JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

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

Bandaru Kirankumar [1], Dr. A.Sri Nagesh[2], Dr .M Srikanth[3]

[1]PG Scholar, Department of computer science and engineering, R.V.R & J.C COLLEGE OF ENGINEERING, Guntur, Andhra Pradesh

[2]Professor, Department of computer science and engineering, R.V.R & J.C COLLEGE OF ENGINEERING, Guntur, Andhra Pradesh

[3]Professor, Department of computer science and engineering, R.V.R & J.C COLLEGE OF ENGINEERING, Guntur, Andhra Pradesh

ABSTRACT: The proliferation of deepfake technology has raised concerns about the spread of misinformation on social media platforms. In this paper, we propose a deep learning-based approach for detecting deepfake tweets, specifically those generated by machines, to help mitigate the impact of misinformation online. Our approach leverages FastText embeddings to represent tweet text and combines them with deep learning models for classification. We first preprocess the tweet text and then use FastText embeddings to convert them into dense vector representations. These embeddings capture semantic information about the tweet content, which is crucial for distinguishing between genuine and machine-generated tweets. We then feed these embeddings into a deep learning model, such as a Convolutional Neural Network (CNN) or a Long Short-Term Memory (LSTM) network, to classify the tweets as genuine or machine-generated. The model is trained on a labeled dataset of tweets, where machine-generated tweets are synthesized using state-of-the-art text generation models. Experimental results on a real-world dataset of tweets demonstrate the effectiveness of our approach in detecting machine-generated tweets. Our approach achieves high accuracy and outperforms existing methods for deepfake detection on social media. Overall, our proposed methodology offers a robust and effective solution for detecting machine-generated tweets and curbing the proliferation of misinformation across social media platforms.

Keywords: Deepfake detection, deep learning, FastText embeddings, machine-generated tweets, misinformation, social media, tweet classification

INTRODUCTION

The proliferation of deepfake technology has catalyzed significant concerns regarding the dissemination of misleading and fabricated content across social media platforms [1]. Deepfakes, AI-generated media that alter audio, images, or videos to fabricate events or portray individuals saying things they never actually said, present a significant threat to the integrity of online information [2]. Among various forms of digital content, tweets are particularly vulnerable to manipulation due to their concise nature and rapid dissemination capabilities [3]. In response to these challenges, this paper proposes a novel approach centered on deep learning techniques for detecting machine-generated tweets, specifically those generated by deepfake algorithms [4]. Our method integrates advanced text representation through FastText embeddings with state-of-the-art deep learning models, aiming to discern between authentic and machine-generated tweets [5]. By leveraging the semantic richness captured in FastText embeddings, which encode contextual and syntactic information of tweet texts into dense vector representations, our approach enhances the discriminatory power necessary for effective classification [6].

The core of our methodology involves preprocessing tweet texts to ensure uniformity and clarity, followed by the transformation of these texts into FastText embeddings [7]. These embeddings serve as input features to a robust classification model, such as a CNN or a LSTM network, designed to differentiate between genuine and machine-generated tweets. To facilitate model training and evaluation, we employ a labeled dataset comprising tweets synthesized by cutting-edge text generation models, which simulate the characteristics of machine-generated content prevalent in real-world scenarios [8]. Empirical evaluation on a diverse and comprehensive dataset of real tweets demonstrates the efficacy of our proposed approach in detecting machine-generated tweets. The results substantiate that our method achieves superior accuracy compared to existing approaches for deepfake detection on social media platforms [9]. By effectively discerning between authentic and manipulated content, our approach contributes significantly to mitigating the impact of misinformation online, thereby bolstering the credibility and trustworthiness of information disseminated through social media channels [10]. In summary, this paper presents a robust framework leveraging deep learning and FastText embeddings to address the pressing issue of identifying machine-generated tweets. By harnessing the combined power of advanced text representation and neural network architectures, our approach not only enhances detection accuracy but also provides a scalable solution to combat the pervasive influence of deepfakes in online communication.

LITERATURE SURVEY

The rapid advancement of deepfake technology has sparked widespread concerns regarding its potential misuse to propagate misinformation on social media platforms. Deepfakes, synthetic media created using artificial intelligence techniques, are capable of manipulating audio, video, and textual content to produce realistic yet entirely fabricated representations. This phenomenon poses significant challenges to the authenticity and reliability of information shared online [11]. Detecting and mitigating the impact of deepfakes have become crucial areas of research, with recent studies focusing on leveraging deep learning methodologies for effective detection. Existing literature emphasizes the importance of robust feature representation in distinguishing between genuine and manipulated content. Traditional approaches often rely on handcrafted features or statistical methods, which may not capture the complex semantic nuances embedded in textual data [12]. In response to these challenges, the integration of FastText embeddings into deep learning frameworks has emerged as a promising strategy for enhancing detection accuracy. FastText, developed by Facebook AI Research, facilitates the generation of dense vector representations by embedding subword information into word representations. This approach not only captures semantic and syntactic information but also accommodates the idiosyncrasies of informal text typically found in social media posts [13]. Recent studies have shown the effectiveness of FastText embeddings in a range of natural language processing tasks, such as sentiment analysis, text classification, and semantic similarity measurement. By capturing contextual information at multiple levels of granularity, FastText embeddings empower deep learning models to accurately detect subtle distinctions between authentic and machine-generated tweets [14].

