International Journal of Communication Networks and Information Security

ISSN: 2073-607X, 2076-0930

Volume 16 Issue 05 Year 2024

DEEPFAKE DETECTION ON SOCIAL MEDIA: LEVERAGING DEEP LEARNING AND FASTTEXT EMBEDDINGS FOR IDENTIFYING MACHINE-GENERATED TWEETS

S. Vinod Kumar [1], Santosh Taruni Annapa Reddy [2], S. Lakshmi devi [2], V. Kalyani [2]

[1]Assistant Professor, [2]UG Student, [1,2]School of Computer Science and Engineering

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

ARTICLE INFO **ABSTRACT**

Received: 10 Nov 2024

Accepted: 15 Dec 2024

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. 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,Twitter text

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.Before the advent of machine learning, detecting deepfakes and AI-generated content was a cumbersome process reliant on human moderators. Traditional approaches often involved extensive fact-checking by agencies, which could not keep pace with the sheer volume of online content. Keyword-based filtering systems frequently led to false positives or negatives due to their inability to understand context and nuance in language. Furthermore, the limited resources available for manual detection meant that misinformation could spread unchecked for extended periods. This inefficiency resulted in a significant gap in the timely identification of harmful content, underscoring the need for automated solutions.The increasing influence of social media in shaping public discourse highlights the critical need to combat misinformation. As deepfake technology becomes more accessible, the potential for misuse grows, creating a landscape where trust in digital content is increasingly jeopardized. This research is motivated by the need to develop an automated solution that can identify deepfakes effectively and efficiently, thus preserving the integrity of information shared online. By harnessing the capabilities of deep learning, the research seeks to create a system that not only enhances detection accuracy but also allows for real-time monitoring of social media platforms. Ultimately, the goal is to empower users and platforms to distinguish between genuine and manipulated content, thereby fostering a safer online environment. The necessity for this work is underscored by the rapid increase in the circulation of deepfake content across social media platforms. As users increasingly rely on digital media for information, the presence of manipulated content poses a significant risk to public perception and trust. This work seeks to fill a crucial gap by providing a real-time detection mechanism that can identify misleading tweets as they emerge. With the stakes so high, particularly during critical events like elections or public health crises, the ability to swiftly combat misinformation is paramount. Moreover, an automated system can operate continuously, monitoring vast amounts of data without the limitations of human resources.

2. LITERATURE SURVEY

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].

3. PROPOSED SYSTEM

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.

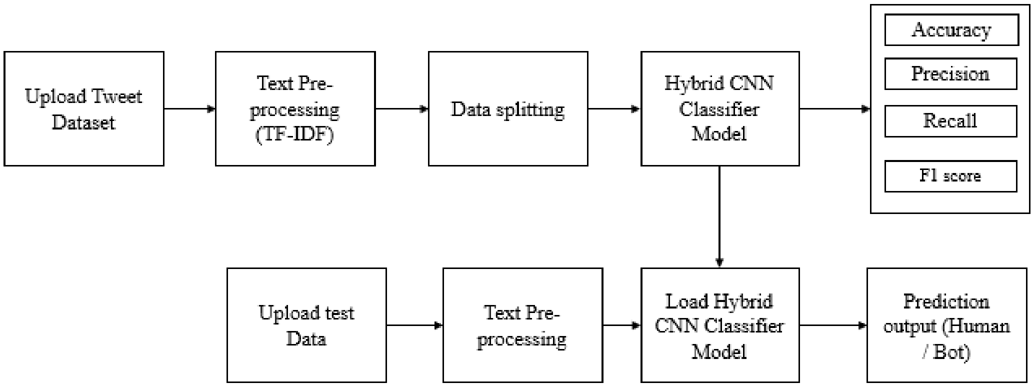
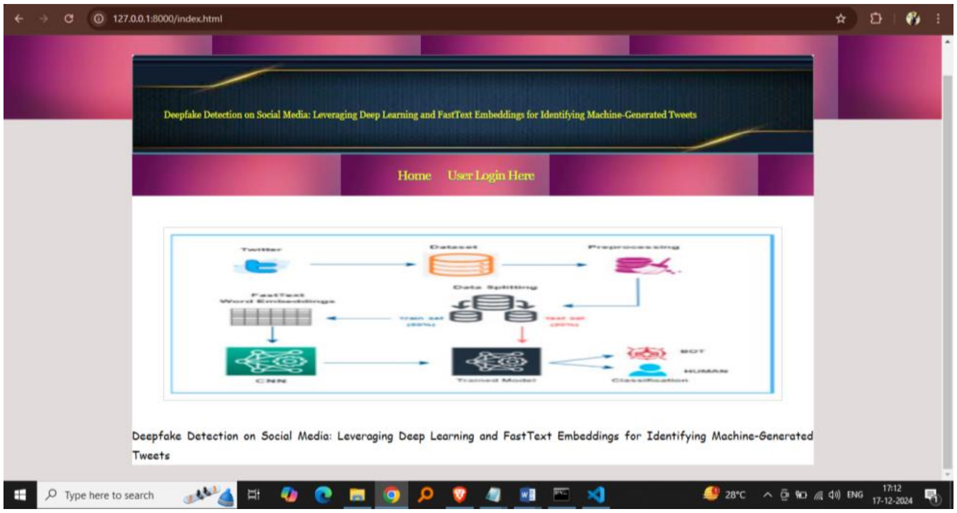

