Deep Fake Detection for Identifying Machine Generated Tweets Using NLP and DL

G. Shabana[1], E. Madhuri[2], A. Nadeem[2],
D. Charan Kumar Reddy[2] and Y. Harshavardhan Lourdu Reddy[2]
[1]*Department of Computer Science and Engineering (AI&ML), Srinivasa Ramanujan Institute of Technology, Rotarypuram Village, B K Samudram Mandal, Anantapur, Andhra Pradesh, India*
[2]*Department of Computer Science and Engineering (Data Science), Srinivasa Ramanujan Institute of Technology, Rotarypuram Village, B K Samudram Mandal, Anantapur, Andhra Pradesh, India*

Keywords: Deepfake Detection, Machine-Generated Text, NLP, Deep Learning, CNN, FastText Embeddings, Bot-Generated Tweets, Social Media Misinformation, Fake Tweet Detection, AI-Generated Content.

Abstract: The rise of advanced natural language generation models has led to an increasing prevalence of deepfake content on social media, particularly machine-generated tweets designed to manipulate public opinion. This research presents a deep learning-based approach for detecting AI-generated tweets using a Convolutional Neural Network (CNN) combined with FastText word embeddings. The proposed model outperforms baseline machine learning methods that utilize Term Frequency (TF), Term Frequency-Inverse Document Frequency (TF-IDF), and traditional FastText embeddings. Experimental results demonstrate the superior accuracy of the CNN-FastText model in distinguishing human-written tweets from bot-generated ones. This study contributes to combating misinformation and enhancing online discourse integrity by providing an effective and scalable detection mechanism.

1 INTRODUCTION

Social media has revolutionized communication, enabling global connectivity and information sharing. However, it has also become a platform for misinformation, with AI-generated content such as deepfake text, images, and videos being used to manipulate public opinion. The rapid evolution of natural language generation models has made it increasingly difficult to distinguish between human-written and machine-generated text, raising concerns about online discourse integrity.

Deepfake text refers to AI-generated content designed to mimic human writing convincingly. Advances in deep learning models like GPT-2 and GPT-3 have enabled the creation of highly realistic tweets, making detection a significant challenge. Unlike long-form deepfake text, which has been widely studied, detecting machine-generated tweets remains an underexplored area, despite their potential for spreading misinformation on a large scale.

Existing approaches to deepfake text detection primarily rely on graph-based and feature-based machine learning techniques. However, these methods are often designed for longer texts such as news articles and lack efficiency in detecting AI-generated content in short social media posts. Additionally, traditional models struggle with challenges such as sarcasm, ambiguous language, and linguistic diversity, limiting their effectiveness in real-world applications.

To address these limitations, this study proposes a deep learning-based approach using a CNN with FastText embeddings to classify tweets as either human-generated or bot-generated. FastText embeddings enhance detection accuracy by capturing sub word-level information, making the model more robust against linguistic variations and complex language structures. The proposed method aims to improve detection performance while ensuring computational efficiency and scalability for real-time applications.

Shabana, G., Madhuri, E., Nadeem, A., Reddy, D. C. K. and Reddy, Y. H. L.
Deep Fake Detection for Identifying Machine Generated Tweets Using NLP and DL.
DOI: 10.5220/0013915700004919

In *Proceedings of the 1st International Conference on Research and Development in Information, Communication, and Computing Technologies (ICRDICCT'25 2025) - Volume 4*, pages 507-511
ISBN: 978-989-758-777-1