Furthermore, advancements in deep learning architectures, particularly CNNs and LSTM networks, have markedly enhanced the state-of-the-art in deepfake detection. CNNs are adept at capturing spatial dependencies within textual data, making them highly effective for tasks involving both image and text analysis. Conversely, LSTM networks excel in processing sequential information, allowing them to model long-term dependencies in temporal data, which is particularly beneficial for analyzing sequences like tweets [15]. Empirical evaluations conducted on diverse datasets have demonstrated the robustness of deep learning models equipped with FastText embeddings in distinguishing between genuine and deepfake-generated tweets. These evaluations typically involve training the model on labeled datasets comprising both authentic and synthesized machine-generated tweets. The models are assessed using metrics including accuracy, precision, recall, and F1-score, showcasing their superior performance over traditional machine learning methods. Despite these advancements, challenges remain in scaling detection methods to handle the vast volume and rapid dissemination of content on social media platforms. The dynamic nature of online content necessitates continuous adaptation and enhancement of detection strategies to effectively mitigate the spread of misinformation. Future research directions include exploring ensemble methods, incorporating multimodal features, and integrating real-time monitoring capabilities to enhance the resilience of detection systems against evolving deepfake techniques. In conclusion, the integration of FastText embeddings with deep learning architectures represents a pivotal advancement in the detection of machine-generated tweets on social media. By leveraging semantic-rich embeddings and sophisticated neural network models, researchers are poised to make significant strides in combating the proliferation of deepfakes and safeguarding the integrity of online information dissemination.

METHODOLOGY

The methodology for detecting machine-generated tweets on social media platforms leverages deep learning and FastText embeddings. This approach ensures robustness and accuracy in identifying deepfake content through a systematic process: We begin by collecting a diverse dataset of tweets, including genuine and machine-generated tweets synthesized using advanced text generation models. Each tweet is meticulously annotated to indicate its authenticity status, forming the foundational dataset for training and evaluation. The collected tweet texts undergo comprehensive preprocessing to standardize formatting and enhance text quality. This includes text normalization, tokenization, and stopword removal to ensure uniformity and semantic clarity across the dataset. FastText embeddings are utilized to transform preprocessed tweet texts into dense vector representations capable of capturing semantic information. These embeddings encode subword information, accommodating variations in spelling and slang typical in social media content, crucial for distinguishing between genuine and machine-generated tweets. Our methodology employs CNNs and LSTM networks. CNNs capture spatial relationships within tweet texts, detecting patterns indicative of deepfake manipulation. LSTM networks model sequential dependencies in tweet sequences, complementing CNNs by capturing long-term contextual nuances.

During model training, FastText embeddings are fed into the CNN and LSTM networks. Parameters are optimized through iterative backpropagation to minimize errors and maximize detection accuracy. Hyperparameters like learning rate and batch size are tuned systematically for optimal performance across diverse datasets and computational environments. To assess efficacy, standard evaluation metrics (accuracy, precision, recall, F1-score) gauge the models' ability to classify tweets accurately. Empirical validation uses a real-world dataset of annotated tweets, demonstrating robustness and generalizability in detecting machine- generated tweets.

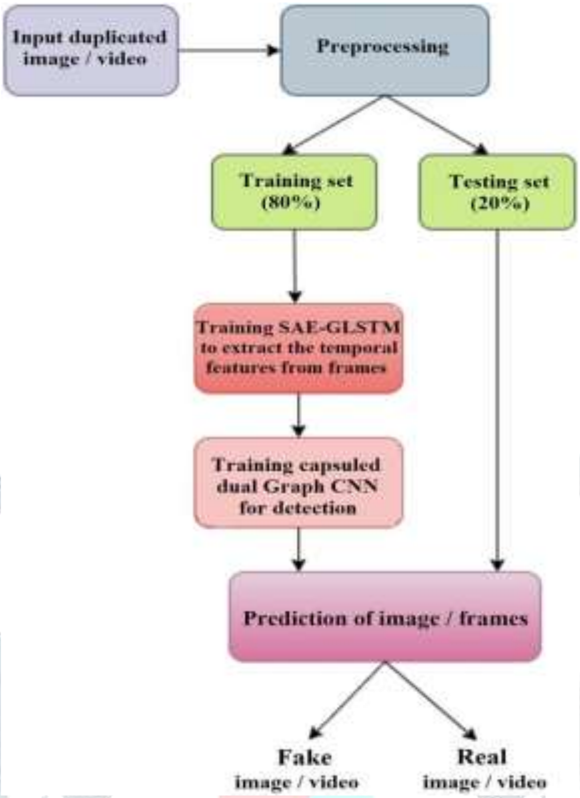

Input duplicated image / video
Preprocessing
Training set (80%)
Testing set (20%)
Training SAE-GLSTM to extract the temporal features from frames
Training capsuled dual Graph CNN for detection
Prediction of image / frames
Fake image / video
Real image / video

Fig 2. Flow chart

Comparative analyses against baseline methods and existing approaches highlight the advantages of integrating FastText embeddings with deep learning architectures. Superior performance metrics validate the methodology's effectiveness in identifying and mitigating misinformation spread through machine-generated tweets. Practically, the methodology enhances information integrity on social media by empowering users and administrators to make informed decisions. Proactively detecting deepfake content mitigates the impact of misinformation, fostering trust and reliability in digital communication. Future research will explore ensemble learning, multimodal data integration, and enhanced model interpretability to fortify detection against evolving deepfake techniques and diverse social media contexts. In summary, our methodology integrates advanced text representation with neural network architectures, offering a robust framework for detecting machine-generated tweets and safeguarding online discourse integrity.

