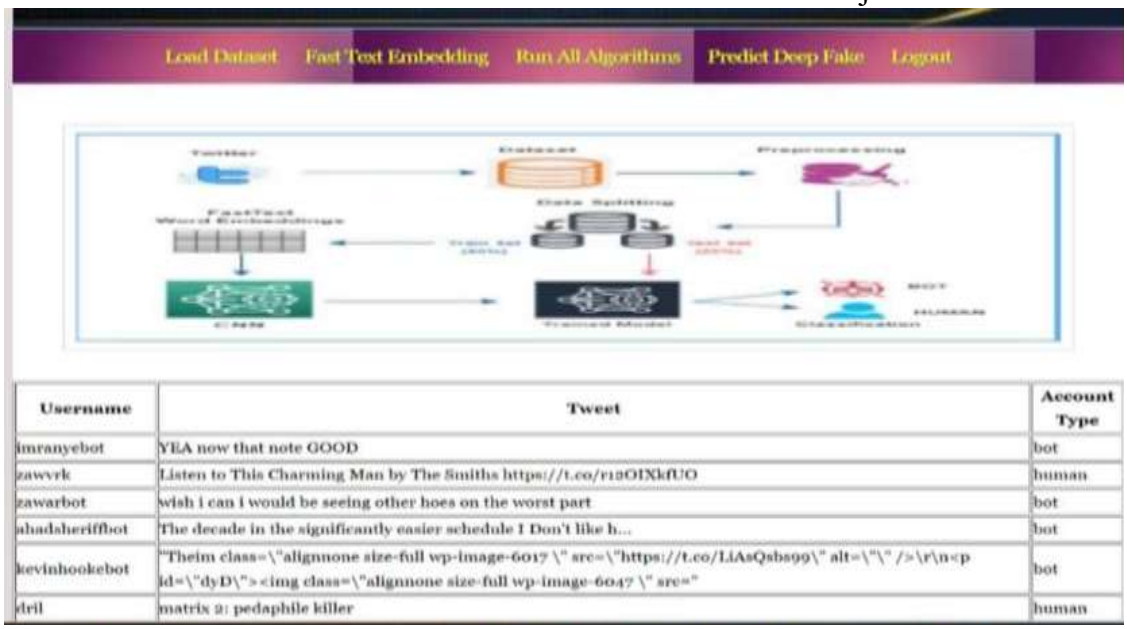


Fig 3. User Login Page

The fig: 3 displays a section of a research document showcasing the user interface for a deepfake detection system on social media. Here, a navigation menu includes links like "Home" and "User Login Here." A flowchart visually represents the model's workflow, covering data collection from Twitter, preprocessing, FastText embeddings, CNN-based training, and classification of tweets as human- or bot-generated. Additionally, the *User Login Page* contains a login form requiring a username and password for system access, ensuring authentication before usage. Dataset loading is a crucial step in the deepfake detection process, as shown in *Fig 3*. This phase involves importing and preprocessing the dataset containing tweets for analysis. The dataset is obtained from sources like Twitter and is structured to include text samples, metadata, and labels indicating whether a tweet is generated by a bot or a human. Once loaded, the data undergoes preprocessing, which includes text cleaning, tokenization, and transformation into FastText embeddings to prepare it for model training. The dataset is then split into training and testing sets, ensuring an optimized balance for deep learning models, such as CNNs, to learn and classify tweets accurately. This process is essential for building a robust model that can effectively differentiate between genuine and machine-generated tweets.