2 LITERATURE SURVEY

- **'Big data analytics: Challenges and applications for text, audio, video, and social media data:** A recommendation system can track a user's watching and reading behavior to locate relevant and entertaining material. A Hadoop-based recommendation engine for internet data on any item is shown in our study. This data collection includes review scores, complaints, notes, feedback, and reviews. We examined movie review and rating site data with Mahout Interfaces.
- **The emergence of deepfake technology: A review:** Modern internet technologies make fake media harder to spot. The latest challenge is deepfakes, lifelike videos that use AI to make people say or do things they didn't. Credible deepfakes can hurt society by reaching millions on social media. Despite the paucity of scientific literature, the paper defines deepfake, names its creators, evaluates its pros and cons, offers examples, and recommends strategies to stop its spread. It analyses 84 public online news stories. The results show that law, corporate policy and volunteer action, education and training, and new technology can detect, authenticate, and prevent deepfakes. Deepfakes plague politics, society, and business. This paper examines deepfakes and gives cybersecurity and AI companies' ways to counteract media forgeries and misinformation.
- **Deepfake warnings for political videos increase disbelief but do not improve discernment: Evidence from two experiments:** Recent advances in ML allow the creation of a "deepfake" a convincing computer-generated picture of a celebrity saying something they never said. No evidence suggests deepfakes alter elections or fool voters. Deepfakes may cause voters to reject all political videos if they are constantly warned about their hazards and existence. Two online studies found that individuals couldn't identify genuine from deepfake videos. Deepfake warnings did not affect participants' ability to spot distorted information. Even though the films were real, participants were led to believe the cautions were fake. Due to the ubiquity of deepfakes, the course warned attendees to be skeptical of internet videos. Our results suggest that politicians and campaigns may influence actual videos using deepfake language, regardless of their credibility.
- **Industrialized disinformation: 2020 global inventory of organized social media manipulation:** Democracy is under jeopardy when social media influence polls. For the last four years, we have closely observed the social media campaigns launched by political parties and governments throughout the globe in an effort to sway public opinion, as well as their links to private companies. How resources, tactics, techniques, and expertise are evolving to influence public opinion is the focus of our 2020 study, which investigates computational propaganda in 81 countries. Three erroneous trends emerge from this year's data:
- **Socialbots: Human like by means of human control?** Opinions are swayed by anonymous social bots. The present election may have been impacted by online propaganda and social media. Strangely, the term "Social Bot" has varying definitions in different branches of science. The essay begins with a fair overview before moving on to detail the technical limitations of Twitter social bots. The advancements in Deep Learning and Big Data have not made bots capable of doing many jobs. We break down the keys to productive human relationships and the ropes to controlling bot capabilities.

3 METHODOLOGY

3.1 Proposed System

The suggested system leverages a DL-based approach to detect machine-generated tweets using a Convolutional Neural Network (CNN) combined with FastText embeddings. Unlike traditional methods that rely on hand-crafted features or feature-based machine learning, this approach enhances detection accuracy by capturing subword-level information, making it effective in handling complex language structures, sarcasm, and ambiguous text.

FastText embeddings improve linguistic adaptability by representing words based on their subword components, allowing the model to recognize rare and misspelled words effectively. This enhances the system's ability to detect bot-generated tweets across diverse linguistic and cultural contexts. The CNN architecture is optimized for feature extraction, providing a scalable and computationally

efficient framework suitable for real-time detection of deepfake tweets.

By integrating FastText embeddings with a CNN-based model, the system overcomes limitations of existing methods, such as high computational costs and poor performance in short-text detection. The proposed solution ensures improved accuracy, adaptability, and scalability, contributing to the fight against misinformation on social media platforms.

3.2 System Architecture

The proposed system follows a structured approach to detecting deepfake tweets using deep learning and natural language processing techniques. The process begins with data collection and preprocessing, where a dataset containing both human-generated and bot-generated tweets, such as the Tweepfake dataset, is gathered. The text undergoes cleaning operations, including tokenization, stopword removal, and normalization, to ensure consistency. FastText embeddings are then applied to convert text into numerical representations, capturing semantic and contextual information at the subword level.

Once the data is preprocessed, feature extraction is performed using FastText embeddings, which provide robust representations by considering word structures and linguistic patterns. Unlike traditional embeddings, FastText effectively handles misspellings, rare words, and subword components, making it particularly suitable for detecting manipulated content in tweets.

The DL model used for classification is based on a CNN architecture. The CNN processes the FastText-embedded tweets using multiple convolutional layers, which extract key features by identifying patterns in the text. Max-pooling layers help reduce dimensionality while retaining essential information, improving the model's efficiency. The extracted features are then passed through fully connected layers that apply activation functions to refine classification accuracy. A final softmax or sigmoid layer determines whether a tweet is human-generated or bot-generated based on learned features.

To ensure optimal performance, the evaluation and optimization phase involves comparing the CNN-FastText model against baseline machine learning techniques, such as TF, TF-IDF, and traditional FastText embeddings. Hyperparameter tuning is performed to enhance accuracy while maintaining computational efficiency.