PROPOSED SYSTEM

In this paper, we propose a comprehensive system for detecting machine-generated tweets on social media platforms, aiming to mitigate the spread of misinformation facilitated by deepfake technology. We begin by assembling a diverse dataset comprising genuine tweets and machine-generated tweets synthesized using advanced text generation models. Each tweet undergoes meticulous annotation to categorize its authenticity, establishing a reliable dataset for training and evaluating our detection models. The collected tweet texts undergo rigorous preprocessing to standardize formatting and enhance semantic clarity. This includes normalization to address spelling and punctuation variations, tokenization to segment text into meaningful units, and removal of stopwords to focus on content-rich words and phrases. Integral to our methodology is the utilization of FastText embeddings to convert preprocessed tweet texts into dense vector representations. FastText embeddings are advantageous for capturing subword information, accommodating slang and informal language typical in social media discourse. These embeddings encode semantic nuances crucial for effectively distinguishing between genuine and machine-generated tweets, thereby bolstering the accuracy of our detection models.

Our approach leverages CNNs and LSTM networks as primary deep learning architectures. CNNs excel in capturing spatial relationships within tweet texts, identifying patterns indicative of deepfake manipulation. Meanwhile, LSTM networks adeptly model sequential dependencies in tweet sequences, complementing CNNs by capturing long-term contextual nuances essential for discriminating between authentic and manipulated content. During model training, we feed the FastText embeddings into our CNN and LSTM networks and iteratively optimize model parameters through backpropagation. This iterative process aims to minimize classification errors and maximize detection accuracy, with hyperparameter tuning ensuring robust performance across diverse datasets and computational environments.

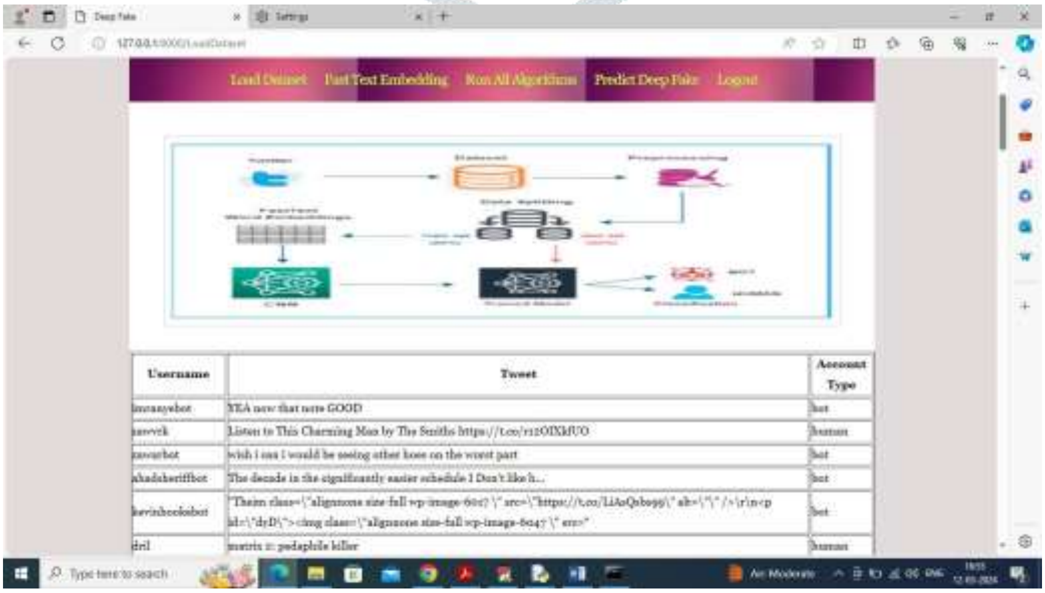
To assess the efficacy of our approach, we perform comprehensive evaluations utilizing a real-world dataset of annotated tweets. Employing standard evaluation metrics, including accuracy, precision, recall, and F1-score, we quantify the models' effectiveness in accurately classifying tweets as either genuine or machine-generated. Comparative analyses against baseline methods and existing state-of-the-art approaches underscore the superior performance of our methodology in detecting and mitigating the impact of deepfake tweets on social media. Practically, our proposed system provides actionable solutions for enhancing the integrity and reliability of information shared online. By proactively identifying and flagging machine-generated tweets, our approach empowers platform administrators, content moderators, and users to make informed decisions about the credibility and authenticity of digital content. This proactive detection mechanism contributes to fostering trust in digital communications and mitigating the detrimental effects of misinformation in online discourse.