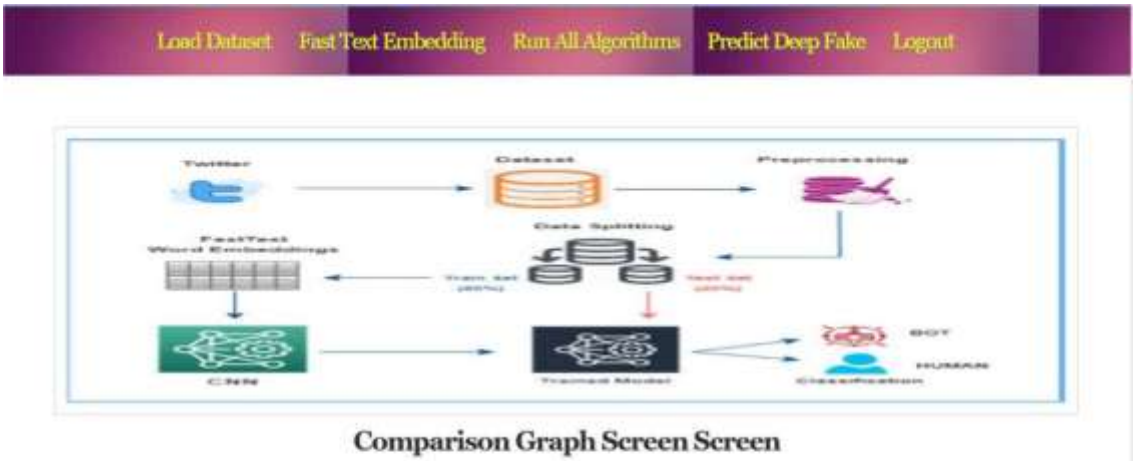
10.48047/jocaaa.2025.34.05.34



Username	Tweet	Account Type
imranyebot	YEA now that note GOOD	bot
zawvrk	Listen to This Charming Man by The Smiths https://t.co/r12OIXkfUO	human
zawarbot	wish i can i would be seeing other hoes on the worst part	bot
ahadsheriffbot	The decade in the significantly easier schedule I Don't like h...	bot
kevinhookebot	"Theim class=\"alignnone size-full wp-image-6017 \" src=\"https://t.co/LiAsQsbs99\" alt=\"\" />\r\n<p id=\"dyD\"><img class=\"alignnone size-full wp-image-6047 \" src="	bot
dril	matrix 2: pedaphile killer	human

Fig 4. DataSet Loaded

The dataset loaded in the fig: 4 appears to be related to the detection of deepfake content on social media using deep learning and text embeddings. It includes columns such as "Username," "Tweet," "Account," and "Type," where the "Type" column categorizes whether the tweet is generated by a bot or a human. The user interface shows options to load the dataset, perform text embedding, run various algorithms, and predict deepfake content. The dataset likely consists of social media posts, usernames, and classification labels that help train or evaluate an AI model for detecting inauthentic or manipulated text content.



Comparison Graph Screen Screen

Algorithm Name	Accuracy	Precision	Recall	FSCORE
Naive Bayes	55.00000000000001	55.76391393196298	54.7954795479548	52.9239460194581
Logistic Regression	56.99999999999999	57.36072206660442	57.095709570957084	56.64885573142454
Decision Tree	56.49999999999999	56.616161616161605	56.55065506550655	56.41173376086576
Random Forest	59.5	59.56443195503813	59.530953095309535	59.47467166979362
Gradient Boosting	62.0	64.17838682533959	61.7961796179618	60.24688774976461
CNN	87.78662804730871	87.86347145178462	87.79485748962493	87.781957779419
Proposed CNN With RFC	95.17490952955367	95.32971204952982	95.17141027879634	95.17051743623801

10.48047/jocaaa.2025.34.05.34

Fig 5. All Algorithms

The fig 5 compares various machine learning algorithms for classifying social media tweets as human- or bot-generated, evaluated using Accuracy, F1-score, Precision, and Recall. The Proposed CNN with RFC model achieves the highest performance, leveraging deep learning's feature extraction and ensemble learning's robustness. CNN also performs well due to its ability to capture contextual relationships in text embeddings. Traditional models like Decision Tree, Gradient Boosting, Logistic Regression, and Naïve Bayes show moderate performance, with Naïve Bayes being the weakest. Random Forest performs well but slightly below the proposed hybrid model. The results indicate that deep learning, particularly CNN combined with RFC, is the most effective approach for deepfake detection in social media.

