Pegem Journal of Education and Instruction, Vol. 13, No. 4 , 2023 (pp. 427-436)

RESEARCH ARTICLE WWW.PEGEGOG.NET

DEEPFAKE DETECTION ON SOCIAL MEDIA TWEETS

[1]Mrs.M.Pragathi Reddy, [2]Husna sultana, [3]G . Deepika, [4]V. Lavanya

[1]Assistant Professor, Department of Computer Science and Engineering, Sridevi Women's Engineering College, Hyderabad, India. yeddula.pragathireddy@gmail.com

[2,3,4,]B.Tech Student, Department of Computer Science and Engineering, Sridevi Women's Engineering College, Hyderabad, India

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 straightforward 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 conventional Convolutional Neural Network (CNN) architecture is devised, leveraging FastText 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 as Long short-term memory (LSTM) and CNN-LSTM displaying the effectiveness and highlighting its advantages in accurately addressing the task at hand. Experimental results indicate that the streamlined design of the CNN architecture, coupled with the utilization of FastText embeddings, allowed for efficient and effective classification of the tweet data with a superior 93%
accuracy.

INTRODUCTION

Using social media, it is easier and faster to propagate false information with the aim of manipulating people's perceptions and opinions especially to build mistrust in a democratic country[5].Accounts with varying degrees of humanness like cyborg accounts to sockpuppets are used to achieve this goal [6]. On the other hand, fully automated social media accounts also known as social bots mimic human

Corresponding Author e-mail:
yeddula.pragathireddy@gmail.com

How to cite this article: 1Mrs.M.Pragathi Reddy, 2Husna sultana, 3G . Deepika, 4V. Lavanya. DEEPFAKE DETECTION ON SOCIAL MEDIA TWEETS.Pegem Journal of Education and Instruction, Vol. 13, No. 4, 2023, 427-436.**Source of support:** Nil **Conflicts of Interest:** None.

DOI: 10.48047/pegegog.13.04.50

Received: 12.10.2023

Accepted: 22.11.2023 **Published:** 24.12.2023

behaviour [7]. Particularly, the widespread use of bots and recent developments in natural language-based generative models, such as the GPT [8] and Grover [9], give the adversary a means to propagate false information more convincingly. The Net Neutrality case in 2017 serves as an illustrative example: millions of duplicated comments played a significant role in the Commission's decision to repeal [10]. The issue needs to be addressed that simple text manipulation techniques may build false beliefs and what could be the impact of more powerful transformer based models. Recently, there have been instances of the use of GPT-2 [11] and GPT-3 [12]: to generate tweets to test the generating skills and automatically make blog articles. A bot based on GPT-3 interacted with people on Reddit using the account "/u/thegentlemetre" to post comments to inquiries on /r/AskReddit [13]. Though most of the remarks made by the bot were harmless. Despite the fact that no harm has been done thus far, OpenAI should be concerned about the misuse of GPT-3 due to this occurrence. However, in order to protect genuine information and democracy on social media, it is important to create a sovereign detection system for machine generated texts, also known as deepfake text. In 2019, a generative model namely GPT-2 displayed enhanced textgenerating capabilities [12] which remained unrecognizable by the humans [14], [15]. Deepfake text on social media is mainly written by the GPT model; this may be due to the fact that the GPT model is better than Grover [16] and CTRL [17] at writing short text [18]. Consequently, it is highly challenging to detect machinegenerated text produced by GPT-2 than by RNN or other previously generated techniques [19]. To address this significant challenge, the present study endeavours to examine deepfakes generated by RNN, as well as GPT-2 and various other bots. Specifically, the study focuses on employing cutting-edge deepfake text detection techniques tailored to the dynamic social media environment. Stateof-the-art research works regarding deepfake text detection include [15], [19], [20]. Authors in [21] improved the detection of deepfake text generated by GPT 2. Deepfake detecting techniques are constantly being improved, including deepfake audio identification techniques [22], [23], deepfake video screening methods [24], and deepfake text detection techniques. Neural network models tend to learn characteristics of machine-generated text instead of discriminating humanwritten text from machine text [25]. Some techniques like replacing letters with homoglyphs and adding commonly misspelled words have made the machinegenerated text detection task more challenging [25]. In addition, previous studies mostly performed deepfake text detection in long text-like stories and news articles. The research claimed that it is easier to identify deepfakes in longer text [26]. The use of

