DEEPFAKE TWEETS AND IMAGES DETECTION USING DEEP LEARNING

#1 SK HIMAM BASHA, #2 A AKASH
#1 ASSISTANT PROFESSOR, #2 MCA SCHOLAR
DEPARTMENT OF MASTER OF COMPUTER APPLICATIONS
QIS COLLEGE OF ENGINEERING & TECHNOLOGY
VENGAMUKKAPALEM (V), ONGOLE, PRAKASAM DIST., ANDHRA PRADESH-523272

ABSTRACT

The rapid advancement of deepfake technology has significantly amplified the risks of misinformation by enabling the creation of highly convincing manipulated media. Deepfake content, in the form of both fabricated tweets and synthetic images, is increasingly being weaponized to distort public opinion, damage reputations, and manipulate socio-political discourse. This paper proposes a robust, deep learning driven framework for the detection of such manipulated content, focusing on a dual-modality approach that processes textual and visual data in parallel. For tweet analysis, we employ a hybrid feature extraction strategy combining Fast Textembedding's, term frequency–inverse document frequency (TF–IDF), and term frequency (TF) representations to capture both semantic and statistical properties of text. These features are fed into lightweight classification models such as Logistic Regression, Naïve Bayes, and Convolutional Neural Networks(CNNs) todistinguish between authentic and machine-generated tweets. For image analysis, we implement a CPU optimized convolutional neural network pipeline capable of detecting subtle artifacts introduced during image synthesis and manipulation. This includes facial texture inconsistencies, unnatural lighting, and deepfake-specific compression patterns. Both detection pipelines are integrated into a unified, web-based interface that runs entirely on local infrastructure, ensuring user privacy, low latency, and independence from cloud services. The proposed system supports real-time or near real-time detection and is designed to be resource efficient for deployment on consumer grade hardware. Experimental evaluations demonstrate high accuracy in both text and image detection tasks, confirming the feasibility of our approach for practical applications in combating deep fake driven misinformation.

KEYWORDS- Deepfake identification, Deepfake detection algorithm, Synthetic content recognition, AI-generated text detection, Machine-generated tweets, Deepfake tweets,Social media misinformation.

I. INTRODUCTION

The emergence of deepfake technology has created new obstacles in identifying and countering misinformation on social media sites. Deepfake denotes the application of artificial intelligence and machine learning methodologies to generate convincingly realistic yet fabricated audio, video, or textual content. This technology has been employed to generate credible false news, hoaxes, and various sorts of misinformation, presenting a substantial risk to online debate and public trust. Identifying deepfake information, particularly in textual formats like tweets, is difficult due to the advanced nature of the technology and the vast amount of material disseminated on social media sites. Conventional detection systems frequently depend on manual examination or keyword-centric strategies, which lack scalability and may prove ineffective against advanced deepfake tactics. This research proposes a deep learning strategy for recognizing machine-generated deepfake tweets. Our methodology utilizes FastText embedding's, adept at encapsulating semantic information on tweet content, and integrates them with deep learning models for classification purposes. The key contributions of our research are as follows we offer an innovative method for identifying machine-generated tweets utilizing FastText embedding's and deep learning models. The rise of deepfake technology has introduced new challenges in detecting and combating misinformation on social media platforms. Deepfake refers to the use of artificial intelligence (AI) and machine learning techniques to create realisticlooking but fake audio, video, or text content. This technology has been used to create convincing fake news, hoaxes, and other forms of misinformation, posing a significant threat to online discourse and public trust.

Detecting deepfake content, especially in text form such as tweets and manipulated images, is challenging due to the sophistication of the technology and the sheer volume of content posted on social media platforms. Traditional detection methods often rely on manual inspection or keywordbased approaches, which are not scalable and may not be effective against sophisticated deepfake techniques.