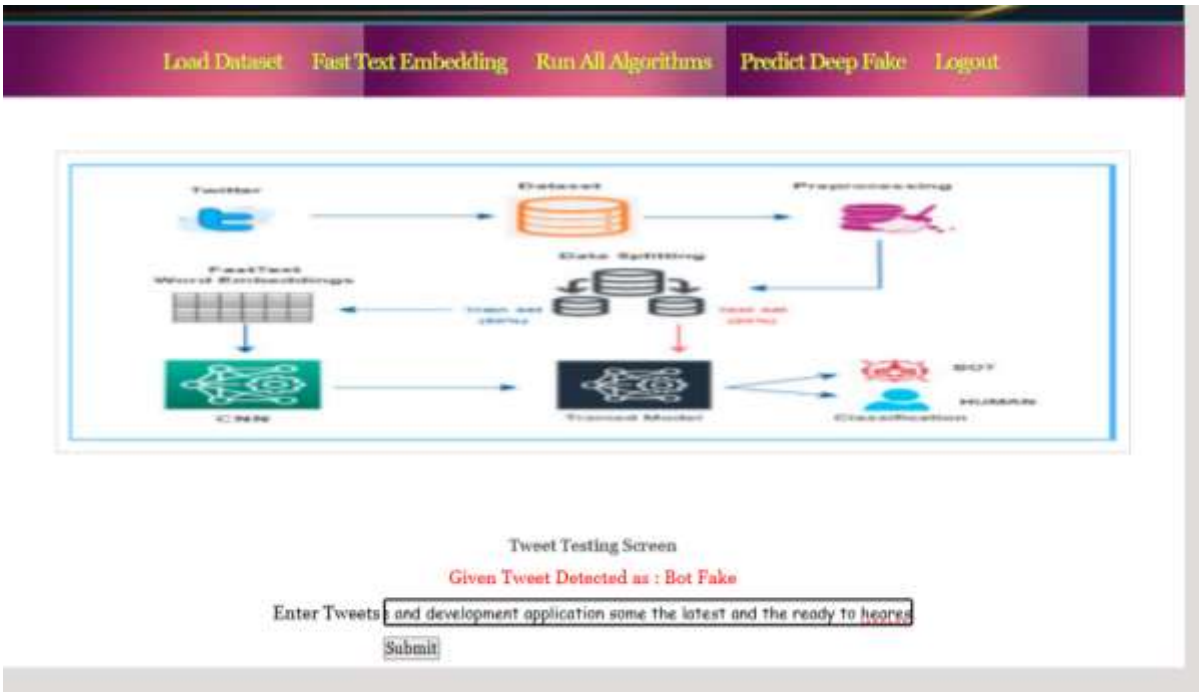


Fig 6. Predict Deep Fake

Figure 6 represents the Deepfake Detection System for social media, which uses deep learning and FastText embeddings to classify tweets as human-generated or bot- generated. The prediction interface allows users to input a tweet into the system, which then processes it through several stages: preprocessing, FastText word embeddings, and CNN-based classification. The trained model analyzes the input tweet and determines whether it originates from a bot or a human. This automated system ensures efficient deepfake detection by leveraging contextual word representations and neural network- based classification, helping mitigate misinformation on social media platforms.