Upload Tweet Dataset
Text Pre-processing (TF-IDF)
Data splitting
Hybrid CNN Classifier Model
Accuracy
Precision
Recall
F1 score
Upload test Data
Text Pre-processing
Load Hybrid CNN Classifier Model
Prediction output (Human / Bot)

Figure 1 : Proposed Block Diagram

In the dataset is first divided into training and testing sets, typically using an 80-20 split. This ensures that the model can be trained on a substantial portion of the data while being evaluated on unseen data for generalization. Preprocessing involves cleaning the text by removing punctuation, special characters, and stop words to reduce noise. Tokenization is applied to break down text into smaller units (tokens). The data is then vectorized, transforming the textual data into numerical representations like word embeddings (FastText), making it suitable for model training.

4. RESULTS AND DISCUSSION

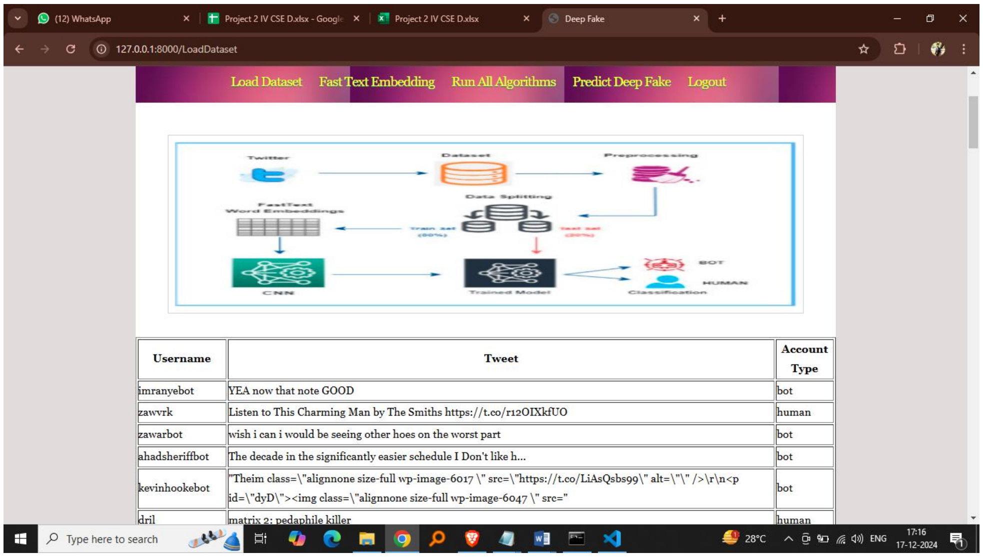
The necessary libraries are imported for text processing, machine learning, and Django utilities. Libraries like pandas, numpy, and sklearn are used for data handling and machine learning, while keras is used for deep learning models. Django utilities like render, messages, and HttpResponse are used to handle web requests and responses.



Deepfake Detection on Social Media: Leveraging Deep Learning and FastText Embeddings for Identifying Machine-Generated Tweets

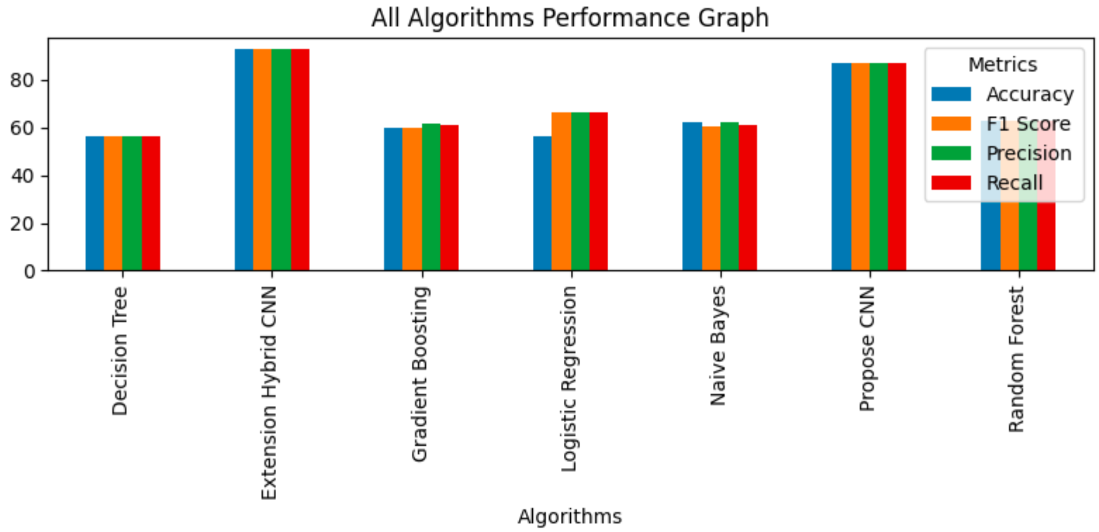
Deepfake Detection on Social Media: Leveraging Deep Learning and FastText Embeddings for Identifying Machine-Generated Tweets