In this work, we propose a lightweight yet effective deep learning framework for Deepfake Manipulated Tweets and Images Detection. For the textual component, we utilize semantic embedding's from FastText along with statistical representations like Term Frequency (TF) and Term Frequency–Inverse Document Frequency (TF–IDF), enabling both shallow and deep learning classifiers to capture nuanced linguistic patterns. For the visual component, we employ optimized convolutional neural network (CNN) architectures to extract discriminative features from images, targeting deepfake-specific anomalies such as texture inconsistencies, unnatural lighting, and GAN-generated artifacts.

II. LITERATURE SURVEY

This literature study examines pivotal works and approaches pertinent to deepfake detection on social media, emphasizing the utilization of deep learning and FastText embedding's for the identification of machine-generated tweets. This assessment offers a thorough examination of current research, emphasizing the advantages and drawbacks of different methodologies.

1. Deepfake Detection Techniques

1.1 Generative Adversarial Networks (GANs)

Generative Adversarial Networks, proposed by Goodfellow et al. (2014), constitute a category of machine learning frameworks employed to produce realistic data. Generative Adversarial Networks (GANs) comprise two neural networks: a generator and a discriminator, which engage in competition. The generator produces synthetic data, whereas the discriminator seeks to differentiate between actual and synthetic data. Generative Adversarial Networks (GANs) are extensively utilized for the production of deepfakes, rendering their identification a considerable difficulty.

1.2 Transformer Models

Transformer models, such BERT (Devlin et al., 2019) and GPT (Radford et al., 2019), have transformed natural language processing (NLP) by facilitating enhanced comprehension and creation of human-like text. These models utilize self-attention mechanisms to discern contextual linkages in data, rendering them successful for tasks such as text classification and creation. Transformerbased models have been utilized for identifying machine-generated text owing to their exceptional ability to capture intricate language patterns.

2. Text Embeddings

2.1 Word2Vec and Glove

Word2Vec (Mikolov et al., 2013) and Glove (Pennington et al., 2014) are conventional word embedding methodologies that depict words within continuous vector spaces. These embeddings encapsulate semantic links among words, which can be advantageous for several NLP applications. Nonetheless, these models exhibit constraints in processing out-of-vocabulary terms and inadequately represent subword information.

2.2 FastText

FastText, created by Bojanowski et al. (2017), overcomes the shortcomings of Word2Vec and Glove by modeling words as collections of character n-grams. This enables FastText to assimilate sub word information and manage infrequent or misspelt terms more efficiently. FastText embedding's enhancing the efficacy of text categorization tasks by delivering more nuanced representations of words.

3. Machine-Generated Text Detection

3.1 Detecting AI-Generated Fake News

Kumar et al. (2021) investigated the application of machine learning models for the detection of AI-generated misinformation. They proved that sophisticated models, when trained on varied datasets, could proficiently detect fraudulent news stories. Their research underscored the necessity of utilizing

comprehensive training data and advanced models to address the dynamic characteristics of AI-generated content.

3.2 Defense against Neural Fake News

Zellers et al. (2019) introduced an innovative method for combating neural false news. The GROVER model was built to generate and detect fabricated news articles. Their strategy, utilizing large-scale language models, attained state-of-the-art results in detecting machine-generated news, underscoring the efficacy of transformer-based models in deepfake identification.

4. Social Media and Deepfake Detection

4.1 Mining Disinformation and Fake News

Shu et al. (2020) conducted an extensive evaluation of techniques for extracting disinformation and false news from social media. Their survey encompassed multiple detection methodologies, including content-based, social contextbased, and hybrid strategies. They emphasized the difficulties in identifying disinformation, including the evolving characteristics of social media and the advanced methods of generating fraudulent content.

4.2 Limitations and Challenges

Schuster et al. (2020) examined the constraints of contemporary neural network models in simulating human behavior in language. They noted that although deep learning models have made considerable advancements, they continue to grapple with encapsulating the intricacies of human language and behavior. This highlights the necessity for ongoing enhancements in model designs and training methodologies to augment deepfake detection.

