Detection of Machine-Generated Social Media Content Using Hybrid Deep Learning and Fast Text Representation

P.Arun Reddy[1], Md Rabbani[2], Chowla Manikya Kalyan[3], Masampally Venkat Ramana[4], Sriperumbuduru Shyam Sundar[5], Gudeti Vivek Vardhan[6]

[1] Assistant Professor, Department Of Computer Science And Engineering(AI& ML), Samskruthi College of Engineering And Technology , Kondapur(V), Ghatkesar(M), Medchal(D),Telangana

[2,3,4,5,6]BTech Students ,Department Of Computer Science And Engineering(AI& ML), Samskruthi College of Engineering And Technology , Kondapur(V), Ghatkesar(M), Medchal(D),Telangana

***Abstract*— The rapid progress in natural language generation has made it easier to create realistic text, which can be misused to influence opinions on social media platforms. Modern deep learning models are capable of generating content that closely resembles human-written text, making it difficult to distinguish between genuine and machine-generated messages. This creates serious concerns regarding misinformation and the spread of misleading content through automated accounts. To address this issue, this study focuses on detecting machine-generated tweets using a deep learning approach. A Convolutional Neural Network (CNN) model is designed and combined with FastText word embeddings to effectively capture semantic patterns in text data. The model is trained and evaluated using the Tweepfake dataset, which contains both human-written and bot-generated tweets. For comparison, several baseline machine learning models and alternative deep learning architectures such as LSTM and CNN-LSTM are also tested. The results show that the proposed CNN with FastText embeddings achieves an accuracy of around 93%, demonstrating its effectiveness in identifying synthetic tweets. This approach offers a reliable solution for improving the detection of fake content on social media.**

***Keywords*— Deep Learning, Deepfake Detection, Machine-Generated Text, Social Media Analysis, FastText Embeddings, Convolutional Neural Network (CNN), Tweet Classification, Natural Language Processing, Fake Content Detection**

I. INTRODUCTION

In recent years, social media platforms have become a major source of communication, where users share opinions, ideas, and information in the form of text, images, and videos. However, along with genuine users, there has been a rapid increase in automated accounts, commonly known as social bots, which are capable of mimicking human behavior [2][7]. These bots can generate and spread content at a large scale, often influencing public opinion. With the rise of deepfake technologies, it has become easier to create misleading content that appears authentic [3]. Such developments raise serious concerns about the reliability of information shared online. As a result, identifying and controlling fake content on social media has become an important area of research to ensure trust and transparency in digital communication.

Advancements in natural language processing and deep learning have significantly improved the ability of machines to generate human-like text. Modern language models such as GPT-based systems can produce highly realistic and context-aware content, making it difficult to distinguish between human-written and machine-generated text [12][8]. These models are often used by automated accounts to generate tweets and online posts that can mislead users or manipulate discussions. There have been real-world instances where such generated content has gone unnoticed by users, highlighting the seriousness of the issue [13]. The increasing sophistication of these models creates a strong need for effective detection systems that can accurately identify synthetic text and prevent the spread of misinformation on social media platforms.

Detecting machine-generated text is a challenging task, especially when dealing with short content like tweets. Unlike long articles, short texts provide limited context, making it harder to identify patterns that differentiate real and fake content. Additionally, techniques such as text manipulation, use of uncommon words, and stylistic variations further

increase the difficulty of detection [20]. Previous studies have mainly focused on longer documents, where detection is relatively easier. However, there is limited research on detecting deepfake content in short social media posts. The availability of well-labeled datasets is also a major concern, as human annotation may not always be accurate in identifying machine-generated text [19]. These challenges highlight the need for improved models that can effectively handle short and dynamic text data.

To address these challenges, researchers have explored various machine learning and deep learning approaches for detecting fake content. Traditional methods rely on features such as term frequency and statistical patterns, while advanced methods use neural networks to learn complex representations of text [1][9]. Deep learning models such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks have shown promising results in text classification tasks. These models are capable of capturing both semantic and contextual information from text data, making them suitable for detecting machine-generated content. In addition, embedding techniques such as FastText help in representing words in a meaningful way, improving the performance of classification models.

