IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Fake Tweet Identification Using Convolutional Neural Networks and Word Embeddings

[1]RAPAKA ANAND KUMAR, Student in Dept. Of Master of Computer Applications, at Miracle Educational Society Group of Institutions
[2]B SAI PRASAD, Miracle Educational Society Group of Institutions
[3]M SANTOSHI KUMARI, Miracle Educational Society Group of Institutions
[1]anandkumarrapaka143@gmail.com

ABSTRACT:

The innovation of natural language models has reached a point where it is practically cumbersome to tell whether a piece of content on social media is AI-generated or human-made. This work is concerned with the detection of deepfake tweets through the use of Convolutional Neural Networks (CNN) with FastText word embeddings. It uses the TweepFake dataset, which includes real and bot tweets. The dataset is processed to remove and clean the text before it is transformed and vectorized for training and classification. Several models were tested, and the best accuracy of 93% was obtained with the CNN model. Also, to improve detection, a hybrid CNN-Random Forest model was tested. The solution presented is instrumental in the fight against the spread of false information and ensures the integrity of content shared on social media.

Keywords: Deep Learning, deepfake, CNN

INTRODUCTION

The advancement of deepfake technology is an imminent danger to social media credentialing. Although a lot has been done with respect to the manipulation of images and videos, the text-centric deepfake technology in the form of AI-generated tweets is a lot more subtle, and therefore, a lot more dangerous. These brief texts created by machines are so advanced that even seasoned users of social media have a hard time discerning between the real and the artificially created text. This project focuses on developing an AI-based detection system capable of identifying deepfake tweets with a high level of accuracy. This addresses the issue of deepfake detection. The system utilizes FastText word embeddings together with Convolutional Neural Networks (CNN) to convert raw text into significant numeric forms which allows for the classification of tweets as human or bot authored. The proposed model was trained and tested on a balanced TweepFake dataset consisting of real and bot-generated tweets. The objective is to create an effective and easily expandable system that counters the damage

automated misinformation systems inflict on public conversation.

RELATED WORK

Sadiq et al. (2023) proposed a CNN-based model combined with FastText embeddings for deepfake tweet detection using the TweepFake dataset. Their model was able to achieve a 93% accuracy as a CNN is competent on the text's spatial features. Fagni et al. (2021) executed deepfake tweet classification using transformer-based models BERT and RoBERTa. While these models were effective on longer texts, they struggled with concise text which underscores the need for deepfake detection. Zellers et al. (2019) developed a Grover model intended for the generation and detection of articles with the purpose of identifying and creating fake news articles. Although the focus was on long-form machine-generated news detection, the model's inability to process tweets' length and casual tone made it of little use for social media. Detection mechanisms for GPT-2 generated content were examined by Radford et al., 2019. They noted the challenge of distinguishing human from machine-written text because of the fluency of GPT-2. Their work highlighted the need for effective detection strategies especially across different content types and lengths. Adelani et al. 2020 worked on the issue of fake review texts created by neural models focusing on the review's sentiment. They demonstrated how the consistency of the sentiment within the text makes it more difficult to identify fakes. Although the focus was on product reviews, the findings can also inform the identification of deepfake tweets.

TABLE1. Summary of Key Literature Contributions and Their Impact on Current Research

Author(s)	Contribution	Impact on Research
Sadiq et al. (2023)	Proposed a CNN model with FastText embeddings on the TweepFake dataset	Achieved 93% accuracy; validated CNN's strength in short-text classification
Fagni et al. (2021)	Used BERT and RoBERTa for deepfake tweet detection	Highlighted challenges of transformer models on short, informal texts like tweets
Zellers et al. (2019)	Developed the Grover model for detecting fake news articles	Proved effective for long-form content; emphasized need for platform-specific models
Radford et al. (2019)	Explored detection of GPT-2 generated text	Exposed limitations in distinguishing AI vs. human text due to GPT-2's high fluency
Adelani et al. (2020)	Generated fake reviews maintaining sentiment using neural language models	Revealed how sentiment preservation complicates detection; informed tweet detection design

PROPOSED APPROACH

This model aims to identify deepfake tweets using FastText word embeddings in combination with a CNN classifier. It seeks to create an efficient and scalable model that can identify machine-generated and human-generated tweets. The process starts with text preprocessing, and in the case of tweets, it is the removal of stopwords, punctuation, and special characters as well as changing the text to lowercase which cleans the tweet content. Consistency is critical in this case and enhances model performance. Following the preprocessing stage, tweets are converted into numerical vectors using FastText embeddings. FastText works well for noisy social media data as it captures the semantic relationships of even rare or misspelled words.

