



Identifying Machine-Generated Tweets with Deep Learning and FastText Embeddings on Social Media

Mr.Mude Sasikanth

Dept. of CSE(AI & ML)

S R Gudlavalleru Engineering College

Gudlavalleru, India

mudesasikanth@gmail.com

Mr.Bathineedi Ajaykumar

Dept. of CSE(AI & ML)

S R Gudlavalleru Engineering College

Gudlavalleru, India

ajaybathineedi@gmail.com

Mr.NalaMolu Sai Mohan

Dept. of CSE(AI&ML)

S R Gudlavalleru Engineering College

Gudlavalleru, India

saimohann134@gmail.com

Mr.Gatta Likith Reddy

Dept. of CSE(AI&ML)

S R Gudlavalleru Engineering College

Gudlavalleru, India

likithreddygatta2003@gmail.com

Mr.Konuganti Gnaneshwar Reddy

Dept. of CSE(AI & ML)

S R Gudlavalleru Engineering College

Gudlavalleru, India

gnaneshwarreddy0816o@gmail.com

Abstract The integrity of information on social media platforms is seriously threatened by deepfake technology, which manipulates media using artificial intelligence. Deepfake content has proliferated in India, particularly in the political and entertainment sectors, where disinformation has been spread by AI-generated movies and fake news. The main goal is to create a strong artificial intelligence model that can reliably identify deepfake content on social media sites, with an emphasis on machine-generated tweets that use FastText embeddings. Conventional techniques included manual social media post filtering using preset criteria and keyword matching, human moderation, and fact-checking organisations. These techniques lacked the scalability to handle the enormous volume of online content and were time-consuming and frequently wrong. Deepfake and AI-generated content detection done by hand is incredibly ineffective, prone to mistakes, and unable to process the enormous amount of social media data in real time. As a result, false and damaging information may proliferate before being discovered or eliminated.

The goal of this study is to counteract false information and protect the integrity of online debate in light of social media's increasing power to shape public opinion. By automating the study of social media information, deep learning models in particular can greatly enhance the detection of deepfakes. While deep learning models can be used to determine if a tweet was created by an AI or a human, FastText embeddings will transform tweets into meaningful

word vectors. This approach enables real-time detection, greater accuracy, and scalability compared to existing approaches

Index terms - Sparrow search algorithm, DL, stock price prediction, LSTM model, sentiment analysis, sentiment dictionary.

1. INTRODUCTION

Significant worries about the spread of false and misleading content on social media platforms have been sparked by the development of deepfake technology. The integrity of internet information is seriously threatened by deep fakes, which are AI-generated media that manipulate audio, photos, or videos to create false events or show people saying things they never actually said. Among the many types of digital content, tweets are especially susceptible to manipulation because of their short length and ease of distribution. To address these issues, this research suggests a novel method for identifying machine-generated tweets—more especially, those produced by deepfake algorithms—that is based on deep learning techniques. In order to distinguish between real and artificially created tweets, our approach combines cutting-edge deep learning models with sophisticated text representation via Fast Text embeddings. Our method improves the discriminatory strength required for efficient classification by utilising the semantic richness obtained by Fast Text embeddings, which incorporate contextual and grammatical information of tweet



sentences into dense vector representations [1]. The foundation of our approach is preparing tweet texts to guarantee consistency and readability, after which they are converted into FastText embeddings. A robust classification model, like a CNN or LSTM network, that distinguishes between real and artificially created tweets uses these embeddings as input features. For training and testing, we use a labelled dataset of tweets generated by state-of-the-art text generation algorithms that mimic the nature of machine-generated content found in real-world contexts [2]. The effectiveness of our suggested method in identifying machine-generated tweets is demonstrated by empirical testing on a wide and extensive dataset of real tweets. The outcomes demonstrate that, when compared to current methods, our strategy for deepfake identification on social media platforms achieves greater accuracy. Our method greatly lessens the impact of false information online by successfully differentiating between real and modified content, which enhances the legitimacy and dependability of information shared on social media. In conclusion, this research tackles the urgent problem of machine-generated tweet identification by presenting a strong architecture that makes use of deep learning and FastText embeddings. Our method improves detection accuracy and offers a scalable way to counteract the widespread impact of deepfakes in online communication by combining the strength of neural network designs and sophisticated text representation..