cutting-edge detection methods on machine-generated text posted on social media is a less explored research area [26]. Text posted on social media is often short, especially on Twitter [27]. There is also a lack of properly labelled datasets containing human and machinegenerated short text in the research community [19]. Researchers in [28] and [29] used a tweet dataset containing tweets generated by a wide range of bots like cyborg, social bot, spam bot, and sock puppet [30]. However, their dataset was human labelled and research claimed that humans are unable to identify machinegenerated text. The authors in [19] provided a labelled dataset namely Tweep fake containing human text and machinegenerated text on Twitter using techniques such as RNN, LSTM, Markov and GPT-2. With the aim of investigating challenges faced in the detection of deepfake text, this study makes use of the same dataset. The dataset containing both bot-generated and human written tweets is used to evaluate the performance of the proposed method. This study employs various machine learning and deep learning models, including Decision Tree (DT), Logistic Regression (LR), AdaBoost Classifier (AC), Stochastic Gradient Descent Classifier (SGC), Random Forest (RF), Gradient Boosting Machine (GBM), Extra tree Classifier (ETC), Naive Bayes (NB), Convolutional Neural Network (CNN), Long Short Term Memory (LSTM), and CNN-LSTM, for tweet classification. Different feature extraction techniques, such as Term Frequency (TF), Term frequency-inverse document frequency (TF-IDF), Fast Text, and Fast Text sub words are also explored to compare their effectiveness in identifying machinegenerated text. This research provides the following contributions:

- 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 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.

LITERATURE REVIEW

IN "Big data analytics: Challenges and applications for text, audio, video, and social media data," All types of machine automated systems are generating large amount of data in different forms like statistical, text, audio, video, sensor, and bio-metric data that emerges the term

Big Data. In this paper we are discussing issues, challenges, and application of these types of Big Data with the consideration of big data dimensions. Here we are discussing social media data analytics, content based analytics, text data analytics, audio, and video data analytics their issues and expected application areas. It will motivate researchers to address these issues of storage, management, and retrieval of data known as Big Data. As well as the usages of Big Data analytics in India is also highlighted. The term big data is used to describe the growth and the availability of huge amount of structured and unstructured data. Big data which are beyond the ability of commonly used software tools to create, manage, and process data within a suitable time. Big data is important because the more data we collect the more accurate result we get and able to optimize business processes. The Big data is very important for business and society purpose. The data came from everywhere like sensors that used to gather climate information, available post or share data on the social media sites, video movie audio etc. This collection of data is called ―BIG DATA‖. Now a days this big data is used in multiple ways to grow business and to know the world [1,2, 15]. In most enterprise scenarios the data is too big or it moves too fast or it exceeds current processing capacity. Big data has the potential to help companies improve operations and make faster, more intelligent decisions. Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time. Big data is a set of techniques and technologies that require new forms of integration to uncover large hidden values from large datasets that are diverse, complex, and of a massive scale. Wal-Mart handles more than 1 million customer transaction every hour. Facebook handles 40 billion photos from its user base. Big data require some technology to efficiently process large quantities of data. It use some technology like, data fusion and integration, genetic algorithms, machine learning, and signal processing, simulation, natural language processing, time series Analytics and visualization [12,13,16]