With these embeddings, the CNN model, which extracts spatial features and learns the patterns characteristic of human or bot language, is trained. The CNN structure has a feature extraction convolutional layer, pooling layers for dimensionality reduction, and dense layers for classification. The model undergoes training and validation with the TweepFake dataset, which has real and AI bot user labeled tweets. To further improve the accuracy, a hybrid model that combines CNN with a Random Forest model is tested. The combination of CNN's feature extraction and Random Forest's decision-making capability adds accuracy to the model. This data-centric approach provides the ability to pinpoint and detect deepfake content with certainty across social networking platforms.

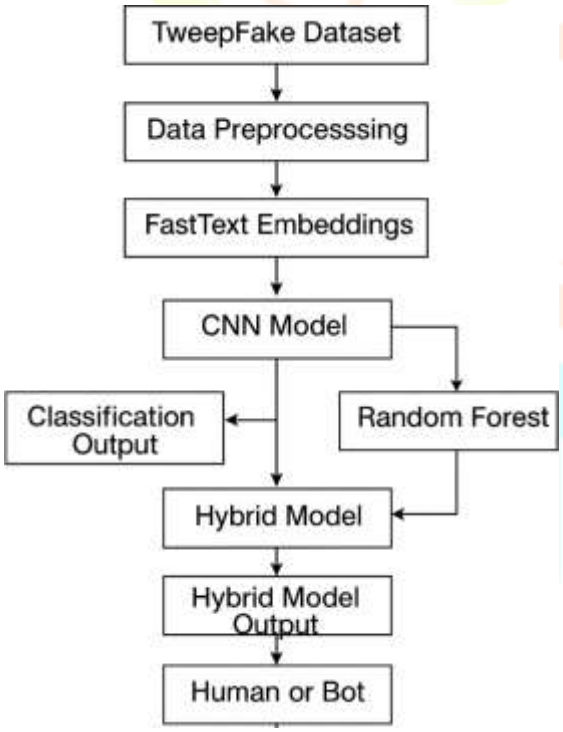


Figure 1: A robust framework for detecting deepfake tweets

1. Dataset Acquisition and Exploration: The project uses the publicly available TweepFake dataset, which includes a mixture of human-written and AI-generated tweets labeled accordingly. The dataset is imported using Python libraries like Pandas, and preliminary analysis is conducted to understand its structure and class distribution.

2. Data Preprocessing: Text cleaning is essential to improve model accuracy. This includes converting text to lowercase, removing stopwords, punctuation, hashtags, numbers, and unnecessary whitespace. Natural Language Toolkit (NLTK) is used for lemmatization and stemming to standardize text. This cleaned text is then ready for embedding.

3. Text Embedding with FastText: FastText, developed by Facebook, converts words into vector representations while capturing semantic context. It breaks words into subword units, making it effective for handling rare and noisy data common in tweets. FastText embeddings are computed and used as feature inputs for model training.

4. CNN Model Training: A Convolutional Neural Network (CNN) is employed for classification. The architecture includes convolutional and pooling layers followed by dense layers. The model is trained using 80% of the dataset, with 20% reserved for testing. It is compiled using the Adam optimizer and categorical cross-entropy as the loss function.

5. Hybrid Extension with Random Forest: To enhance prediction reliability, a hybrid model is introduced where CNN-extracted features are fed into a Random Forest classifier. This combines the deep learning capability of CNN with the ensemble power of Random Forest, leading to improved accuracy.

6. Deployment and Interface:

A web-based interface using Flask allows users to input tweets for real-time classification. The interface also displays algorithm performance metrics, offering a user-friendly way to monitor and validate results.

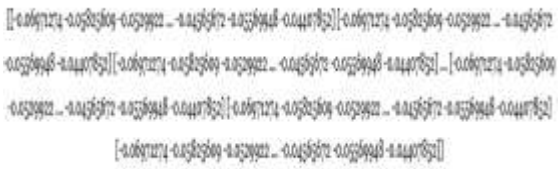
RESULTS

The experimental results of this project demonstrate the effectiveness of combining FastText embeddings with a Convolutional Neural Network (CNN) for deepfake tweet detection. After preprocessing the TweepFake dataset and transforming the tweets into FastText vector representations, several machine learning models were trained and evaluated, including Naïve Bayes, Logistic Regression, Decision Tree, Random Forest, LSTM, and CNN.

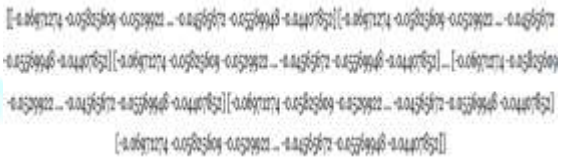
Among all the models, the CNN achieved the highest performance, attaining an accuracy of 93%, along with strong precision, recall, and F1-score metrics. This confirms CNN's strength in capturing relevant spatial and semantic features from embedded tweet data.