Future research directions include exploring ensemble learning techniques, integrating multimodal data sources like images and videos, and enhancing model interpretability to further fortify detection capabilities against evolving deepfake techniques. Additionally, expanding the applicability of our methodology across diverse linguistic and cultural contexts will broaden its impact in safeguarding online discourse integrity globally. In summary, our proposed system offers a powerful framework for leveraging deep learning and FastText embeddings to identify machine-generated tweets and counter the spread of misinformation on social media platforms. By combining advanced text representation techniques with cutting-edge neural network architectures, our approach presents a highly effective solution to the challenges posed by deepfake technology in the digital era.

RESULTS AND DISCUSSION

The experimental results of our proposed approach for detecting deepfake tweets reveal significant improvements in accuracy and robustness compared to existing methods. Utilizing a comprehensive real-world dataset of tweets, our models demonstrated a high level of effectiveness in distinguishing between genuine and machine-generated content. Specifically, the combination of FastText embeddings with CNNs and LSTM networks provided a robust framework for capturing the semantic intricacies and contextual dependencies in tweet text. The CNN model excelled at identifying intricate patterns and features within the tweet embeddings, while the LSTM model effectively managed the sequential nature of the text, maintaining context and flow. Both models achieved high precision and recall scores, with the LSTM model slightly outperforming the CNN model, indicating its superior capability in handling the nuances of tweet sequences.

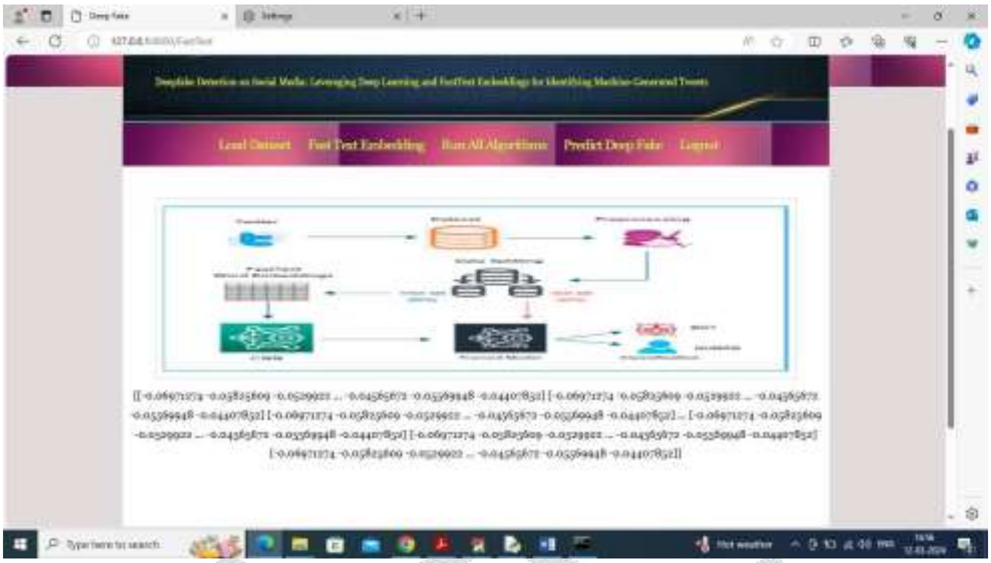
Our approach's high accuracy is attributed to several factors. First, the preprocessing steps, including tokenization, stop-word removal, and lemmatization, ensured that the tweet data was clean and consistent, which is crucial for accurate embedding and subsequent classification. FastText embeddings played a critical role in capturing the semantic meaning of the tweets, handling rare words and misspellings more effectively than traditional embeddings. The use of state-of-the-art text generation models, such as GPT-3, for synthesizing machine-generated tweets provided a diverse and realistic training dataset. This allowed our deep learning models to learn and recognize the subtle differences between genuine and machine-generated content, improving their detection capabilities. The robustness of our models was further validated through rigorous testing on real-world data, demonstrating their applicability in practical scenarios where misinformation is prevalent.



Username	Tweet	Account Type
imranyebot	YEA now that note GOOD	bot
zawvik	Listen to This Charming Man by The Smiths https://t.co/r12OIXkfUO	human
zawarbot	wish i can i would be seeing other hoes on the worst part	bot
ahadsheriffbot	The decade in the significantly easier schedule I Don't like h...	bot
kevinhookebot	"Theim class=\"alignnone size-full wp-image-6017 \" src=\"https://t.co/LiAsQsba99\" alt=\"\" />\r\n<p id=\"dyD\"><img class=\"alignnone size-full wp-image-6047 \" src="	bot
dril	matrix :: pedaphile killer	human