IN "The emergence of deepfake technology: A review," Novel digital technologies make it increasingly difficult to distinguish between real and fake media. One of the most recent developments contributing to the problem is the emergence of deepfakes which are hyper-realistic videos that apply artificial intelligence (AI) to depict someone say and do things that never happened. Coupled with the reach and speed of social media, convincing deepfakes can quickly reach millions of people and have negative impacts on our society. While scholarly research on the topic is sparse, this study

analyzes 84 publicly available online news articles to examine what deepfakes are and who produces them, what the benefits and threats of deepfake technology are, what examples of deepfakes there are, and how to combat deepfakes. The results suggest that while deepfakes are a significant threat to our society, political system and business, they can be combatted via legislation and regulation, corporate policies and voluntary action, education and training, as well as the development of technology for deepfake detection, content authentication, and deepfake prevention. The study provides a comprehensive review of deepfakes and provides cybersecurity and AI entrepreneurs with business opportunities in fighting against media forgeries and fake news.

Existing System :

social media platforms were created for people to connect and share their opinions and ideas through texts, images, audio, and videos [1]. A bot is computer software that manages a fake account on social media by liking, sharing, and uploading posts that may be real or forged using techniques like gap-filling text, search-and- replace, and video editing or deepfake [2]. Deep learning is a part of machine learning that learns feature representation from input data. Deepfake is a combination of "deep learning" and "fake" and refers to artificial intelligence-generated multimedia (text, image, audio and video) that may be misleading [3]. Deepfake multimedia's creation and sharing on social media have already created problems in a number of fields such as politics [4] by deceiving viewers into thinking that they were created by humans. Using social media, it is easier and faster to propagate false information with the aim of manipulating people's perceptions and opinions especially to build mistrust in a democratic country [5]. Accounts with varying degrees of humanness like cyborg accounts to sockpuppets are used to achieve this goal [6]. On the other hand, fully automated social media accounts also known as social bots mimic human behaviour [7]. Particularly, the widespread use of bots and recent developments in natural language-based generative models, such as the GPT [8] and Grover [9], give the adversary a means to propagate false information more convincingly. The Net Neutrality case in 2017 serves as an illustrative example: millions of duplicated comments played a significant role in the Commission's decision to repeal [10]. The issue needs to be addressed that simple text manipulation techniques may build false

beliefs and what could be the impact of more powerful transformer-based models. Recently, there have been instances of the use of GPT-2 [11] and GPT-3 [12]: to generate tweets to test the generating skills and automatically make blog articles. A bot based on GPT-3 interacted with people on Reddit using the account "/u/thegentlemetre" to post comments to inquiries on /r/AskReddit [13]. Though most of the remarks made by the bot were harmless. Despite the fact that no harm has been done thus far, OpenAI should be concerned about the misuse of GPT-3 due to this occurrence. However, in order to protect genuine information and democracy on social media, it is important to create a sovereign detection system for machine-generated texts, also known as deepfake text.

Drawback in Existing System

- **Data Bias:** The effectiveness of deepfake detection models heavily relies on the quality and diversity of the training data. If the training data is biased or not representative of the entire range of deepfake techniques, the model may struggle to generalize to new and unseen types of deepfakes.
- **Generalization to New Deepfake Techniques:** Deep learning models may struggle to generalize to new and emerging deepfake techniques that were not present in the training data. Deepfake technology evolves rapidly, and models may become obsolete if they are not regularly updated with new data.
- **Explainability and Interpretability:** Deep learning models, especially complex ones, often lack transparency and interpretability. Understanding how the model reaches a particular decision can be challenging, making it difficult to trust and explain the detection results, which is important for user acceptance and legal considerations.
- **False Positives and Negatives:** Deepfake detection models may produce false positives (incorrectly flagging genuine content as deepfake) or false negatives (failing to detect actual deepfakes). Striking a balance between sensitivity and specificity is crucial to avoid the negative impact of both types of errors.

Proposed System

- **Data Preprocessing:** Clean and preprocess the collected data, including text normalization, removing irrelevant information, and handling missing or noisy data. Tokenize the text into

words or sub-word units for input to the deep learning model.