10.48047/jocaaa.2025.34.05.34

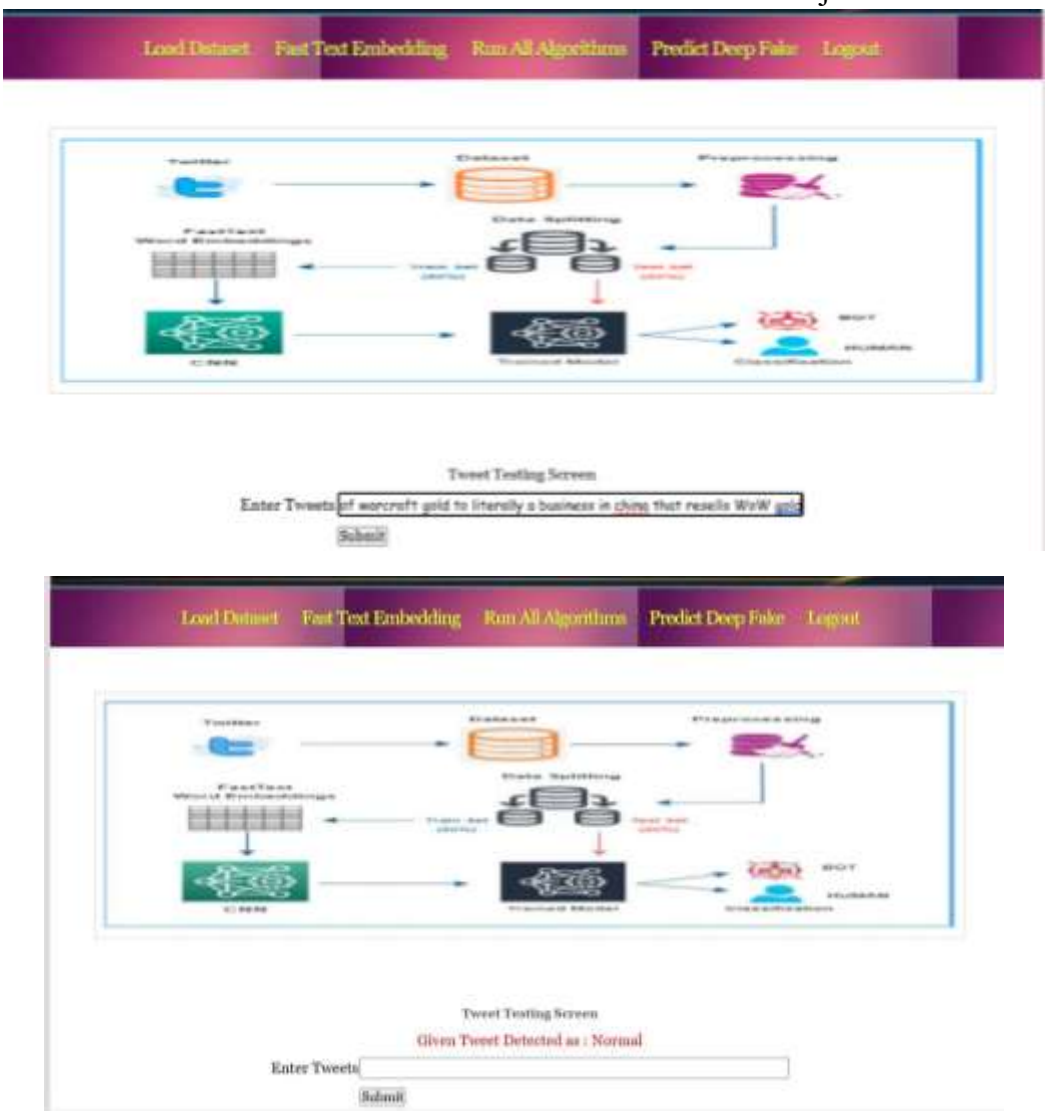


Fig 7. Output for the given label

The fig: 7 represents the Deepfake Detection System's output for a given tweet. The system takes an input tweet, processes it through FastText embeddings and a CNN-based classification model, and determines whether the tweet is generated by a bot or a human. In this specific case, the system has classified the given tweet as "Normal," indicating that it is likely human-generated rather than a deepfake or bot-generated content. The detection workflow involves data preprocessing, feature extraction using FastText, training with CNN, and final classification, ensuring an accurate prediction based on learned patterns.

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

10.48047/jocaaa.2025.34.05.34

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 FastText, 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. Gurusai, V. V. N. V., and M. Ravi Kumar.(2025)."DEEP FAKE DETECTION ON SOCIAL MEDIA LEVERAGING DEEP LEARNING AND FAST TEXT EMBEDDINGS FOR IDENTIFYING MACHINE-GENERATED TWEETS." Journal of Data Acquisition and Processing.

2. Sadiq, Saima, Turki Aljrees, and Saleem Ullah.(2025).Deepfake Detection on Social Media: Leveraging Deep Learning and Fast Text Embeddings for Identifying Machine-Generated Tweets.