In this study, a deep learning-based approach is proposed to detect machine-generated tweets on social media. The model combines a CNN architecture with FastText word embeddings to effectively analyze short text data and identify patterns associated with synthetic content. The system is evaluated using a labeled dataset containing both human-written and machine-generated tweets, allowing a fair comparison of performance. The results are compared with other machine learning and deep learning models to demonstrate the effectiveness of the proposed approach. By improving the detection of fake text, this work aims to reduce the spread of misinformation and enhance the reliability of social media platforms. It also highlights the importance of developing advanced detection systems to keep pace with evolving text generation technologies [14][15].

II. LITERATURE SURVEY

Vosoughi et al., [2018] [5] examined how information spreads on social media and found that false news travels faster than true news. Their study showed that misleading content often reaches a larger audience because it appears more interesting and emotionally engaging. The authors analyzed real social media data and discovered that people tend to share surprising or unusual information more quickly. This makes it easier for fake news to influence public opinion. They also found that humans play a major role in spreading misinformation, not just automated bots. The research highlights the serious impact of false information in online platforms and the need for better detection systems. This work helps in understanding why fake content spreads easily and provides useful insights for building systems that can control misinformation and improve the reliability of social media content.

Zellers et al., [2019] [9] worked on detecting fake news generated by advanced neural networks. Their research focused on identifying text that is created by machines but appears similar to human writing. They developed methods to analyze patterns in generated text and compare them with real content. The study showed that as text generation models improve, it becomes more difficult to detect fake content. The authors also introduced tools to test how realistic machine-generated text can be. Their work highlights the growing challenge of distinguishing between real and artificial content. It also shows that detection systems must continue to evolve along with generation models. This research is important because it demonstrates how deep learning can be used not only to create content but also to identify and control fake information.

Dale et al., [2021] [12] discussed the abilities of modern language models like GPT-3 and how they can generate high-quality text. The study explained that these models can produce content that is very similar to human writing, making them useful for many applications. However, the author also pointed out that this technology can be misused to create fake or misleading content. Since the generated text is highly natural, it becomes difficult for people to identify whether it is written by a human or a machine. This creates challenges in maintaining trust on digital platforms. The study emphasizes the need for effective detection systems to handle such risks. It provides a clear understanding of both the advantages and limitations of advanced language models, making it valuable for research in detecting machine-generated text.

Fagni et al., [2021] [19] introduced a dataset called TweepFake, which is designed to study machine-generated tweets. The dataset includes both human-written and bot-generated tweets, making it useful for training detection models. The authors collected tweets created using different techniques and compared them with real user posts. Their work highlights the difficulty of detecting fake content in short text formats like tweets, where there is limited information available. The dataset helps researchers understand how machine-generated text differs from human-written text. It also supports the development of better models for identifying fake content. This

contribution is important because it provides a reliable dataset for testing and improving detection systems in social media environments.

Stiff et al., [2022] [20] studied methods to detect computer-generated fake content using machine learning techniques. Their research focused on identifying differences in writing patterns between human and machine-generated text. The authors explained that even though generated text looks natural, it still contains certain patterns that can be analyzed for detection. They also discussed how improving language models make detection more challenging over time. The study suggests that combining different features and using advanced models can improve the accuracy of detection systems. Their work provides useful insights into how fake content can be identified and controlled. It also highlights the importance of developing strong detection techniques to handle the increasing use of artificial intelligence in content generation.

III. DATASET DETAILS

The dataset used in this project consists of tweet data collected from social media platforms, which includes both human-written and machine-generated content. Each record in the dataset mainly contains the tweet text along with a corresponding class label that indicates whether the tweet is generated by a human or a bot. The human class represents genuine user-written content, while the bot class represents tweets created automatically using machine learning models. This dataset helps in understanding the differences between real and synthetic text. The data is organized in a structured format where each row represents a single tweet and its label, making it suitable for analysis and model training. Since social media text is usually short and informal, it presents unique challenges for classification. This dataset forms the foundation of the system, allowing the model to learn patterns and identify fake or machine-generated tweets effectively.

