ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940
www.ijasem.org
Vol 18, Issue.2 April 2024

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

Kalva Sabitha [1], Kayita Pavani [2],
Kuchipudi Bhavani [3], Binde Sai Chandana[4],
Dr. S. Sri Bindu[5]
[1,2,3,4] UG Student, Dept. of ECE, CMR Institute of Technology, Hyderabad
[5]Associate Professor, Dept. of ECE,
CMR Institute of Technology, Hyderabad

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 conventional Convolutional Neural Network (CNN) architecture 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 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 design of the CNN architecture coupled with the utilization of FastText embeddings is suitable for efficient and effective classification of the tweet data with a superior 93% accuracy.

INTRODUCTION

ISSN2454-9940
www.ijasem.org
Vol 18, Issue.2 April 2024

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.

In 2019, a generative model namely GPT-2 displayed enhanced text-generating 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 machine-generated 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. State-of-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.

ISSN2454-9940
www.ijasem.org
Vol 18, Issue.2 April 2024

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 human-written text from machine text [25]. Some techniques like replacing letters with homoglyphs and adding commonly misspelled words have made the machine-generated 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 machine-generated 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 machine-generated text. The authors in [19] provided a labelled dataset namely Tweep fake containing human text and machine-generated 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 machine-generated text. This research provides the following contributions:

- Presenting a deep learning framework combined with word embeddings that

ISSN2454-9940
www.ijasem.org
Vol 18, Issue.2 April 2024

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

"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 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

ISSN2454-9940
www.ijasem.org
Vol 18, Issue.2 April 2024

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.

"Deepfake warnings for political videos increase disbelief but do not improve discernment: Evidence from two experiments,"

The rapid advancement of 'deepfake' video technology— which uses deep learning artificial intelligence algorithms to create fake videos that look real—has given urgency to the question of how policymakers and technology companies should moderate inauthentic content. We conduct an experiment to measure people's alertness to and ability to detect a high-quality deepfake among a set of videos. First, we find that in a natural setting with no content warnings, individuals who are exposed to a deepfake video of neutral content are no more likely to detect anything out of the ordinary (32.9%) compared to a control group who viewed only authentic videos (34.1%). Second, we find that when individuals are given a warning that at least one video in a set of five is a deepfake, only 21.6% of respondents correctly identify the deepfake as the only inauthentic video, while the remainder erroneously select at least one genuine video as a deepfake.

"The spread of true and false news online,"

The intentional and non-intentional use of social media platforms resulting in digital wildfires of misinformation has increased significantly over the last few years. However, the factors that influence this rapid spread in the online space remain largely unknown. We study how believability and intention to share information are influenced by multiple factors in addition to confirmation bias. We conducted an experiment where a mix of true and false articles were evaluated by study participants. Using hierarchical linear modelling to analyze our data, we found that in addition to confirmation bias, believability is influenced by source endorser credibility and argument quality, both of which are moderated by the type of information – true or false. Source likeability also had a positive main effect on believability. After controlling for belief and confirmation bias, intention to share information was affected by source endorser credibility and information source likeability.

"Social bots: Humanlike by means of human control?"

ISSN2454-9940
www.ijasem.org
Vol 18, Issue.2 April 2024

Social bots are currently regarded an influential but also somewhat mysterious factor in public discourse and opinion making. They are considered to be capable of massively distributing propaganda in social and online media, and their application is even suspected to be partly responsible for recent election results. Astonishingly, the term social bot is not well defined and different scientific disciplines use divergent definitions. This work starts with a balanced definition attempt, before providing an overview of how social bots actually work (taking the example of Twitter) and what their current technical limitations are. Despite recent research progress in Deep Learning and Big Data, there are many activities bots cannot handle well. We then discuss how bot capabilities can be extended and controlled by integrating humans into the process and reason that this is currently the most promising way to realize meaningful interactions with other humans. This finally leads to the conclusion that hybridization is a challenge for current detection mechanisms and has to be handled with more sophisticated approaches to identify political propaganda distributed with social bots.

"GPT understands, too,"

Prompting a pretrained language model with natural language patterns has been proved effective for natural language understanding (NLU). However, our preliminary study reveals that manual discrete prompts often lead to unstable performance—e.g., changing a single word in the prompt might result in substantial performance drop. We propose a novel method P-Tuning that employs trainable continuous prompt embeddings in concatenation with discrete prompts. Empirically, P-Tuning not only stabilizes training by minimizing the gap between various discrete prompts, but also improves performance by a sizeable margin on a wide range of NLU tasks including LAMA and SuperGLUE. P-Tuning is generally effective for both frozen and tuned language models, under both the fully-supervised and few-shot settings.

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 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

ISSN2454-9940
www.ijasem.org
Vol 18, Issue.2 April 2024

several speakers after five seconds of listening [34].

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

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.

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 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.

CONCLUSION

Deepfake text detection is a critical and challenging task in the era of misinformation and manipulated content. This study aimed to address this challenge

ISSN2454-9940
www.ijasem.org
Vol 18, Issue.2 April 2024

by proposing an approach for deepfake 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 TF-IDF 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 deepfake text. Furthermore, the results of the proposed approach are compared with other state-ofthe-art transfer learning models from previous literature.

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] [9] Karne, R. K. ., & Sreeja, T. K. . (2023). PMLC- Predictions of Mobility and Transmission in a Lane-Based Cluster VANET Validated on Machine Learning. International Journal on Recent and Innovation Trends in Computing and Communication, 11(5s), 477–

ISSN2454-9940
www.ijasem.org
Vol 18, Issue.2 April 2024

483.
https://doi.org/10.17762/ijritcc.v11i5s.7109

[12] [10] Radha Krishna Karne and Dr. T. K. Sreeja (2022), A Novel Approach for Dynamic Stable Clustering in VANET Using Deep Learning (LSTM) Model. IJEER 10(4), 1092-1098. DOI: 10.37391/IJEER.100454.

[13] [11] Reddy, Kallem Niranjan, and Pappu Venkata Yasoda Jayasree. "Low Power Strain and Dimension Aware SRAM Cell Design Using a New Tunnel FET and Domino Independent Logic." International Journal of Intelligent Engineering & Systems 11, no. 4 (2018).

[14] [12] Reddy, K. Niranjan, and P. V. Y. Jayasree. "Design of a Dual Doping Less Double Gate Tfet and Its Material Optimization Analysis on a 6t Sram Cells."

[15] [13] Reddy, K. Niranjan, and P. V. Y. Jayasree. "Low power process, voltage, and temperature (PVT) variations aware improved tunnel FET on 6T SRAM cells." Sustainable Computing: Informatics and Systems 21 (2019): 143-153.

[16] [14] Reddy, K. Niranjan, and P. V. Y. Jayasree. "Survey on improvement of PVT aware variations in tunnel FET on SRAM cells." In 2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC), pp. 703-705. IEEE, 2017

[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,andM. 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.