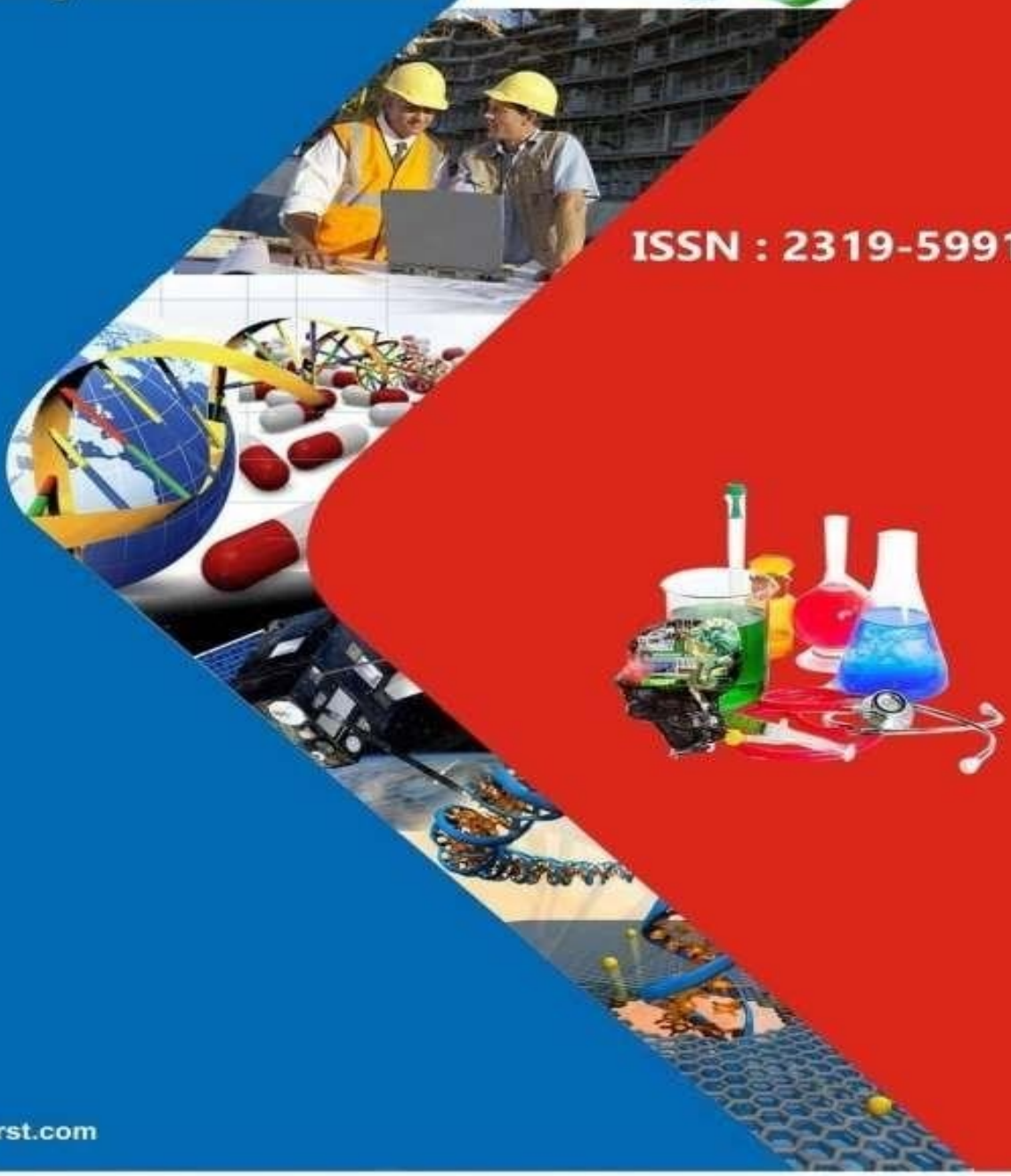
IJERST

International Journal of Engineering Research and Science & Technology

ISSN : 2319-5991



www.ijerst.com

Email: editor@ijerst.com or editor.ijerst@gmail.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 2, 2025