To make the dataset suitable for model training, several preprocessing steps are applied to improve its quality and consistency. Initially, the text data is cleaned by removing unnecessary elements such as special characters, stop words, and irrelevant symbols. After cleaning, the textual data is converted into numerical form using techniques like FastText, TF-IDF, or Term Frequency, which represent each word as a numeric vector. This transformation is important because machine learning models cannot process raw text directly. The dataset is then normalized and divided into training and testing sets, allowing the model to learn patterns from one portion and evaluate performance on unseen data. Shuffling is also applied to ensure a balanced distribution of both human and bot tweets. These preprocessing steps play a key role in improving model accuracy and ensuring reliable classification results in detecting deepfake tweets.

IV. PROPOSED METHODOLOGY

The proposed system follows a structured approach to detect machine-generated tweets using deep learning and text embedding techniques. Initially, the tweet dataset containing both human-written and bot-generated content is loaded into the system and analyzed. Data preprocessing is then performed to clean the text by removing stop words, special characters, and unnecessary symbols. After cleaning, the textual data is converted into numerical form using Fast Text embeddings, which represent words as meaningful vectors. The dataset is then normalized and shuffled to ensure balanced distribution of both classes. Following preprocessing, the data is divided into training and testing sets, allowing the model to learn patterns from the training data and evaluate its performance on unseen data.

Once the data is prepared, multiple machine learning and deep learning algorithms such as Naïve Bayes, Logistic Regression, Decision Tree, Random Forest, CNN, and LSTM are applied for classification. Each model is trained and tested using the same dataset, and their performance is evaluated using metrics such as accuracy, precision, recall, and F1-score. Among all models, the CNN combined with FastText embeddings shows the best performance. The system also presents results using graphs and tables for better comparison. Finally, the trained model is used to predict whether new tweet input is human-written or machine-generated.

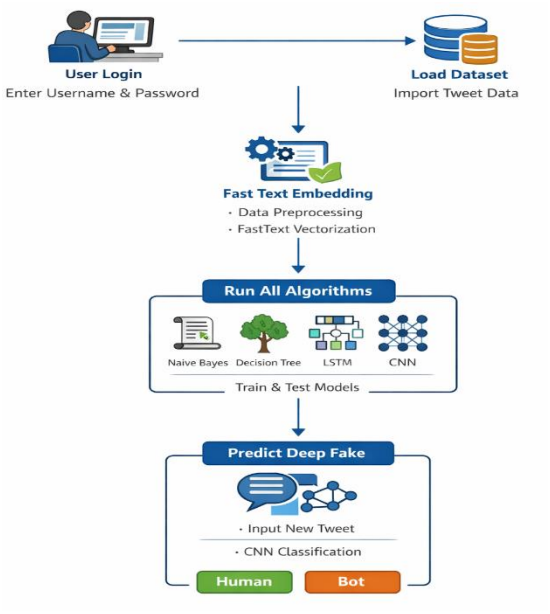

Figure [1]: Deep Fake Tweet Detection System

Figure [1] shows the workflow of the system. It starts with user login and dataset loading, followed by preprocessing and Fast Text embedding. Models are then trained and evaluated, and finally the CNN model predicts whether a tweet is human-written or bot-generated.

V. RESULT AND DISCUSSION

The experimental results of this project demonstrate the effectiveness of deep learning techniques in detecting machine-generated tweets. After preprocessing the dataset and converting the text into numerical vectors using FastText embeddings, multiple machine learning and deep learning models were trained and evaluated. Among all the models, the Convolutional Neural Network (CNN) achieved the highest accuracy of around 93%, showing strong performance in classifying tweets as human-written or bot-generated. Evaluation metrics such as precision, recall, and F1-score also indicated consistent and reliable results for the CNN model. In comparison, traditional machine learning models like Naïve Bayes, Logistic Regression, and Decision Tree showed lower accuracy, as they were less effective in capturing complex patterns in text data. The results highlight that deep learning models, especially CNN, are better suited for handling short and unstructured text like tweets. Performance comparisons presented through tables and graphs clearly show the superiority of the proposed approach. Overall, the system proves to be efficient and accurate in identifying deepfake tweets, helping to reduce the spread of misleading content on social media platforms.