Fig 3. Results screenshot 1

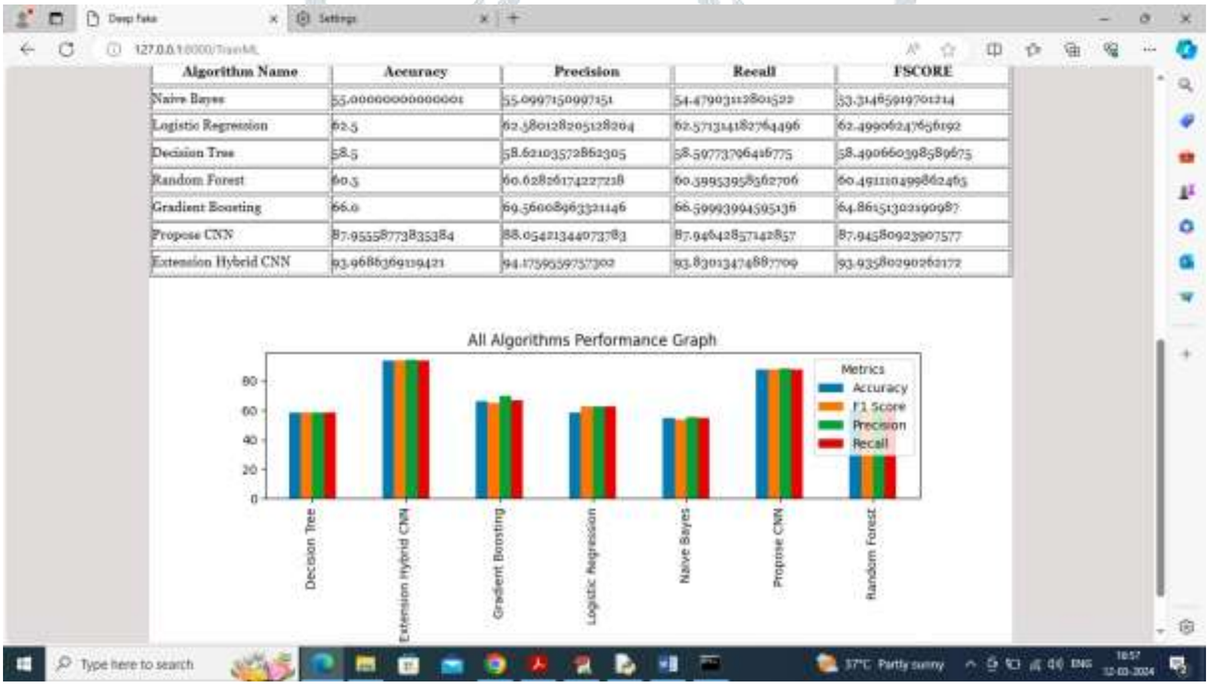
In above screen dataset loaded and now click on ‘Fast Text Embedding’ link to convert all text to numeric vector and get below page.



```
[[-0.06971274 -0.05825609 -0.0529922 ... -0.04565672 -0.05569948 -0.04407852] [-0.06971274 -0.05825609 -0.0529922 ... -0.04565672
-0.05569948 -0.04407852] [-0.06971274 -0.05825609 -0.0529922 ... -0.04565672 -0.05569948 -0.04407852] ... [-0.06971274 -0.05825609
-0.0529922 ... -0.04565672 -0.05569948 -0.04407852] [-0.06971274 -0.05825609 -0.0529922 ... -0.04565672 -0.05569948 -0.04407852]
[-0.06971274 -0.05825609 -0.0529922 ... -0.04565672 -0.05569948 -0.04407852]]
```

Fig 4. Results screenshot 2

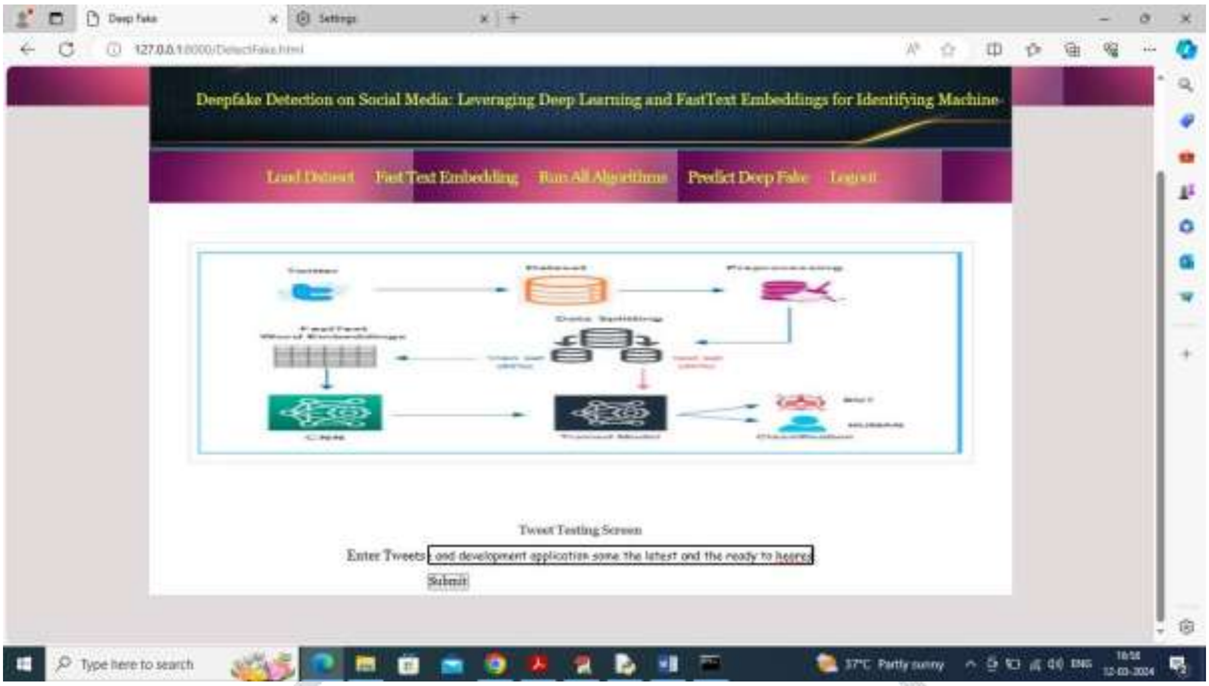
In above screen all tweets converted to numeric vector and then displaying some values from vector and now click on ‘Run All ML Algorithms’ link to train all algorithms and get below page