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

[1]MR. K.POORNACHANDRA RAO, [2]SADHU RAKESH, [3]ERIGINENI MANISAI, [4]ELURI SAI VARSHAN REDDY, [5]AVULA DINESH KUMAR REDDY

[1](ASSISTANT PROFESSOR), [2345]B.TECH STUDENTS

DEPARTMENT OF CSE, RISE KRISHNA SAI PRAKASAM GROUP OF INSTITUTIONS

ABSTRACT

Recent advancements in natural language production provide an additional tool to manipulate public opinion on social media. Furthermore, advancements in language modelling have significantly strengthened the generative capabilities of deep neural models, empowering them with enhanced skills for content generation. Consequently, text-generative models have become increasingly powerful allowing the adversaries to use these remarkable abilities to boost social bots, allowing them to generate realistic deepfake posts and influence the discourse among the general public. To address this problem, the development of reliable and accurate deepfake social media message-detecting methods is important. Under this consideration, current research addresses the identification of machine-generated text on social networks like Twitter. In this study, a simple deep learning model in combination with word embeddings is employed for the classification of tweets as human- generated or bot-generated using a publicly available Tweepfake dataset. A Naïve Bayes is devised, leveraging Fast Text word embeddings, to undertake the task of identifying deepfake tweets. To showcase the superior performance of the proposed method, this study employed several machine learning models as baseline methods for comparison. These baseline methods utilized various features, including Term Frequency, Term Frequency- Inverse Document Frequency, FastText, and FastText subword embeddings. Moreover, the performance of the proposed method is also compared against other deep learning models such SVM displaying the

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

effectiveness and highlighting its advantages in accurately addressing the task at hand. Experimental results indicate that the design of the Naïve Bayes architecture coupled with the utilization of FastText embeddings is suitable for efficient and effective classification of the tweet data with a superior 93% accuracy.

1.INTRODUCTION

Nowadays, social media is one of the most popular tools used by people to communicate with one another. It is also largely used by organizations to reach out to customers. In [1], it has been reported that there are 3.5 billion active social media users globally. Face book, Twitter, LinkedIn, and other social media networks are used by organizations to improve brand visibility and boost their sales. Twitter is one of the most popular social media platforms. It has 340 million active users who are allowed to communicate at a large scale and share their

opinions about different topics. Twitter could be targeted by various kinds of attacks. For example, a spear phishing attack in July 2020 led to the hijack of high pro_le Twitter accounts [2]. Also, fraudulent accounts could be created to impersonate legitimate users and organizations. Twitter can also be exploited by bot net, which is a set of malicious accounts that operate under a bot master, and are controlled by software programs rather than human users. The tweet-based social media bots pose serious security risks to Twitter users. These bots are used to spread fake contents, phishing links, and spams. Although they are not used as bots to launch DDOS attacks, they could be utilized as Command and Control (C&C) infrastructure to coordinate DDOS attack [3], [4]. They are capable of interacting with human accounts to deceive the users and hijack their accounts. These bots are also used as tools to launch large- scale manipulation campaigns to influence public opinions. According to a study [5], 52% of online traffic is generated by botnets, and the rest is produced by actual users. It is also worthy to note that some bots are found with over 350,000 fake followers. To deal with the above issues, there is a need to develop detection systems that can accurately distinguish between Twitter bot accounts and human accounts. Twitter data represent one of the examples of big data as around 500 million tweets are generated every day,

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

i.e., 6,000 tweets every second [6]. Big data analytics has been widely used in different fields [7]_[11] to

process large amount of data, discover hidden patterns, and find correlations among data points. Artificial intelligence techniques are increasingly leveraged by big data analysis. In particular, shallow (conventional) and deep learning techniques have received considerable attention from the academia and industry due to their success in dealing with heterogeneous and complex data, automatic learning of models, revealing unseen patterns, identifying dependencies, and

getting insights from analyzing data. Artificial intelligence has been extensively used by Twitter to determine tweet recommendations for users. In fact, deep neural networks are applied on Twitter data to determine the relevant content for users, and hence improve their experience on the platform [12]. Artificial intelligence has played an important role in fighting inappropriate content. In 2017, about 300,000 accounts were suspended and identified with the help of artificial intelligence tools rather than humans. This review aims at providing an overview of different tweet-based bot detection methods that use shallow and deep learning techniques to distinguish between human accounts and bot accounts. In particular, the main contributions of the paper are the following: 1) A taxonomy, which classifies the state-of-the-art on machine learning techniques for tweet-based bot detection, is presented. 2) A comprehensive review is presented on shallow and deep learning techniques for tweet-based bot detection, which covers the solutions up to year 2020. 3) The challenges and open issues related to tweet-based bot detection ar highlighted and discussed.