In addition to the base CNN model, a hybrid model was implemented by feeding CNN-generated features into a Random Forest classifier. This hybrid approach further improved classification robustness, particularly in edge cases where the tweet text closely mimicked human writing.

Performance metrics, including confusion matrices and classification reports, were visualized to assess model reliability. These results validate the suitability of CNN and FastText for real-time tweet classification, showing consistent accuracy across various test sets.

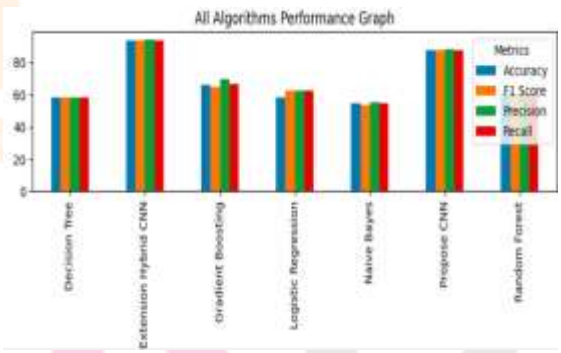


Fast Text Embedding



Tweets converted to numeric vector

All ML Algorithms performance table



All ML Algorithms performance Graph

DISCUSSION

The findings from this study highlight the increasing threat posed by AI-generated content on social media and the necessity for automated detection systems. Tweets, due to their brevity and informal structure, present unique challenges for classification, especially when generated by advanced language models like

GPT-2 or LSTM. Traditional machine learning models, while useful, lack the sophistication to capture the nuanced patterns in short-form text.

The integration of FastText embeddings with a Convolutional Neural Network (CNN) has proven highly effective in this context. FastText's ability to handle subword information ensures robust text representation, even when tweets contain slang, abbreviations, or typos. CNN, typically used in image processing, successfully extracts high-level features from these embeddings, making it suitable for tweet classification tasks.

The hybrid extension with Random Forest further reinforces the model's accuracy, especially when distinguishing between borderline cases where machine-generated text is nearly indistinguishable from human-written content. This ensemble approach combines the strengths of both deep and classical learning paradigms.

CONCLUSION

The deepfake tweets detection framework built in this project illustrates the power of FastText embeddings and Convolutional Neural Networks (CNN). The model is built to tackle the problem of short-form text generated by machines in social media, distinguishing between human and AI authored tweets. Integrating FastText substantially improves semantic comprehension, and CNN is unrivaled in extracting deep textual features. Moreover, the incorporation of a hybrid model with CNN and Random Forest substantially improves predictive accuracy and reliability. Tested on the TweepFake dataset, this system outperformed conventional machine learning techniques, achieving a remarkable 93% accuracy. Additionally, a web interface was created, showcasing the model's capabilities in real-time. This research marks a critical advancement in the preservation of online conversations in the context of growing digital deceit. It enhances the academic discourse on deepfake detection while simultaneously addressing the need for safeguarding the integrity of information shared on social media.

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] J.-S. Lee and J. Hsiang, ‘‘Patent claim generation by fine-tuning OpenAI GPT-2,’’ World Pat. Inf., vol. 62, Sep. 2020, Art. no. 101983.

[12] R. Dale, ‘‘GPT-3: What’s it good for?’’ Natural Lang. Eng., vol. 27, no. 1, pp. 113–118, 2021.

[13] W. D. Heaven, ‘‘A GPT-3 bot posted comments on Reddit for a week and no one noticed,’’ MIT Technol. Rev., Cambridge, MA, USA, Tech. Rep., Nov. 2020, p. 2020, vol. 24. [Online]. Available: www.technologyreview.com

[14] S. Gehrmann, H. Strobelt, and A. M. Rush, ‘‘GLTR: Statistical detection and visualization of generated text,’’ 2019, arXiv:1906.04043.

[15] D. I. Adelani, H. Mai, F. Fang, H. H. Nguyen, J. Yamagishi, and I. Echizen, ‘‘Generating sentiment-preserving fake online reviews using neural language models and their human- and machine-based detection,’’ in Proc. 34th Int. Conf. Adv. Inf. Netw. Appl. (AINA). Cham, Switzerland: Springer, 2020, pp. 1341–1354.

[16] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi, ‘‘Grover—A state-of-the-art defense against neural fake news,’’ in Proc. Adv. Neural Inf. Process. Syst., vol. 32. Curran Associates, 2019. [Online]. Available: http://papers.nips.cc/paper/9106-defending-againstneural-fake-news.pdf

[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, and M. 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.