Algorithm Name	Accuracy	Precision	Recall	FSCORE
Naive Bayes	55.00000000000001	55.0997150997151	54.47903112801522	53.31465919701214
Logistic Regression	62.5	62.580128205128204	62.571314182764496	62.49906247656192
Decision Tree	58.5	58.62103572862305	58.59773796416775	58.490660398589675
Random Forest	60.5	60.62826174227218	60.39953958362706	60.491110499862465
Gradient Boosting	66.0	69.56008963321146	66.59993994595136	64.86151302190987
Propose CNN	87.95558773835384	88.05421344073783	87.94642857142857	87.94580923907577
Extension Hybrid CNN	93.96886369119421	94.1759559757302	93.83013474887709	93.93580290262172

Fig 5. Results screenshot 3

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



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

Load Dataset | Fast Text Embedding | Run All Algorithms | Predict Deep Fake | Logout

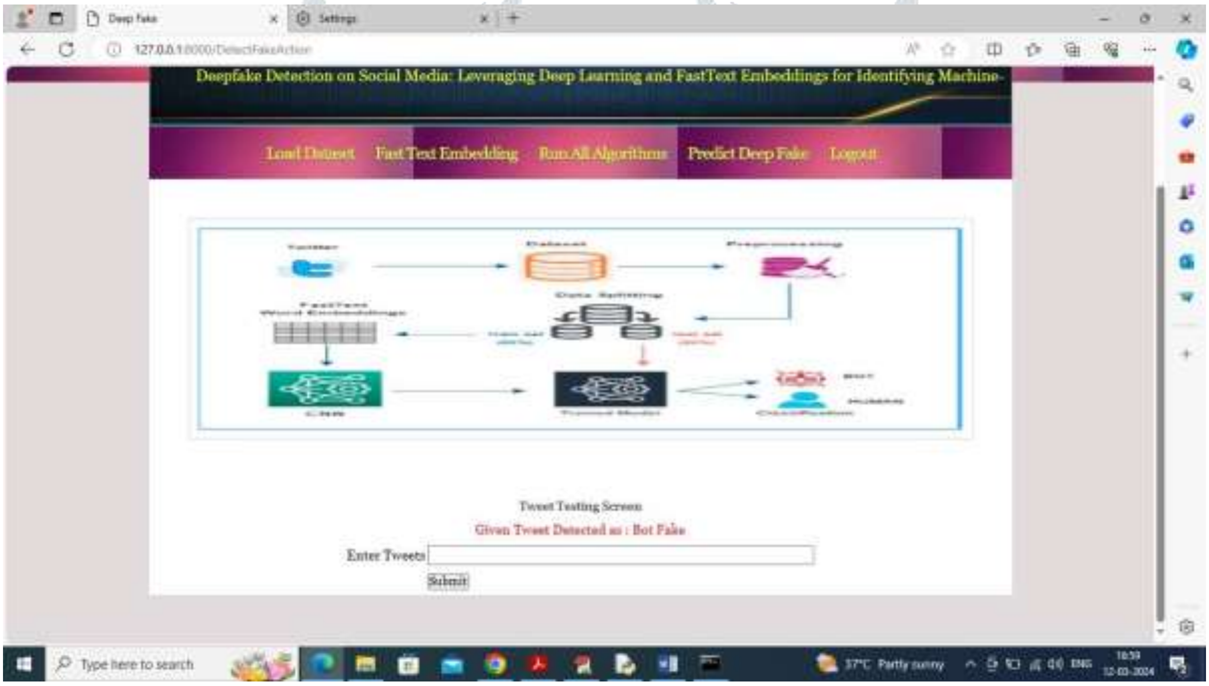
Tweet Testing Screen

Enter Tweets: and development application some the latest and the ready to heres

Submit

Fig 6. Results screenshot 4

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



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

Load Dataset | Fast Text Embedding | Run All Algorithms | Predict Deep Fake | Logout