2.LITERATURE SURVEY

Giant Language Model Test Room (GLTR), is a visual tool that aids people in spotting deepfake texts [42]. The generated text is sampled word per word from a next token distribution; this distribution typically differs from the one that people unconsciously use when they write or speak (many sampling approaches may be employed, but the simplest option is to take the most probable token). In order to help individuals distinguish between human-written text samples and machine- generated ones, GLTR aims to display these statistical

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

linguistic distinctions. Authors in [15] carried out the sole study on identifying deep-fake social media messages on GPT-2-based Amazon evaluations. The Grover-based detector, GLTR, the RoBERTa-based detector from OpenAI, and a straightforward ensemble that combined these detectors using logistic regression at the score level were among the human-machine discriminators that were assessed. The aforementioned deepfake text detection techniques have two drawbacks: aside

from the study [15], they focused on creating news stories, which are lengthier than social media communications. Additionally, only one known adversarial generating model is often used to produce deepfake text samples (GPT-2 or

GROVER). We are not sure about the number and type of generative architectures

employed in a real-world scenario. Existing research in deepfake text detection includes methods like graph-based approach [45], feature-based approach [46], and deep learning models like

BiLSTM [47] and RoBERTa [19]. In a survey [48], the researchers offered a moreprofound insight into the creation and identification of deepfakes, the prevailingpatterns and progressions in this field, and the limitations of existing defence mechanisms. These studies have focused on creating and detecting news stories, which are typically longer than social media communications. This raises concerns about the generalizability of such methods to the specific challenges posed by short text on social media. Some studies [47], [49] used the PAN dataset which focuses on determining profiles of fake accounts. Others [46], [50], [51] used the Cresci dataset, which used profile features like tweet content, activity patterns, and network characteristics to find bot accounts. To aid the research community in identifying shorter deepfake texts created by different generating approaches, our Tweepfake dataset offers a collection of tweets generated by several generative models.

3. EXISTING SYSTEM

Deepfake technologies initially emerged in the realm of computer vision [31], [32],[33], advancing towards effective attempts at audio manipulation [34], [35] and

text synthesis [36]. In computer vision, deepfakes often involve face manipulation, including whole-facial synthesis, identity

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

swapping, attribute manipulation, and emotion switching [22]-as well as body reenactment [37]. Audio deepfakes, which have recently been used, generate spoken audio from a text corpus using the voices of several speakers after five seconds of listening [34]. The upgrading of the language models was made possible in 2017 because of the development of the self-attention mechanism and the transformer. Language modelling estimates the likelihood that a given sequence of words will appear in a sentence using various statistical and probabilistic methodologies. The succeeding transformer-based language models (GPT [38], BERT [39], GPT-2 [36], etc.) improved not only language-generating tasks but also natural language interpretation tasks. In 2019, Radord et al. [36] created GPT-2, a pre-trained

language model that can create paragraphs of text that are coherent and human-like on their own with just one short sentence as input. The same year, authors [9] developed GROVER, a novel method for quickly and effectively learning and creating multi-field documents like journal articles. The conditional language model CTRL, which employs control codes to produce text with a particular style, content, and task-specific behaviour, was published shortly after [17]. Researchers [40] introduced OPTIMUS which included a variational autoencoder in the text production process. The GPT-2 research team conducted an internal detection study [41] using text samples generated by the GPT-2. First, they assessed a conventional machine- learning method that trains a logistic regression discriminator on TF-IDF unigram and bigram characteristics. Following that, they tested a simple zero-shot baseline using an overall probability threshold: a text excerpt is classified as machine- generated if, according to GPT-2, its likelihood is closer to the mean likelihood over all machine-generated texts than to the mean of human written texts.

