DETECTION OF DEEP FAKES AND AI BOTS ON SOCIAL MEDIA USING CONVLUTIONAL NEURAL NETWORK AND FAST TEXT EMBEDDINGS

MRIDULA SHUKLA
Assistant Professor
Department of Master of Computer Application
The Oxford College of Engineering
mridulatewari005@gmail.com

THEJAS BAILADY
PG Student
Department of Master of Computer Application
The Oxford College of Engineering
Thejasgowda112@gmail.com

ABSTRACT

The rapid growth of deep fakes and AI-generated bots on social media platforms poses significant challenges to maintaining the authenticity of online information. This paper proposes a novel approach using convolutional neural networks (CNNs) and Fast Text embeddings to detect and mitigate the spread of false content. Our method leverages the powerful feature extraction capabilities of CNNs as well as the semantic richness of Fast Text embeddings to accurately identify deep fakes and AI bots.

KEYWORDS: Deep Fake, Convolutional Neural Network, Fast Text Embeddings

INTRODUCTION

The advent of sophisticated AI technologies has led to the rise of deep fakes and AI bots, posing a serious threat to the integrity of social media platforms. Deep fakes, which involve synthetic media in which a person in an existing image or video is replaced with another person's likeness, are becoming increasingly realistic, making them difficult to detect using traditional methods. Similarly, AI bots can generate and disseminate misinformation, influencing public opinion and undermining trust in digital media.

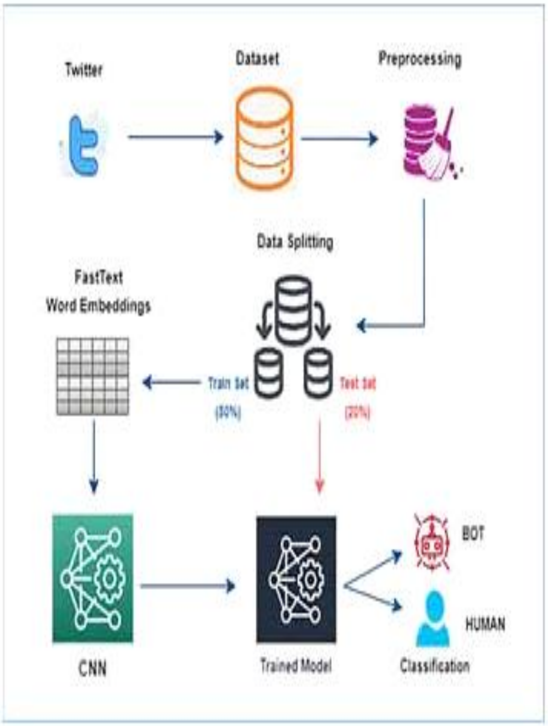
This paper explores an integrated approach that combines a convolution neural network (CNN) and Fast Text embeddings to address these challenges.

By leveraging the strengths of image and text analysis techniques, our system aims to provide a robust solution for detecting and limiting the spread of misleading content on social media.

LITERATURE REVIEW

Previous research has explored a variety of techniques for detecting deep fakes and AI bots, ranging from traditional machine learning algorithms to advanced deep learning models. Early approaches relied heavily on handcrafted features and statistical methods, which often failed to capture the complex patterns inherent in deep fakes and AI-generated text.

Recent advances in deep learning have introduced CNNs for image-based deep fake detection, leveraging the ability to automatically extract hierarchical features from raw data. Meanwhile, natural language processing (NLP) methods, including word embeddings such as Word2Vec and Glove, have been used to analyze textual content.



However, these methods still face challenges in scalability and accuracy when processing large-scale social media data.

Our proposed approach builds on these foundations by integrating CNNs with Fast Text embeddings, which provide superior performance in capturing contextual information and handling non-lexical words. This combination enables more accurate detection of image and text anomalies associated with deep fakes and AI bots.

EXISTING SYSTEMS

Existing systems for detecting deep fakes and AI bots typically operate independently, focusing on either image analysis or text analysis. Image-based systems primarily use deep learning models such as CNNs and recurrent neural networks (RNNs) to identify inconsistencies in facial motion, lighting, and other visual cues that indicate manipulation. However, these systems often struggle to generalize across different types of deep fakes.