Figure 2: Home

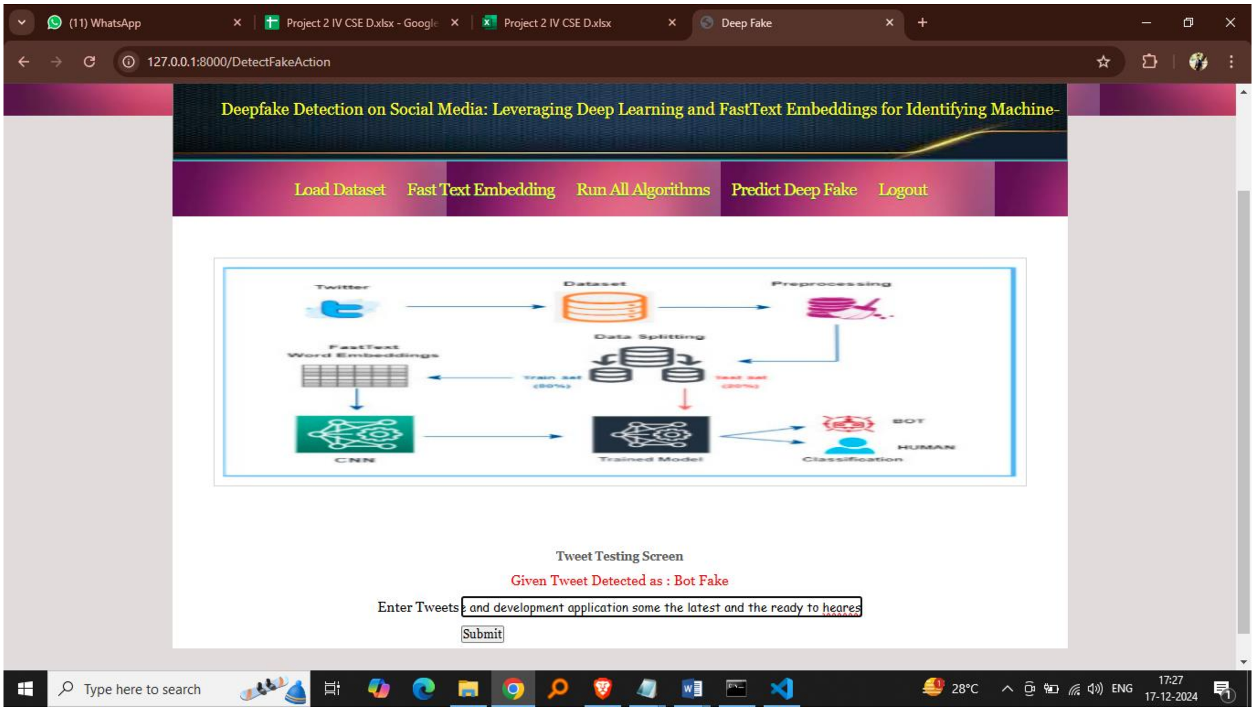


Username	Tweet	Account Type
imranyebot	YEA now that note GOOD	bot
zawvrk	Listen to This Charming Man by The Smiths https://t.co/r12OIXkfUO	human
zawarbot	wish i can i would be seeing other hoes on the worst part	bot
ahadsheriffbot	The decade in the significantly easier schedule I Don't like h...	bot
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

3: After Load Dataset



4: Comparison graph of all algorithm



Tweet Testing Screen

Given Tweet Detected as : Bot Fake

Enter Tweets and development application some the latest and the ready to heares

Submit

5: 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 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.While this research demonstrates the potential of deep learning and FastText embeddings in detecting machine-generated tweets, there are several areas for future enhancement and exploration. One key direction for future work is improving the model's ability to handle various types of deepfake content, such as videos, images, and audio. Expanding the framework to analyze multimodal content could help provide a more comprehensive solution for detecting deepfakes across different media formats.

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.