DISADVANTAGES:

- In an existing system,the system never develops Deep learning models like CNN which can automatically learn significant features from text input.
- An existing systems not are capable of capturing hierarchical patterns, local

relationships, and long-term connections, allowing the model to extract usable representations from the incoming text and

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

by stacking multiple layers of CNN, dependencies of text cannot be captured.

3.1 PROPOSED SYSTEM

In the proposed framework, a labelled dataset is collected from a public repository. The collected dataset contains tweets from human and bot accounts. In order to simplify the text and enhance its quality, a series of preprocessing steps are employed to clean the tweets. The dataset is divided into 80:20 ratios for training and testing. The next step involves transforming the text into vectors using FastText word embedding. Subsequently, these vector representations are fed into the CNN model. The proposed methodology, which leverages FastText word embedding in conjunction with a 3-layered CNN, is employed for the training process. The efficacy of this approach is assessed through the utilization of four evaluation metrics: Accuracy, Precision, Recall,and F1-score. The fixed vocabulary size of transfer learning models can pose challenges when encountering out-of vocabulary terms. In contrast, the CNN model used in this study does not suffer from such limitations. It can effectively handle out-of- vocabulary terms without compromising performance, as it does not rely on pre- defined vocabularies. This flexibility allows the model to better adapt to diverse and evolving textual data. By incorporating CNN features and harnessing the power of Fast Text, the proposed model achieves higher accuracy and demonstrates better performance compared to the selected state-of-the-art approaches. These findings highlight the effectiveness and superiority of the proposed model in tackling the challenge of deepfake text detection. Overall, the results obtained from the comparison with existing studies validate the proficiency of the proposed model and establish its competitive advantage in the field of deepfake text detection.

ADVANTAGES

• Presenting a deep learning framework combined with word embeddings that

effectively identifies machine-generated text on social media platforms.

• Comprehensive evaluation of various machine learning and deep learning models

for tweet classification.

• Investigation of different feature extraction techniques for detecting deepfake

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 2, 2025

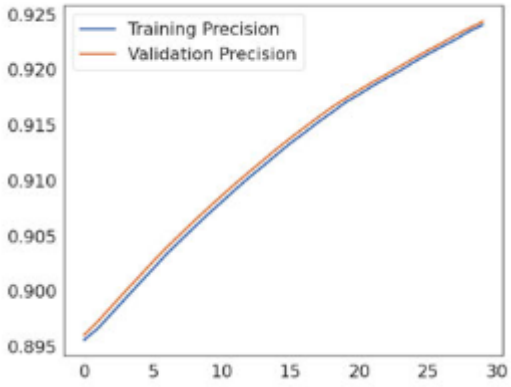
text, with a focus on short text prevalent on social media.

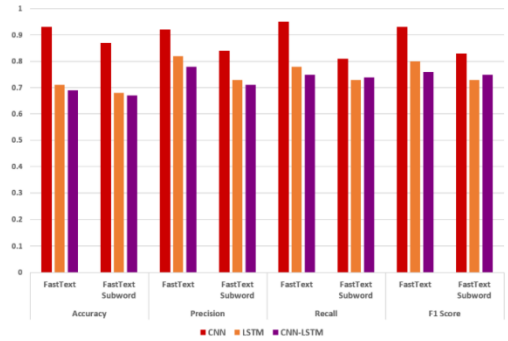
• Demonstrating the superiority of our proposed method, incorporating CNN with

Fast Text embeddings, over alternative models in accurately distinguishing

machine generated text in the dynamic social media environment.

4. OUTPUT SCREENS





5. CONCLUSIONS

Twitter is one of the most popular social media platforms that allows connecting people and helps organizations reaching out to customers. Tweet-based botnet can compromise Twitter and create malicious accounts to launch large-scale attacks and manipulation campaigns. In this review, we have focused on big data analytics, especially shallow and deep learning to fight against tweet-based botnets, and to accurately distinguish between human accounts and tweet-based bot accounts. We have discussed related surveys, and have

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

also provided a taxonomy that classifies the state-of-the-art tweet-based bot detection techniques up to 2020. In addition, the shallow and deep learning techniques are described for tweet-based botdetection, along with their performance results. Finally, we presented and discussed the open issues and future research challenges.

6. REFERENCE

[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.