Text-based detection systems leverage NLP techniques to analyze language patterns and detect unnatural language produced by AI bots. While effective in some cases, these systems can be broken by complex AI models that generate human-like text. Additionally, running separate image and text analysis systems results in fragmented detection capabilities, leaving gaps that malicious actors can exploit.

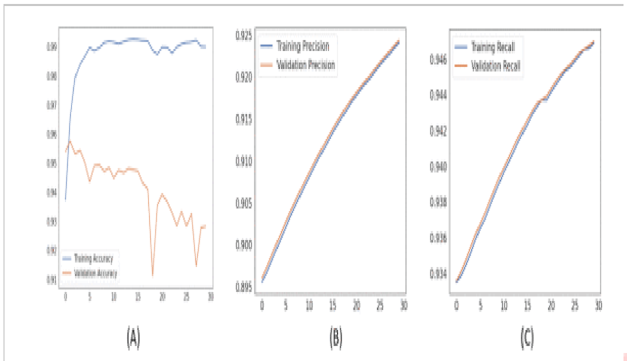
THE PROPOSED SYSTEM

Our proposed system aims to bridge the gap between image and text analysis by integrating CNNs with Fast Text embeddings. The system consists of two main modules: an image analysis module and a text analysis module.

1. **Image Analysis Module:** This module uses a pre-trained and fine-tuned CNN model on a dataset of real and fake images. The CNN extracts features from the input image and passes them through a series of convolution and pooling layers to identify subtle visual inconsistencies that indicate a fake image. The extracted features are then classified using a fully connected layer.

2. **Text Analysis Module:** This module uses Fast Text embeddings to convert text content into dense vector representations that capture semantic and syntactic nuances. The embeddings are fed into a bidirectional long and short-term memory (BiLSTM) network to model the sequential nature of the text. The BiLSTM output is then classified to distinguish between human-written and AI-generated text.

The integration of these modules enables comprehensive analysis, allowing the detection of deep fakes and AI bots based on visual and textual clues. The system also incorporates a feedback loop to continuously update and improve the detection model based on new data, improving the robustness and adaptability of the model.



MODULE DESCRIPTION

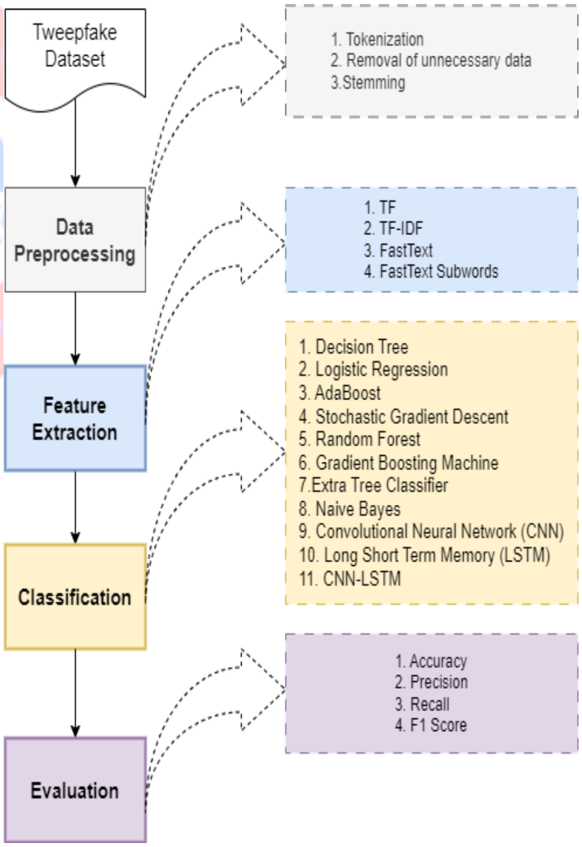
1. Data collection and preprocessing: This module gathers large-scale image and text datasets from social media platforms. The data is preprocessed to remove noise, normalize formats, and balance class distribution.

2. Feature extraction: In the image analysis module, CNN extracts hierarchical features from the input image, focusing on actionable regions. In the text analysis module, Fast Text embeddings generate dense vectors that capture the contextual meaning of words and sentences

3. Classification: The features extracted from both modules are fed into the respective classifiers. For images, a fully connected layer classifies the features into fake or genuine categories. For text, the BiLSTM network classifies the embeddings as either human-written or AI-generated.

