DEEP FAKE ON SOCIAL MEDIA: FOR IDENTIFYING MACHINE GENERATED TWEETS

G. NARSIMHA REDDY, P. ROHIT, K. NAGARAJU, J. AKHIL, B. RAMBABU

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 adversaries to use these remarkable abilities to boost social bots, allowing them to generate realistic deepfake posts and influence the discourse among the 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 was employed for the classification of tweets as human-generated or bot-generated using a publicly available Tweep fake 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, we employed several machine learning models as baseline methods for comparison. These baseline methods utilized various features, including term frequency, term frequency-inverse document frequency, Fast Text, and Fast Text sub word embeddings.

Keywords—Deepfake Detection, Machine-generated, Twitter, Social Media, GPT-3.

INTRODUCTION

In recent years, the rise of fake content on social media platforms has become a pressing concern, with machine-generated text posing a significant challenge. This proliferation of deepfake text, generated by advanced language models like GPT-2 and GPT-3, has raised alarms regarding its potential to deceive and manipulate public opinion. Detecting machine-generated text, especially in the fast-paced and dynamic environment of social media, has become imperative to safeguard against misinformation and maintain the integrity of online discourse. In this context, our study focuses on Deepfake Detection on Social Media, specifically leveraging deep learning techniques and Fast Text embeddings to identify machine-generated tweets. The prevalence of short-form content on platforms like Twitter presents unique challenges for detection algorithms, necessitating innovative approaches and robust methodologies. The primary objective of our research is to develop a framework capable of accurately distinguishing between human-written and machine- generated tweets. By employing a combination of machine learning and deep learning models, along with advanced feature extraction techniques, we aim to enhance the efficacy of detection mechanisms in the realm of social media. We conduct a thorough assessment of various machine learning and deep learning models for tweet classification, considering factors such as accuracy, efficiency, and scalability. We explore different methods for feature extraction, focusing on their effectiveness in identifying machine-generated text within the constraints of short-form content prevalent on social media platforms. Our approach integrates deep learning architectures, such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM), to harness the power of neural networks in detecting subtle patterns indicative of machine-generated text. We employ Fast Text embeddings, known for their ability to capture semantic relationships and contextual information, to enhance the representation of textual data and improve detection accuracy. Through empirical evaluations, we demonstrate the superiority of our proposed method in accurately detecting machine-generated tweets, thereby mitigating the spread of misinformation, and safeguarding the integrity of social media discourse. By addressing the challenges posed by deepfake text on social media and leveraging advanced technologies like deep learning and Fast Text embeddings, our research contributes to the development of robust detection mechanisms vital for maintaining trust and authenticity in online communication.

METHODOLOGY

1. **Data Collection and Preprocessing:** To ensure representativeness, we assemble balanced datasets of tweets produced by machines and humans. We ready the data for additional analysis after cleaning it up by eliminating noise and tokenizing the text.

2. **Feature extraction:** We extract significant characteristics from the text by utilizing both conventional techniques such as TF-IDF and FastText embeddings. Semantic linkages are captured by FastText embeddings, while word importance is highlighted by TF-IDF.

3. **Model Development:** We use both deep learning and machine learning models in our approach. We investigate many algorithms and customize them for tweet classification, including logistic regression, decision trees, and deep learning architectures like CNN and LSTM.

4. **Model Training and Hyperparameter Tuning:** Using the prepared datasets, models are trained. Grid search techniques are used to fine-tune hyperparameters, and k-fold cross-validation is used to validate the results. This guarantees both generalizability and optimal performance.

5. **Evaluation:** Using measurements like accuracy, precision, recall, F1 score, and AUC-ROC, performance evaluation entails a thorough examination. To find the optimal strategy, we evaluate the efficacy of several models and feature extraction strategies.

6. **Deployment and Monitoring:** The best-performing model is put into use, integrated into social networking sites for on-the-spot detection, and its performance is regularly tracked. Based on user comments and identified errors, we create a feedback loop for refinement.