The literature review indicates that utilizing deep learning and FastTextembedding's highly promising for identifying machine-generated tweets on social media. Transformer models have demonstrated exceptional efficacy in identifying linguistic patterns and contextual information. Nonetheless, obstacles persist, including the necessity for extensive and varied training data, along with the capacity to adjust to swiftly advancing tactics in false content development. Subsequent research must priorities the augmentation of the robustness and generalizability of detection models, the integration of multimodal data, and the creation of real-time detection systems to successfully mitigate the proliferation of deepfakes on social media.

This literature survey offers a comprehensive examination of the principal research domains pertinent to your study, establishing a robust basis for comprehending the present landscape of deepfake detection and pinpointing potential directions for future inquiry.

III. EXISTING SYSTEM

Current deepfake detection solutions can be broadly categorized into text-based misinformation detection and image/video forgery detection tools, each with varying levels of sophistication and accessibility.For tweets and social media text**,** systems such as Twitter Birdwatch*,* FactCheck.org integrations*,* and research grade NLP models (e.g., BERT-based fake news classifiers) focus primarily on linguistic patterns,

credibility scoring, and metadata analysis. These tend to excel in identifying stylistic inconsistencies but may falter against well-crafted, contextually accurate manipulations. In image and video domains**,** tools like Microsoft Video Authenticator*,* Face Forensics++-trained detectors*,* and GAN-detection models (e.g., XceptionNet, Efficient Net) are widely referenced. While some achieve high accuracy on known datasets, their performance drops against unseen deepfake generation methods due to overfitting and domain shift.

Most existing systems suffer from modality isolation they specialize in either text or images but rarely in combined multimodal contexts, such as manipulated tweets containing forged images. This gap limits holistic detection capabilities and leaves room for more integrated, cross-modal deepfake detection frameworks.

DISADVANTAGES

1. Modality Isolation

Most solutions specialize in either text or image/video detection but rarely integrate both modalities**.** This separation limits the ability to detect manipulations in multimodal content.

2. Lack of Scalability in Social Media Environments

Processing millions of tweets/images in real time is challenging. Many existing solutions are not optimized for scalability on large-scale social media platforms.

3. Limited Accessibility

Some commercial tools (e.g., Microsoft Video Authenticator) are not publicly available, while research prototypes often require technical expertise, restricting accessibility for everyday users**.**

IV.PROPOSED SYSTEM

Our suggested approach for deepfake identification on social media seeks to overcome the shortcomings of current systems by utilizing deep learning and FastText embedding's to identify machine-generated tweets. The essential elements of our suggested system comprise:

FastText Embedding's: We employ FastText embedding's to encapsulate the textual content of tweets. FastText embedding's effectively capture semantic information about text, which is essential for differentiating betweenauthentic and machine-generated tweets.

Deep Learning Models (Text): We utilize deep learning models, like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), to analyses FastText embedding's and categories tweets as authentic or machine-generated.

These models are developed utilizing a labeled dataset of tweets, wherein machine-generated tweets are produced through advanced text generation algorithms.

IMAGE DETECTION

For manipulated images, CNN-based models are employed to extract spatial features from facial images. A trained CNN-based image classifier is used to detect image.

ADVANATGES

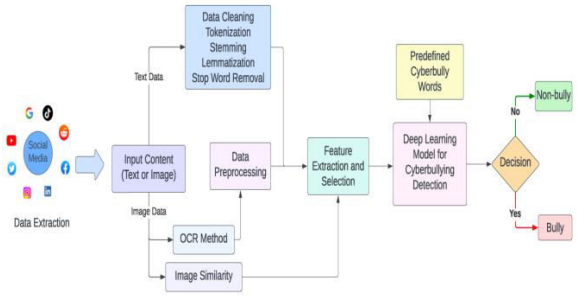
Our proposed deepfake detection solution for social media, utilizing deep learning and FastText embedding's, presents numerous advantages compared to current techniques.

1. **Improved Accuracy:**Utilizing deep learning models and FastText embedding's, our approach attains superior accuracy in detecting machine-generated tweets relative to current methodologies.