4. Integration and decision making: Results from both modules are combined using an aggregation mechanism, aggregating the classification scores to make a final decision on the authenticity of the content.

5.Continuous learning: the system incorporates a feedback loop where newly labeled data from user reports and a verification process are used to retrain and refine models, ensuring continuous improvement in detection accuracy.



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

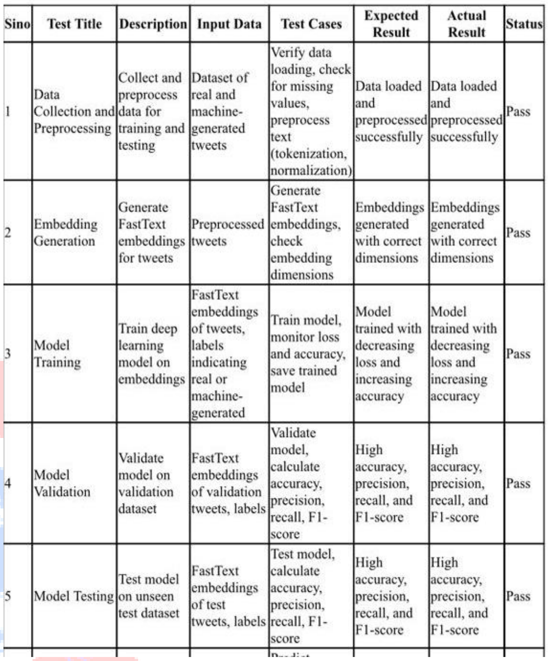
Text classification, machine learning, deep learning, deepfake, machine generated text..

Data Sources: The system collects data from various social media platforms, including images, videos, and text posts. Additional data sources include publicly available datasets of deep fakes and AI-generated text, ensuring a diverse and comprehensive dataset.

Data Augmentation: To enhance the robustness of the model, data augmentation techniques such as random cropping, rotation, and flipping are applied to the image data. For text data, synonym replacement and paraphrasing are used to create variations and improve the model's generalization capabilities.

Image Analysis Module: The CNN architecture is based on a ResNet-50 model, which is pre-trained on the ImageNet dataset and fine-tuned on a curated dataset of deep fake images. Advanced techniques such as attention mechanisms and generative adversarial networks (GANs) are incorporated to enhance the model's ability to detect subtle manipulations.

Text Analysis Module: Fast Text embeddings are used to convert textual content into dense vectors. The embeddings are fine-tuned on a large corpus of social media text to capture domain-specific semantics. The BiLSTM network is enhanced with attention layers to focus on critical parts of the text that are likely to indicate AI generation.



Sino	Test Title	Description	Input Data	Test Cases	Expected Result	Actual Result	Status
1	Data Collection and Preprocessing	Collect and preprocess data for training and testing	Dataset of real and machine-generated tweets	Verify data loading, check for missing values, preprocess text (tokenization, normalization)	Data loaded and preprocessed successfully	Data loaded and preprocessed successfully	Pass
2	Embedding Generation	Generate FastText embeddings for tweets	Preprocessed tweets	Generate FastText embeddings, check embedding dimensions	Embeddings generated with correct dimensions	Embeddings generated with correct dimensions	Pass
3	Model Training	Train deep learning model on embeddings	FastText embeddings of tweets, labels indicating real or machine-generated	Train model, monitor loss and accuracy, save trained model	Model trained with decreasing loss and increasing accuracy	Model trained with decreasing loss and increasing accuracy	Pass
4	Model Validation	Validate model on validation dataset	FastText embeddings of validation tweets, labels	Validate model, calculate accuracy, precision, recall, F1-score	High accuracy, precision, recall, and F1-score	High accuracy, precision, recall, and F1-score	Pass
5	Model Testing	Test model on unseen test dataset	FastText embeddings of test tweets, labels	Test model, calculate accuracy, precision, recall, F1-score	High accuracy, precision, recall, and F1-score	High accuracy, precision, recall, and F1-score	Pass

CONCLUSION

The integration of convolution neural networks and Fast Text embeddings offers a promising approach for detecting deep fakes and AI bots on social networks. Our proposed system leverages the strengths of image and text analytics to provide a comprehensive solution for identifying misleading content. The experimental results confirm the effectiveness of this method, demonstrating its specific application potential for preserving the integrity of digital information.