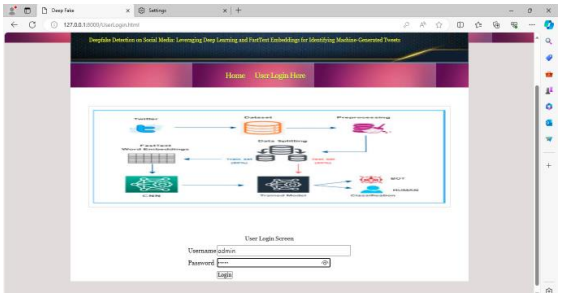

Figure [2]: User Login Interface

Figure [2] shows the login page of the system where users enter their username and password to access the application. This step ensures that only authorized users can use the system features. After successful login, the user is redirected to the main dashboard to continue further operations.

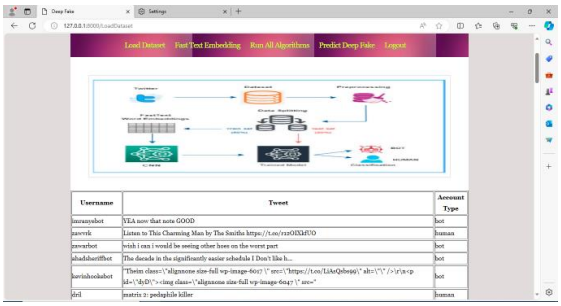

Figure [3]: Dataset Loading Interface

Figure [3] displays the dataset loading screen where the tweet dataset is imported into the system. It shows the structure of the dataset, including tweet text and corresponding labels such as human or bot. This step prepares the data for further processing and model training.

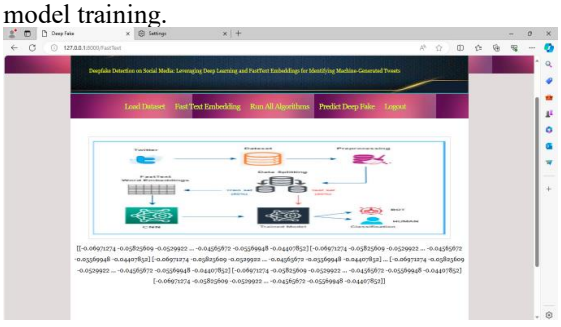

Figure [4]: Fast Text Embedding Output

Figure [4] illustrates the conversion of tweet text into numerical vectors using Fast Text embedding. The cleaned text data is transformed into numeric form so that machine learning models can process it. This step is important for capturing the meaning of words in the dataset.

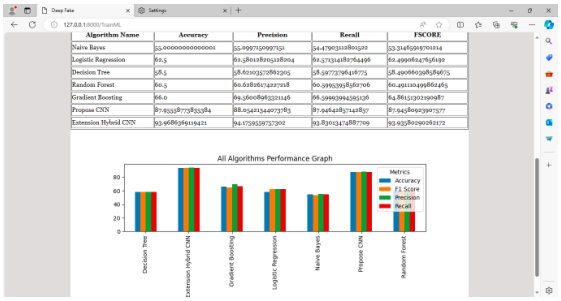

Figure [5]: Model Training and Performance Evaluation

Figure [5] shows the results of different machine learning and deep learning models after training. It includes evaluation metrics such as accuracy, precision, recall, and F1-score. This comparison helps identify the best-performing model for detecting deepfake tweets.

Figure [6]: Deep Fake Tweet Prediction Output

Figure [6] represents the final prediction output of the system. Users can enter a tweet, and the system classifies it as either human-written or bot-generated. The result is displayed clearly, showing the effectiveness of the model in identifying fake content.