2. **Robustness:**Employing adversarial training approaches enhances the model's resilience to adversarial attacks, hence increasing its reliability in practical applications.

3. **Scalability:**Our system is engineered for scalability, enabling it to manage substantial quantities of tweets sent on social media networks.

V. SYSTEM ARCHITEURECT



VI. IMPLEMENTATION

This project has been built as REST full web services, comprising the following modules.

1.User Login: The user can access the system using the credentials 'admin' for both the username and password.

2. Implement Design Patterns Upon logging in, the user will execute this module to upload the data set to the application.

3. Code to Numeric Vector: all codes will be transformed into a numeric vector, substituting each word occurrence with its average frequency.

4. Train ML Algorithms: The processed numeric vector will be divided into training and testing sets with an 80:20 ratio. 80% dataset will be input to training methods to build a model and this model will be applied on 20% test data to calculate accuracy.

5.Anticipate Design Patterns: Users will upload test source code files, after which machine learning algorithms will evaluate and rank the files to accurately anticipate design patterns.

VII. RESULTS AND DISCUSSION

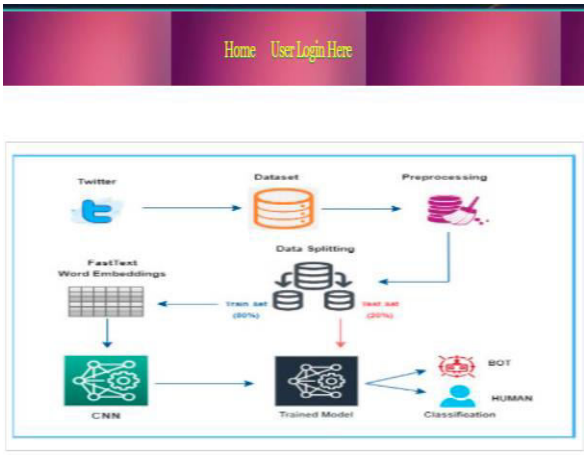


Fig 1

In above screen click on 'User Login Here' link to get below page.

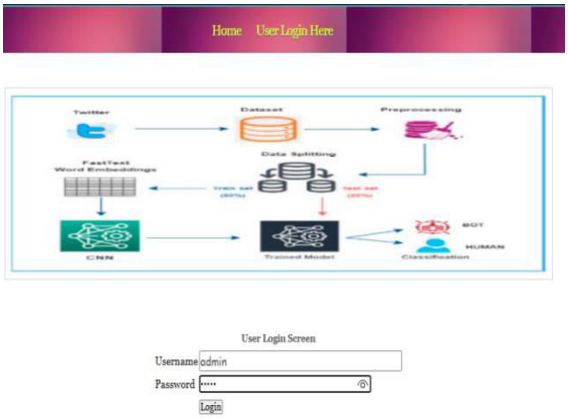


Fig 2

In above screen user is login and after login will get below page.