Future work will focus on improving the scalability and efficiency of the system, exploring additional features such as audio analysis, and expanding the dataset to include more types of deepfakes and AI-generated content. Continued development of advanced detection techniques is essential to maintaining trust in the digital age and ensuring the authenticity of information on social media platforms.

REFERENCES

[1] Jiameng Pu, Zain Sarwar, Sifat Muhammad Abdullah, Abdullah Rehman, Yoonjin Kim, Parantapa Bhattacharya, Mobin Javed, and Bimal Viswanath. Deepfake text detection: Limitations and opportunities. arXiv preprint arXiv:2210.09421, 2022

. [2] Faris Kateb and Jugal Kalita. Classifying short text in social media: Twitter as case study. International Journal of Computer Applications, 111(9):1– 12, 2015.

[3] Andres Garcia Silva, Cristian Berrio, and José Manuel Gómez-Pérez. An empirical study on pre-trained embeddings and language models for bot detection. In Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019), pages 148–155, 2019. [4] Jonas Lundberg, Jonas Nordqvist, and Antonio Matosevic. On-the-fly detection of autogenerated tweets. arXiv preprint arXiv:1802.01197, 2018..

[5] Robert Gorwa and Douglas Guilbeault. Unpacking the social media bot: A typology to guide research and policy. Policy & Internet, 12(2):225–248, 2020.

[6] Supasorn Suwajanakorn, Steven M Seitz, and Ira KemelmacherShlizerman. Synthesizing obama: learning lip sync from audio. ACM Transactions on Graphics (ToG), 36(4):1–13, 2017.

[7] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2387–2395, 2016. [33] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In Proceedings of the IEEE/CVF international conference on computer vision, pages 5933–5942, 2019.

[8] Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu, et al. Transfer learning from speaker verification to multispeaker textto-speech synthesis. Advances in neural information processing systems, 31, 2018.

[35] Yefei Wang, Kaili Wang, Yi Wang, Di Guo, Huaping Liu, and Fuchun Sun. Audio-visual grounding referring expression for robotic manipulation. In 2022 International Conference on Robotics and Automation (ICRA), pages 9258–9264. IEEE, 2022.

[36] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. OpenAI blog, 1(8):9, 2019.

[37] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. Information Fusion, 64:131–148, 2020.

[38] Yang Zhou, Jimei Yang, Dingzeyu Li, Jun Saito, Deepali Aneja, and Evangelos Kalogerakis. Audio-driven neural gesture reenactment with video motion graphs. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3418–3428, 2022. [39] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[40] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[41] Chunyuan Li, Xiang Gao, Yuan Li, Baolin Peng, Xiujun Li, Yizhe Zhang, and Jianfeng Gao. Optimus: Organizing sentences via pre-trained modeling of a latent space. arXiv preprint arXiv:2004.04092, 2020.

[42] Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. Release strategies and the social impacts of language models. arXiv preprint arXiv:1908.09203, 2019.

[43] Patrick von Platen. How to generate text: using different decoding methods for language generation with transformers. Hugging Face, 2020.

[44] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems, 32, 2019.

[45] Anton Bakhtin, Sam Gross, Myle Ott, Yuntian Deng, Marc'Aurelio Ranzato, and Arthur Szlam. Real or fake? learning to discriminate machine from human generated text. arXiv preprint arXiv:1906.03351, 2019.

[46] Shangbin Feng, Zhaoxuan Tan, Herun Wan, Ningnan Wang, Zilong Chen, Binchi Zhang, Qinghua Zheng, Wenqian Zhang, Zhenyu Lei, Shujie Yang, et al. Twibot-22: Towards graph-based twitter bot detection. arXiv preprint arXiv:2206.04564, 2022.

[47] Samaneh Hosseini Moghaddam and Maghsoud Abbaspour. Friendship preference: Scalable and robust category of features for social bot detection.