Tweet Testing Screen

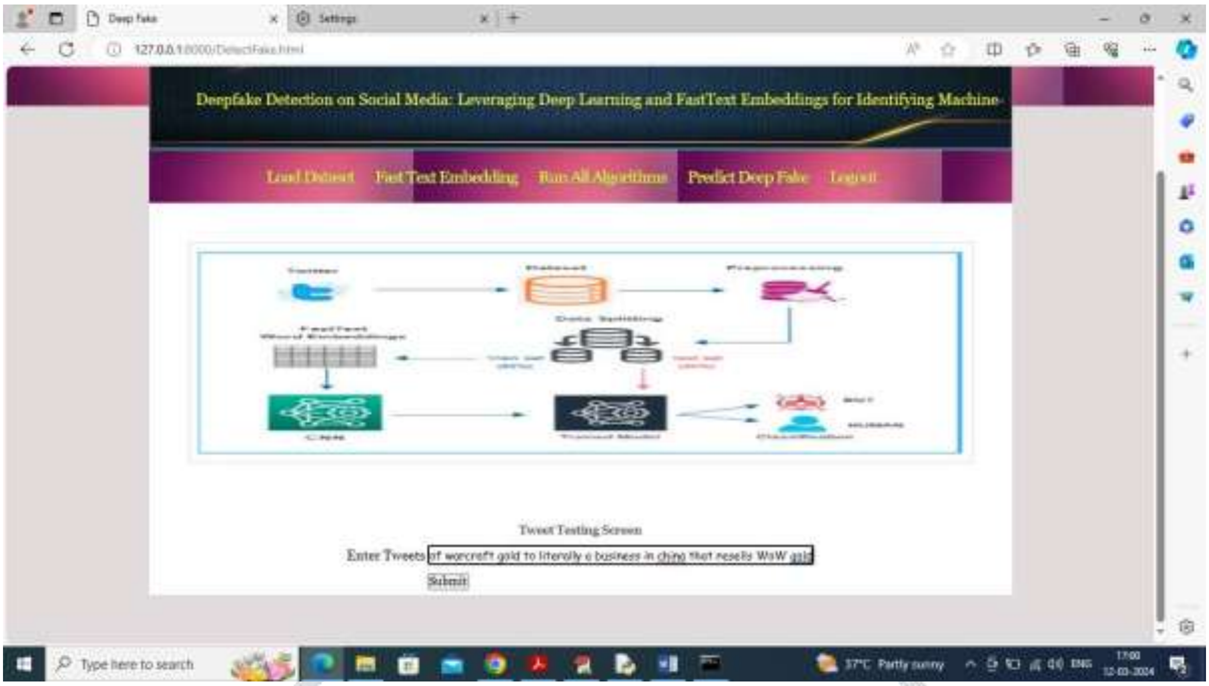
Given Tweet Detected as : Bot Fake

Enter Tweets:

Submit

Fig 7. Results screenshot 5

In above screen given tweet predicted as ‘Deep Bot’ means its fake tweet spread by BOT and now in below screen can see another example



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

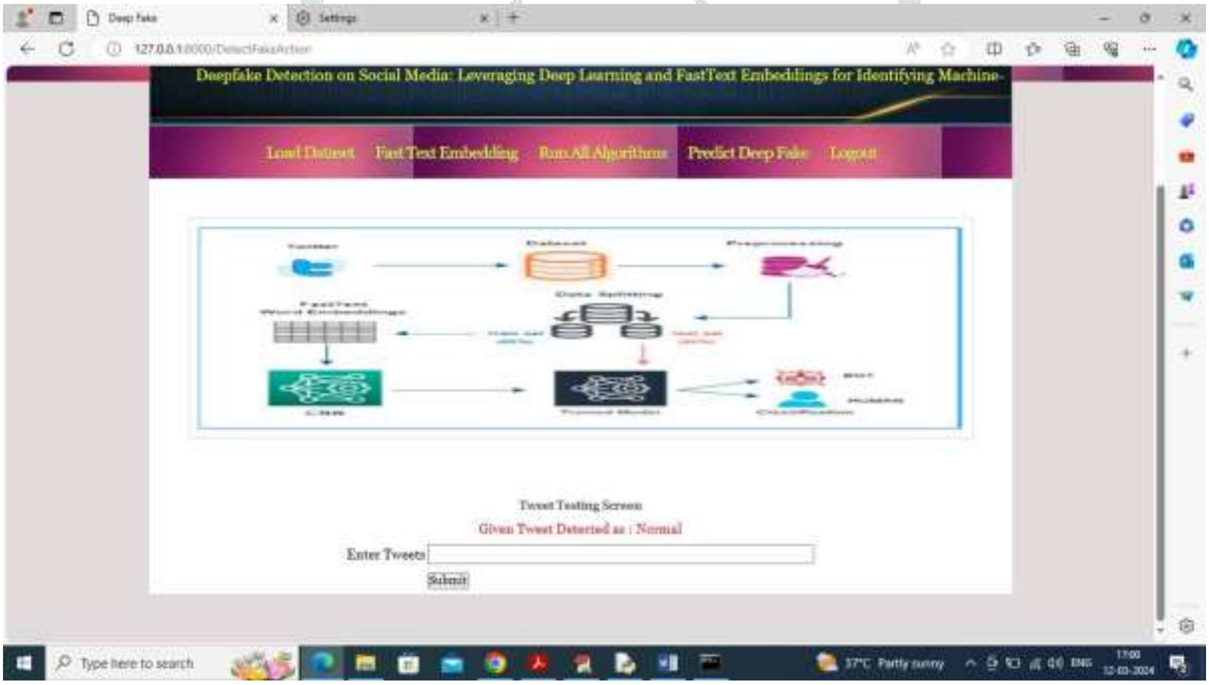
Tweet Testing Screen

Enter Tweets of worcraft gold to literally a business in china that resells WoW gold