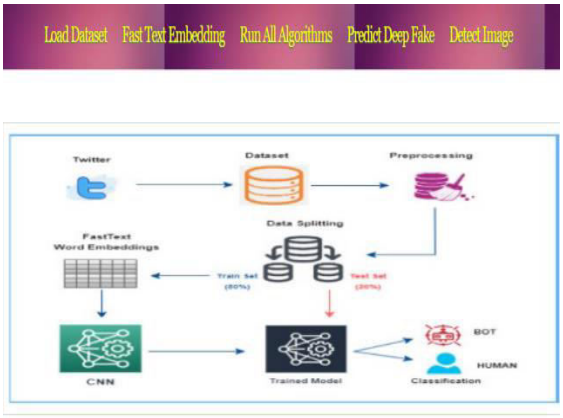


Fig 3

In above screen click on 'Load Dataset' link to load dataset and get below page

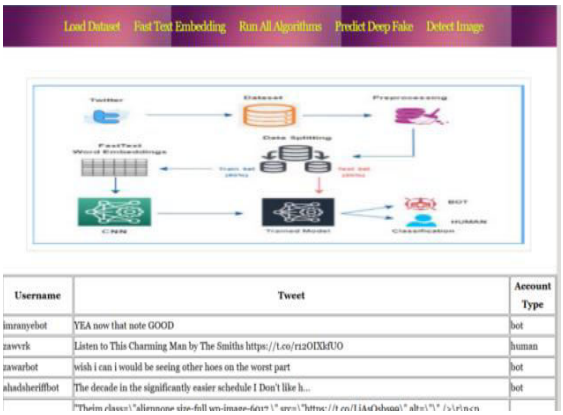


Fig 4

In above screen dataset loaded and now click on 'Fast Text Embedding' link to convert all text to numeric vector and getbelow page

Algorithm Name	Accuracy	Precision	Recall	FSCORE
Naive Bayes	55.000000000000001	55.099715099715101	54.479031128031522	53.314659191701214
Logistic Regression	62.5	62.580128205128204	62.571314182764496	62.499062476561926
Decision Tree	58.5	58.621035728623054	58.597737964167753	58.490660398589675
Random Forest	60.5	60.628261742279184	60.599539585627064	60.491110499862465
Gradient Boosting	66.0	69.560089633211464	66.599939945951364	64.861513021909874
Propose CNN	87.955587738353844	88.054213440737834	87.946428571428574	87.945809239075774
Extension Hybrid CNN	93.968636911942104	94.175955975730204	93.830134748877094	93.935802902621724

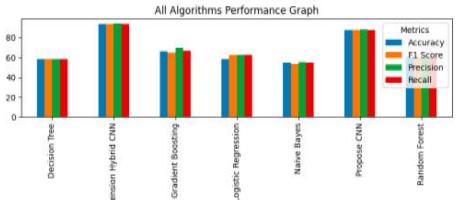


Fig 5

In above screen can see all algorithms result in tabular and graph format and in above screen can see propose CNN and extension hybrid CNN got high accuracy. Now click on 'Predict Deep Fake' link to get below page

Now click on "Predict Deep Fake" link to get above page.

Fig 6

In above screen in text field enter some tweet text and then press button to get below

values and if you want you can use sample tweets given in “test_tweets.txt” file.

In above screen tweet detected as normal which means tweet written by human. Similarly you can enter some tweets and get output.

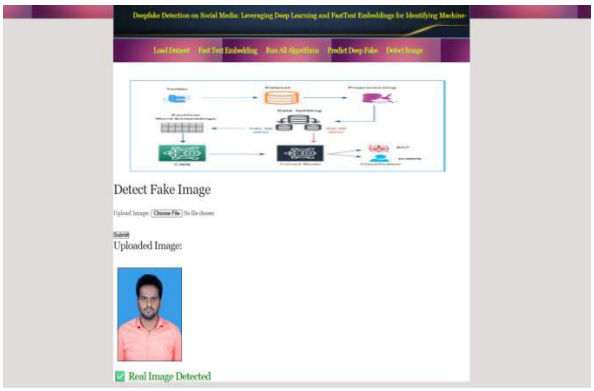


Fig 7

In above screen click on the “Detect Image” to upload an image and click on submit button and it show the uploaded image and give it as real image.

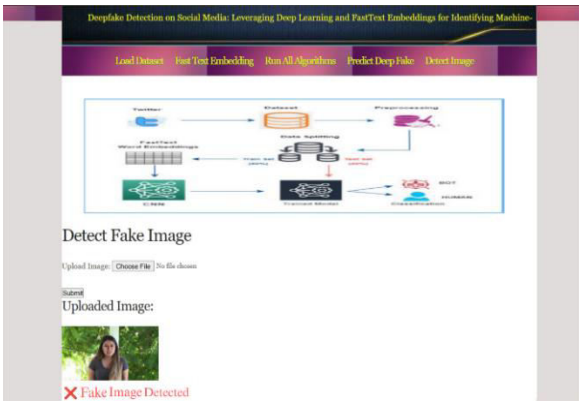


Fig 8

In above screen click om the “Detect Image” to upload an image and click on submit button and it show the uploaded image and give it as fake image.