DISCUSSION

The findings of this project highlight the importance of using suitable deep learning techniques for detecting machine-generated content on social media. Among all the models tested, the Convolutional Neural Network (CNN) performed better because of its ability to capture complex patterns and relationships within short text data like tweets. The use of FastText embeddings also played a significant role in improving performance, as it helped represent words in a meaningful numerical form. In comparison, traditional machine learning models such as Naïve Bayes and Logistic Regression showed lower accuracy, as they are less effective in handling the complexity of natural language data. Proper preprocessing steps, including text cleaning and vectorization, contributed greatly to the overall effectiveness of the system. Splitting the dataset into training and testing sets ensured fair evaluation of the models. Additionally, performance metrics such as accuracy, precision, recall, and F1-score provided a clear understanding of model behavior. Overall, the project demonstrates that combining deep learning with text embedding techniques can offer a reliable solution for detecting fake or bot-generated tweets on social media platforms.

VI. CONCLUSION

This project presents an effective approach for identifying machine-generated tweets using deep learning methods. The data was carefully prepared by cleaning the text, removing unnecessary elements, and converting it into numerical form with the help of FastText embeddings. Several machine learning and deep learning models were tested to understand their performance on tweet classification. Among them, the Convolutional Neural Network (CNN) showed the best results, as it was able to capture hidden patterns in short text more efficiently than other models. The evaluation metrics, including accuracy, precision, recall, and F1-score, confirmed the reliability of the proposed approach. The system is also designed in a way that allows users to input new tweets and receive predictions instantly. This makes it practical for real-time use. Overall, the project shows that combining deep learning with text embedding techniques can help in reducing the spread of fake content on social media. It also provides a strong base for future improvements and development of more advanced detection systems.

REFERENCES

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

[2] 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.

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

[4] 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.

[5] 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.

[6] 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.

[7] C. Grimme, M. Preuss, L. Adam, and H. Trautmann, ‘‘Social bots: Humanlike by means of human control?’’ Big Data, vol. 5, no. 4, pp. 279–293, Dec. 2017.

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

[9] 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.

[10] 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.

[11] 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.

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

[13] 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

[14] S. Gehrmann, H. Strobelt, and A. M. Rush, ‘‘GLTR: Statistical detection and visualization of generated text,’’ 2019, arXiv:1906.04043. VOLUME 11, 2023 95019 IEEE Transaction on Machine Learning,Volume:11,Issue Date:24.August.2023

[15] D. I. Adelani, H. Mai, F. Fang, H. H. Nguyen, J. Yamagishi, and I. Echizen, ‘‘Generating sentiment-preserving fake online reviews using neural language models and their human- and machine-based detection,’’ in Proc. 34th Int. Conf. Adv. Inf. Netw. Appl. (AINA). Cham, Switzerland: Springer, 2020, pp. 1341–1354.

[16] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi, ‘‘Grover—A state-of-the-art defense against neural fake news,’’ in Proc. Adv. Neural Inf. Process. Syst., vol. 32. Curran Associates, 2019. [Online]. Available: http://papers.nips.cc/paper/9106-defending-against neural-fake-news.pdf

[17] N. S. Keskar, B. McCann, L. R. Varshney, C. Xiong, and R. Socher, ‘‘CTRL: A conditional transformer language model for controllable generation,’’ 2019, arXiv:1909.05858.

[18] A. Uchendu, Z. Ma, T. Le, R. Zhang, and D. Lee, ‘‘TURINGBENCH: A benchmark environment for Turing test in the age of neural text generation,’’ 2021, arXiv:2109.13296.

[19] T. Fagni, F. Falchi, M. Gambini, A. Martella, and M. Tesconi, ‘‘TweepFake: About detecting deepfake tweets,’’ PLoS ONE, vol. 16, no. 5, May 2021, Art. no. e0251415.

[20] H. Stiff and F. Johansson, ‘‘Detecting computer-generated disinformation,’’ Int. J. Data Sci. Anal., vol. 13, no. 4, pp. 363–383, May 2022.