3. Shifna, N. Fathima Shrene, K. Baalaji, Pasapala Niharika, Pasam Swathi, Pasangulapati Rama Krishna, and Reddimala Leena.(2025)."Identifying Machine Generated Tweets: Deepfake Detection on Social Media." In 2024 3rd Edition of IEEE Delhi Section Flagship Conference.

4. Alzahrani, Amani.(2025)."Misinformation Detection in the Social Media Era." PhD diss., Howard University.

5. Bhogi, Aditya, Anjani Dasari, Nikita Sharma, and Neha Garg.(2025)."Machine Learning for Fake News Detection on Social Media Text." In 2024 International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT), pp. 642-647. IEEE, 2023.

6. Mouratidis, Despoina, Maria Nefeli Nikiforos, and Katia Lida Kermanidis.(2025)."Deep learning for fake news detection in a pairwise textual input schema.

7. Mouratidis, Despoina, Maria Nefeli Nikiforos, and Katia Lida Kermanidis.(2025).Deep Learning for Fake News Detection in a Pairwise Textual Input Schema.

8. Shifna, N. Fathima Shrene, K. Baalaji, Pasapala Niharika, Pasam Swathi, Pasangulapati Rama Krishna, and Reddimala Leena.(2025)."Identifying Machine Generated Tweets: Deepfake Detection on Social Media." In 2024 3rd Edition of IEEE Delhi Section Flagship Conference (DELCON), pp. 1-5.

10.48047/jocaaa.2025.34.05.34

9. Hirlekar, Vaishali Vaibhav, and Arun Kumar.(2025)."An empirical analysis of fake tweet detection using statistical and deep learning approaches."In Information and Communication Technology for Competitive Strategies (ICTCS 2020) ICT: Applications and Social Interfaces, pp. 1033-1044.

10. Vallileka, N., P.Sundaravadivel, U.Karthikeyan, R.Santhana Krishnan, K. Lakshmi Narayanan, and S. Sundararajan.(2025)."DeepTweet:leveraging transformer-based models for enhanced fake news detection in twitter sentiment analysis." In 2024 7th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), pp.

11. Nair, Vinita, Jyoti Pareek, and Sanskruti Bhatt.(2024)."A Knowledge-Based Deep Learning Approach for Automatic Fake News Detection using BERT on Twitter." Procedia Computer Science 235.

12. Kotiyal, Bina, Heman Pathak, and Nipur Singh. .(2024)."Debunking multi-lingual social media posts using deep learning." International Journal of Information Technology 15.

13. Saeed,Ammar,and Eesa Al Solami.(2024)."Fake News Detection Using Machine Learning and Deep Learning Methods."CMC-COMPUTERS MATERIALS & CONTINUA 77, no. 2: 2079-2096.

14. Alghamdi,Jawaher,Yuqing Lin, and Suhuai Luo.(2024)."Enhancing hierarchical attention networks with CNN and stylistic features for fake news detection." Expert Systems with Applications 257: 125024.

15. Dhaka, Deepali,Saima Saleem, and Monica Mehrotra.(2024)."USE-Based Deep Learning Approach to Detect Spammers on Twitter." In International Conference on Smart Trends in Computing and Communications, pp. 407-416. Singapore: Springer Nature Singapore.

16. Sahin, Muammer Eren, Chunyang Tang, and Mohammad A. Al-Ramahi.(2024)."Fake News detection on social media: a word embedding-based approach."

17. Widjaja, Elbert.(2024)."Fake news detection using social media data."

18. AlJamal, Mahmoud, Rabee Alquran, Ayoub Alsarhan, Mohammad Aljaidi, Wafa'Q. Al-Jamal, and Ali Fayez Alkoradees.(2025)."Optimized Novel Text Embedding Approach for Fake News Detection on Twitter X: Integrating Social Context, Temporal Dynamics, and Enhanced Interpretability.

19. Sinha, Harsh, Sakshi, and Yashvardhan Sharma.(2025)."Text-convolutional neural networks for fake news detection in Tweets." In Evolution in Computational Intelligence: Frontiers in Intelligent Computing: Theory and Applications (FICTA 2020), Volume 1, pp. 81-90. Springer Singapore.