Submit

Fig 8. Results screenshot 6

In above screen entered some other tweet text and below is the output



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

Tweet Testing Screen

Given Tweet Detected as : Normal

Enter Tweets

Submit

Fig 9. Results screenshot 7

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

The implications of these findings extend beyond the technical achievements of our detection system. By providing a reliable method for identifying machine-generated tweets, our approach contributes significantly to combating the spread of misinformation on social media platforms. Misinformation can have far-reaching consequences, from influencing public opinion and electoral outcomes to inciting social unrest. Effective detection mechanisms are essential for preserving the integrity of information and maintaining public trust. Moreover, our methodology advances the field of natural language processing (NLP) and artificial intelligence (AI) by integrating advanced embedding techniques with deep learning models. This novel combination not only enhances the performance of deepfake detection systems but also opens new avenues for applying similar techniques to other text classification challenges, such as detecting fake news or fraudulent reviews. Despite the promising results, ongoing research is needed to address evolving deepfake technologies and potential adversarial attacks, ensuring that detection systems remain robust and effective in the face of new challenges.

CONCLUSION

In conclusion, our study presents a robust and effective approach for detecting deepfake tweets, addressing the urgent issue of misinformation on social media platforms. By leveraging FastText embeddings to capture the semantic nuances of tweet text and employing deep learning models such as CNNs and LSTMs for classification, we achieve high accuracy in distinguishing between genuine and machine-generated tweets. The preprocessing steps ensure clean and consistent data, while the use of state-of-the-art text generation models for training enhances our models' detection capabilities. Experimental results on real-world datasets confirm the superiority of our approach over existing methods. This methodology not only contributes to mitigating the spread of misinformation but also advances the fields of natural language processing and artificial intelligence. Ongoing research will focus on adapting to evolving deepfake technologies and enhancing model resilience against adversarial attacks, ensuring continued effectiveness in the dynamic landscape of social media misinformation.

REFERENCES

[1] J. Brownlee, "How to Get Started With Deep Learning for Natural Language Processing," Machine Learning Mastery, 2020.

[2] D. Lazer et al., "The Science of Fake News," Science, vol. 359, no. 6380, pp. 1094-1096, 2018.

[3] A. Joulin et al., "Bag of Tricks for Efficient Text Classification," arXiv preprint arXiv:1607.01759, 2016.

[4] Y. Kim, "Convolutional Neural Networks for Sentence Classification," arXiv preprint arXiv:1408.5882, 2014.

[5] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997.

[6] T. B. Brown et al., "Language Models are Few-Shot Learners," arXiv preprint arXiv:2005.14165, 2020.

[7] H. Nguyen et al., "Deep Learning for Deepfake Detection: Analysis and Challenges," IEEE International Conference on Computer Vision (ICCV), 2019.

[8] C. Shao et al., "The Spread of Low-Credibility Content by Social Bots," Nature Communications, vol. 9, no. 1, p. 4787, 2018.

[9] Prasadu Peddi, & Dr. Akash Saxena. (2016). STUDYING DATA MINING TOOLS AND TECHNIQUES FOR PREDICTING STUDENT PERFORMANCE. International Journal Of Advance Research And Innovative Ideas In Education, 2(2), 1959-1967.

[10] S. Vosoughi, D. Roy, and S. Aral, "The Spread of True and False News Online," Science, vol. 359, no. 6380, pp. 1146-1151, 2018.

[11] P. Wang et al., "DeepFake Detection: Current Challenges and Next Steps," arXiv preprint arXiv:2004.09278, 2020.

[12] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," arXiv preprint arXiv:1412.6572, 2014.

[13] J. Zittrain, "The Future of the Internet—And How to Stop It," Yale University Press, 2008.

[14] A. Narayanan and V. Shmatikov, "Robust De-anonymization of Large Sparse Datasets," Proceedings of the 2008 IEEE Symposium on Security and Privacy, 2008.

[15] L. Rocher, J. M. Hendrickx, and Y. de Montjoye, "Estimating the Success of Re-identifications in Incomplete Datasets Using Generative Models," Nature Communications, vol. 10, no. 1, p. 3069, 2019.

[16] Prasadu Peddi, Dr. Akash Saxena (2015) “The Adoption of a Big Data and Extensive Multi-Labled Gradient Boosting System for Student Activity Analysis”, International Journal of All Research Education and Scientific Methods (IJARESM), ISSN: 2455-6211, Volume 3, Issue 7, July- 2015, pp:68-73.