- **Feature Extraction with FastText Embeddings:** Utilize FastText embeddings to convert the textual content of tweets into dense vector representations. FastText embeddings capture semantic information and can handle out-ofvocabulary words, providing a robust representation for machine-generated text.
- **Deep Learning Model Architecture:** Design a deep learning model for tweet classification. This model should take the FastText embeddings as input and output a probability score indicating the likelihood of the tweet being machinegenerated. Consider using architectures like recurrent neural networks (RNNs), long short-term memory networks (LSTMs), or transformer models for capturing sequential dependencies in the text.
- **Integration with Social Media Platforms:** Develop an interface or integration with social media platforms to enable real-time or batch processing of tweets. Ensure compliance with the platforms' APIs and privacy policies. Consider providing feedback mechanisms for users to report false positives or negatives.

Algorithm

- **FastText Embeddings:** Utilize the FastText algorithm to generate word embeddings for the textual content of tweets. FastText is capable of capturing sub-word information, making it effective for handling misspellings, out-of-vocabulary words, and variations in language.
- **Explainable AI Techniques:** Incorporate techniques for explainability, such as attention mechanisms or LIME (Local Interpretable Model-agnostic Explanations), to provide insights into the model's decision-making process. Explainability is essential for building trust and understanding the model's behavior.
- **Evaluation Metrics:** Use appropriate evaluation metrics such as precision, recall, F1-score, and area under the Receiver Operating Characteristic (ROC) curve to assess the performance of your deepfake detection model. Consider the trade-off between false positives and false negatives based on the application's requirements.

Advantages

- **Robust Textual Representations:** FastText embeddings provide robust representations of

textual content by capturing semantic relationships and subword information. This can enhance the model's ability to understand the nuances of language, including misspellings, slang, and variations.

- **Adaptability to New Deepfake Techniques:**

  Deep learning models are capable of learning complex patterns from data, enabling them to adapt to new and emerging deepfake techniques. Regular updates and retraining can ensure the model remains effective against evolving threats.

- **Model Generalization:** The use of FastText embeddings and deep learning models enables the system to generalize well to new and unseen data. This is important for accurately detecting machine-generated content across a variety of contexts.

- **Continuous Improvement:**

  The system can be designed for continuous learning and improvement. Regular updates to the model based on new data and emerging trends in deepfake techniques contribute to the long-term effectiveness of the deepfake detection system.

CONCLUSION

Deep fake text detection is a critical and challenging task in the era of misinformation and manipulated content. This study aimed to address this challenge by proposing an approach for deep fake text detection and evaluating its effectiveness. A dataset containing tweets of bots and humans is used for analysis by applying several machine learning and deep learning models along with feature engineering techniques. Well-known feature extraction techniques: Tf and TFIDF and word embedding techniques: Fast Text and Fast Text sub words are used. By leveraging a combination of techniques such as CNN and Fast Text, the proposed approach demonstrated promising results with a 0.93 accuracy score in accurately identifying deep fake text. Furthermore, the results of the proposed approach are compared with other state-of the- art transfer learning models from previous literature. Overall, the adoption of a CNN model structure in this study shows its superiority in terms of simplicity, computational efficiency, and handling out-of-vocabulary terms. These advantages make the proposed approach a compelling option for text detection tasks, demonstrating that sophisticated performance can be achieved without the need for complex and time-consuming transfer learning models. The findings of this study contribute to advancing the field of deep fake detection and provide valuable insights for future research and practical applications. As social media

continues to play a significant role in shaping public opinion, the development of robust deep fake text detection techniques is imperative to safeguard genuine information and preserve the integrity of democratic processes. The challenges and opportunities in the emerging trend of quantum machine learning are highlighted in [76] and the quantum approach to detect deep fake text is presented in [77]. In future research, the quantum NLP and other cutting-edge methodologies will be applied for more sophisticated and efficient detection systems, to fight against the spread of misinformation and deceptive content on social media platforms.

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 GPT2,’’ *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*.

[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 machinebased detection,’’ in *Proc. 34th Int. Conf. Adv. Inf. Netw. Appl. (AINA)*. Cham, Switzerland: Springer, 2020, pp. 1341– 1354.