3.3 Modules

3.3.1 Data Collection and Preprocessing

- Collects human-generated and bot-generated tweets from datasets like Tweepfake.
- Cleans text by removing stopwords, special characters, and performing tokenization.
- Converts text into numerical representations using FastText embeddings.

3.3.2 Feature Extraction

- Uses FastText embeddings to capture subword-level and contextual information.
- Enhances handling of misspellings, rare words, and linguistic variations.

3.3.3 Deep Learning Model: CNN Architecture

- Processes extracted features using a Convolutional Neural Network (CNN).
- Uses convolutional layers to detect patterns and relationships in text.
- Applies max pooling to retain key information while reducing dimensionality.

3.3.4 Classification and Prediction

- Fully connected layers process extracted features for classification.
- Uses a softmax or sigmoid layer to classify tweets as human-generated or bot-generated.

3.3.5 Model Evaluation and Optimization

- Compares CNN-FastText performance with baseline models (TF, TF-IDF, traditional embeddings).
- Tunes hyperparameters to improve accuracy and efficiency.

3.3.6 Real-Time Detection and Deployment

- Ensures scalability for large-scale tweet classification.
- Integrates the model into social media monitoring tools for real-time detection.

3.4 Algorithms

CNN: The Convolutional Neural Network (CNN) Algorithm processes the FastText embeddings to

extract meaningful patterns from the text. CNNs use convolutional layers to detect relationships between words, followed by max-pooling layers that reduce dimensionality while preserving essential features. The extracted features are passed through fully connected layers, which refine the classification process and improve accuracy.

FastText Embeddings Algorithm: The FastText Embeddings Algorithm is used to convert words into vector representations while capturing subword-level information. Unlike traditional word embeddings, FastText breaks words into character n-grams, enabling better handling of rare words, misspellings, and linguistic variations. This approach ensures richer contextual representation, improving the model's ability to differentiate between human-generated and machine-generated tweets.

Hyperparameter Tuning Algorithm: To further enhance performance, the Hyperparameter Tuning Algorithm is employed. This involves optimizing key parameters such as learning rate, batch size, and convolutional filter size to improve detection accuracy and computational efficiency. By fine-tuning these parameters, the model achieves better generalization and scalability for real-time deepfake tweet detection.

4 EXPERIMENTAL RESULTS

Accuracy: How well a test can differentiate between healthy and sick individuals is a good indicator of its reliability. Compare the number of true positives and negatives to get the reliability of the test. Following mathematics:

$$Accuracy = TP + TN / (TP + TN + FP + FN) \quad (1)$$

Precision: The accuracy rate of a classification or number of positive cases is known as precision. Accuracy is determined by applying the following formula:

$$Precision = TP / (TP + FP) \quad (2)$$

Recall: The recall of a model is a measure of its capacity to identify all occurrences of a relevant machine learning class. A model's ability to detect class instances is shown by percentage of correctly anticipated positive observations relative to total positives.

$$Recall = TP / (TP + FN) \quad (3)$$

Table 1: Evaluation metircs.

Algorithm Name	Accuracy	Precision	Recall	F1 Score
Navie Bayes	56	61.191	59.078	54.844
Logistic Regression	59.5	59.037	59.134	59.048
Decision Tree	58.5	57.5	57.323	57.338
Random Forest	58.5	57.655	57.588	57.608
Gradient Boosting	51	55.507	54.256	49.359
Propose CNN	87.545	87.707	87.525	87.527
Extension Hybrid CNN	94.089	94.193	94.014	94.079

F1-Score: A high F1 score indicates that a machine learning model is accurate. Improving model accuracy by integrating recall and precision. How often a model gets a dataset prediction right is measured by the accuracy statistic.

$$F1\ Score = 2 / ((1 / precision) + (1 / recall)) \quad (4)$$

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (5)$$

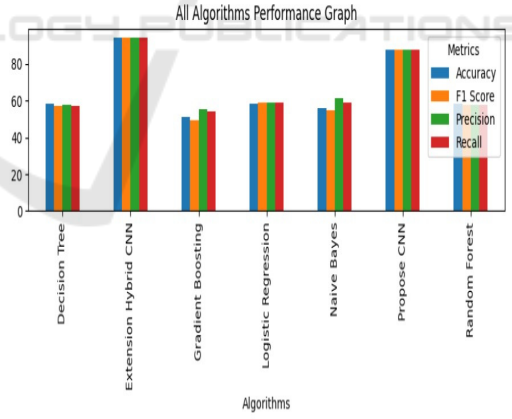

Figure 1: Accuracy graph.