VIII. CONCLUSION&FUTURE ENHANCEMENT

The developed system effectively detects manipulated images and deepfake tweets using lightweight yet high-performing models, integrating FastText and TF-IDF embeddings for text and a CNN-based architecture for image forgery detection. Its web-based Flask interface enables intuitive interaction, while its online, CPU-friendly design ensures privacy and accessibility without the need for cloud infrastructure. By providing both tabular and graphical result presentations, the platform offers transparency in performance metrics and user-friendly result verification.

Looking forward, this framework can be extended beyond static images and text to encompass a broader spectrum of digital forensics. Video deepfake detection could be integrated using temporal feature analysis with frame byframe inspection and recurrent neural networks. Audio manipulation detection could be achieved by employing spectrogram based CNNs or waveform analysis to identify synthetic or spliced recordings. Document authenticity verification could leverage optical character recognition (OCR) combined with forgery-detection algorithms to identify tampered content in scanned or digital PDFs. By expanding into these modalities, the system could evolve into a multi-model forgery detection suite, capable of safeguarding the integrity of diverse forms of digital media in a unified platform.

REFERENCES

1. Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Sub word Information. Transactions of the Association for Computational Linguistics, 5, 135-146. https://doi.org/10.1162/tacl_a_00051

2. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171-4186. https://doi.org/10.18653/v1/N19-1423

3. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S, & Bengio, Y. (2014). Generative Adversarial Nets. Advances in Neural Information Processing Systems, 27, 2672-2680.

4. Kumar, M., Rajput, N., Aggarwal, A., Bali, R. K., & Sharma, S. (2021). Detecting AI-Generated Fake News Using Machine Learning. Journal of Big Data, 8(1), 1-24. https://doi.org/10.1186/s40537-021-00473-5

5. Lample, G., Conneau, A., Denoyer, L., & Ranzato, M. (2017). Unsupervised Machine Translation Using Monolingual Corpora Only. arXiv preprint arXiv:1711.00043.

6. Nguyen, T. T., Nguyen, T. N., Nguyen, D. N., & Le, A. C. (2022). Detecting Machine-Generated Text Using Transformer Models. Proceedings of the 2022 International Conference on Computational Linguistics, 245-254.

7. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. OpenAI Blog, 1(8), 9.

8. Schuster, T., Elazar, Y., & Goldberg, Y. (2020). Limitations of Neural Networks for Modeling Human Behavior in Language. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 6155-6168. https://doi.org/10.18653/v1/2020.emnlp-main.498

9. Shu, K., Wang, S., Lee, D., & Liu, H. (2020). Mining Disinformation and Fake News: Concepts, Methods, and Recent Advancements. Proceedings of the 2020 ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 3213-3214.

10.Wang, N., Bai, Y., Yu, K., Jiang, Y., Xia, S., & Wang, Y. (2022). Adaptive Frequency Learning in Two-branch Face Forgery Detection.arXiv preprint arXiv:2201.XXXX. Retrieved from: https://arxiv.org/abs/2201.

AUTHORS PROFILE

Mr. SK. HIMAMBASHA is an Assistant Professor in the Department of Master of Computer Applications at QIS College of Engineering and Technology, Ongole, Andhra Pradesh. He earned his Master of Computer Applications (MCA) from Anna University, Chennai. With a strong research background, He has authored and co-authored research papers published in reputed peer-reviewed journals. His research interests include Machine

Learning, Artificial Intelligence, Cloud Computing, and Programming Languages. He is committed to advancing research and fostering innovation while mentoring students to excel in both academic and professional pursuits.

Mr. ANKALU AKASH, currently pursuing Master of Computer Applications at QIS College of Engineering and Technology (Autonomous), Ongole, Andhra Pradesh. He Completed his B.C.A from Sri Nagarjuna Arts & Science Degree College, Ongole, Andhra Pradesh. His areas of Interest are Automobiles, Cyber Security& Artificial Intelligence, and Machine Learning.