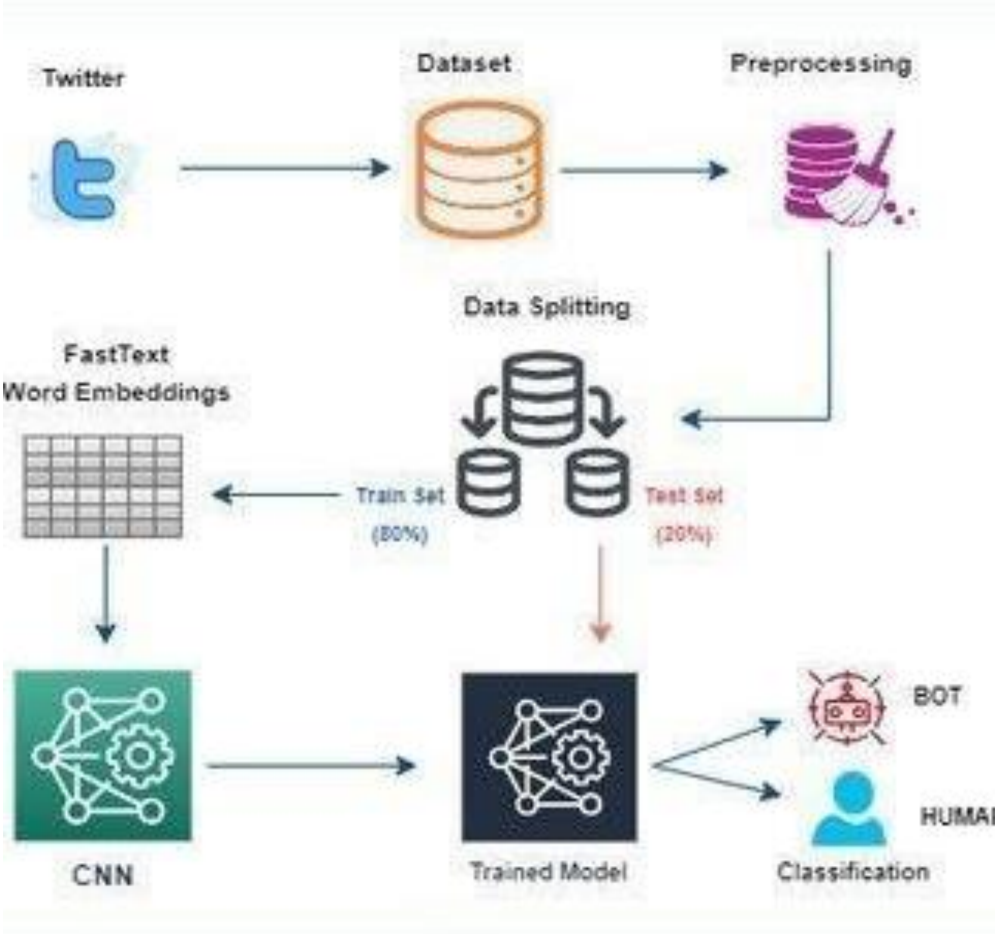

FIG-1 Architecture Diagram

HARDWARE REQUIREMENTS

- ✓ CPU (Central Processing Unit): Minimum: - Intel Core i5 or AMD Ryzen 5 - Recommended: Intel Core i7 or AMD Ryzen 7
- ✓ RAM: Minimum: - 16 GB - Recommended: 32 GB or higher
- ✓ Storage: - Minimum: 256 GB SSD - Recommended: 512 GB SSD or larger
- ✓ Network Connectivity: - Minimum: 100 Mbps Ethernet or Wi-Fi - Recommended: 1 Gbps Ethernet for data transfer.
- ✓ Cooling System: - Minimum: Standard cooling solution - Recommended: High-performance cooling system for GPU-intensive tasks.
- ✓ Redundancy and Backup (if required):- Minimum: RAID 1 for data redundancy - Recommended: Additional

backup servers and data replication solutions.

SOFTWARE REQUIREMENTS

- ✓ Operating System: Ubuntu or Windows.
- ✓ Programming Language: Python 3.8 or later for the latest features and support.
- ✓ Deep Learning Framework: TensorFlow 2.5 or PyTorch 1.8 for improved performance and features.
- ✓ Libraries: OpenCV 4.5 or later, NumPy, and pandas for up-to-date functionalities.
- ✓ Integrated Development Environment (IDE): Visual Studio Code 1.60 or later, PyCharm, etc.
- ✓ Version Control: Git 2.30 or later for efficient code tracking and collaboration.

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 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 is 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 deepfake 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 deepfake text detection techniques is imperative to safeguard genuine information and preserve the integrity of democratic processes. 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] Jai Prakash Verma, Smita Agrawal, Bankim Patel, and Atul Patel. Big data analytics: Challenges and applications for text, audio, video, and social media data. International Journal on Soft Computing, Artificial Intelligence and Applications (IJSCAI), 5(1):41–51, 2016.

[2] Husna Siddiqui, Elizabeth Healy, and Aspen Olmsted. Bot or not. In 2017 12th international conference for internet technology and secured transactions (ICITST), pages 462–463. IEEE, 2017.

[3] Mika Westerlund. The emergence of deepfake technology: A review. Technology innovation management review, 9(11), 2019.

[4] John Ternovski, Joshua Kalla, and Peter M Aronow. Deepfake warnings for political videos increase disbelief but do not improve discernment: Evidence from two experiments. 2021.

[5] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. science, 359(6380):1146–1151, 2018.

[6] Samantha Bradshaw, Hannah Bailey, and Philip N Howard. Industrialized disinformation: 2020 global inventory of organized social media manipulation. Computational Propaganda Project at the Oxford Internet Institute, 2021.

[7] Christian Grimme, Mike Preuss, Lena Adam, and Heike Trautmann. Social bots: Human-like by means of human control? Big data, 5(4):279– 293, 2017.

[8] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. Gpt understands, too. arXiv preprint arXiv:2103.10385, 2021.

[9] R. Dinesh Kumar, Prof. Dr.J.Suganthi (2018); A Research Survey on Sanskrit Offline Handwritten Character Recognition; Int J Sci Res Publ 3(1) (ISSN: 2250-3153)

[10] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. Advances in neural information processing systems, 32, 2019.

[11] R. Dinesh Kumar, E. Golden Julie, Y. Harold Robinson, S. Vimal, Gaurav Dhiman, Murugesh Veerasamy, "Deep Convolutional Nets Learning Classification for Artistic Style Transfer", Scientific Programming, vol. 2022, Article ID 2038740, 9 pages, 2022.

[12] Logan Beckman. The inconsistent application of internet regulations and suggestions for the future. Nova L. Rev., 46:277, 2021.

[13] Jieh-Sheng Lee and Jieh Hsiang. Patent claim generation by fine-tuning openai gpt-2. World Information, 62:101983, 2020.

[14] Dinesh Kumar, R., Kalimuthu, M., Jayaram, B. (2022). Character Recognition System Using CNN for Sanskrit Text. In: Satyanarayana, C., Gao, XZ., Ting, CY., Muppalaneni, N.B. (eds) Proceedings of the International Conference on Computer Vision, High Performance Computing, Smart Devices and Networks. Advanced Technologies and Societal Change. Springer, Singapore.