5 CONCLUSIONS

This research presents a deep learning-based approach for detecting machine-generated tweets using CNN architecture with FastText embeddings. The proposed system effectively differentiates between human-generated and bot-generated tweets

by leveraging subword-level text representations and deep feature extraction techniques. Experimental data shows that the CNN model outperforms standard machine learning methods in accuracy, durability, and scalability. By addressing the challenges of misinformation and manipulated content on social media, this study contributes to the development of reliable tools for safeguarding online discourse and ensuring the integrity of digital interactions. Table 1 and figure 1 shows the results obtained.

6 FUTURE SCOPE

The proposed deep learning model for detecting machine-generated tweets can be extended and enhanced in several ways. Future research can focus on improving detection accuracy by incorporating transformer-based architectures like BERT or GPT for more advanced contextual understanding. Additionally, expanding the dataset to include multilingual tweets will improve the model's adaptability to diverse linguistic patterns.

Integrating real-time detection mechanisms with social media platforms can help in the proactive identification of deepfake content, reducing misinformation spread. Further, combining textual analysis with multimodal data such as images and videos can enhance the detection of complex deepfake content. The system can also be extended to detect evolving AI-generated text by continuously updating training data with the latest generative models.

REFERENCES

C.Grimme, M.Preuss, L.Adam, and H.Trautmann, ‘‘Socialbots: Human like by means of human control?’’ Big Data, vol. 5, no. 4, pp. 279–293, Dec. 2017.

H. Siddiqui, E. Healy, and A. Olmsted, ‘‘Bot or not,’’ in Proc. 12th Int. Conf. Internet Technol. Secured Trans. (ICITST), Dec. 2017, pp. 462–463.

J. P. Verma and S. Agrawal, ‘‘Big data analytics: Challenges and applica tions for text, audio, video, and social media data,’’ Int. J. Soft Comput., Artif. Intell. Appl., vol. 5, no. 1, pp. 41–51, Feb. 2016.

J. Ternovski, J. Kalla, and P. M. Aronow, ‘‘Deepfake warnings for political videos increase disbelief but do not improve discernment: Evidence from two experiments,’’ Ph.D. dissertation, Dept. Political Sci., Yale Univ., 2021.

J.-S. Lee and J. Hsiang, ‘‘Patent claim generation by fine-tuning OpenAI GPT-2,’’ World Pat. Inf., vol. 62, Sep. 2020, Art. no. 101983.

L. Beckman, ‘‘The inconsistent application of internet regulations and suggestions for the future,’’ Nova Law Rev., vol. 46, no. 2, p. 277, 2021, Art. no. 2.

M. Westerlund, ‘‘The emergence of deepfake technology: A review,’’ Technol. Innov. Manage. Rev., vol. 9, no. 11, pp. 39–52, Jan. 2019.

R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi, ‘‘Defending against neural fake news,’’ in Proc. 33rd Int. Conf. Neural Inf. Process. Syst. (NIPS), Dec. 2019, pp. 9054–9065, Art. no. 812.

R. Dale, ‘‘GPT-3: What’s it good for?’’ Natural Lang. Eng., vol. 27, no. 1, pp. 113–118, 2021.

S. Vosoughi, D. Roy, and S. Aral, ‘‘The spread of true and false news online,’’ Science, vol. 359, no. 6380, pp. 1146–1151, Mar. 2018.

S. Bradshaw, H. Bailey, and P. N. Howard, ‘‘Industrialized disinformation: 2020 global inventory of organized social media manipulation,’’ Comput. Propaganda Project Oxford Internet Inst., Univ. Oxford, Oxford, U.K., Tech. Rep., 2021.

W. D. Heaven, ‘‘A GPT-3 bot posted comments on Reddit for a week and no one noticed,’’ MIT Technol. Rev., Cambridge, MA, USA, Tech. Rep., Nov. 2020, p. 2020, vol. 24. [Online]. Available: www.technologyreview.com

X. Liu, Y. Zheng, Z. Du, M. Ding, Y. Qian, Z. Yang, and J. Tang, ‘‘GPT understands, too,’’ 2021, arXiv:2103.10385.