2. LITERATURE SURVEY

Deepfake technology's widespread use has raised serious worries about the spread of false and misleading content on social media platforms [1]. The integrity of internet information is seriously threatened by deepfakes, AI-generated media that manipulate audio, photos, or videos to create false events or show people saying things they never actually said [2]. Because they are brief and have the potential to spread quickly, tweets are one of the most manipulable types of digital communication [3]. To address these issues, this work suggests a unique method for identifying machine-generated tweets—more especially, those produced by deepfake algorithms—that is based on deep learning techniques [4]. In order to distinguish between real and artificially created tweets, our approach combines cutting-edge deep learning models with sophisticated text representation via FastText embeddings [5]. Our

<https://doi.org/10.36893/iej.2025.v54i05.05>

method improves the discriminatory strength required for efficient classification by utilising the semantic richness recorded in FastText embeddings, which incorporate contextual and grammatical information of tweet messages into dense vector representations [6]. Our methodology is based on preparing tweet texts to make them consistent and readable, and then converting them into FastText embeddings [7]. A robust classification model, like a CNN or LSTM network, that distinguishes between real and artificially created tweets uses these embeddings as input features. For training and testing, we use a labelled dataset of tweets generated by state-of-the-art text generation algorithms that mimic the nature of machine-generated content found in real-world contexts [8]. The effectiveness of our suggested method in identifying machine-generated tweets is demonstrated by empirical testing on a wide and extensive dataset of real tweets. The outcomes confirm that our methodology outperforms current methods for detecting deepfakes on social media platforms in terms of accuracy [9]. Our method greatly reduces the impact of false information on the internet by successfully differentiating between real and altered content, which increases the legitimacy and dependability of information shared on social media platforms [10]. In conclusion, this research tackles the urgent problem of machine-generated tweet identification by presenting a strong architecture that makes use of deep learning and FastText embeddings. Our method not only improves detection accuracy but also offers a scalable way to counteract the widespread impact of deepfakes in online communication by combining the strength of neural network designs and sophisticated text representation. Deepfake technology's quick development has raised many worries about how it can be abused to promote false information on social media. Deepfakes are artificial intelligence-generated synthetic media that can manipulate text, audio, and video to create realistic-looking but completely fake representations. The legitimacy and dependability of information provided online are seriously threatened by this phenomenon [11]. With recent studies concentrating on utilising deep learning approaches for efficient detection, identifying and reducing the impact of deepfakes have emerged as critical study issues. The material now under publication highlights how crucial strong feature representation is for differentiating between authentic and altered content. Conventional methods frequently depend on statistical techniques or manually created characteristics, which fail to capture



the intricate semantic subtleties present in textual data [12]. A viable method for improving detection accuracy in response to these issues is the use of FastText embeddings into deep learning frameworks. By incorporating subword information into word representations, FastText, created by Facebook AI Research, makes it easier to create dense vector representations. In addition to capturing syntactic and semantic information, this method takes into account the peculiarities of informal text, which are frequently encountered in social media posts [13]. In a variety of natural language processing applications, including sentiment analysis, text categorisation, and semantic similarity assessment, recent research has demonstrated the efficacy of FastText embeddings. FastText embeddings enable deep learning models to precisely identify minute differences between real and artificially generated tweets by capturing contextual information at various levels of granularity [14]. Additionally, the state-of-the-art in deepfake detection has significantly improved due to developments in deep learning architectures, including CNNs and LSTM networks. CNNs are very useful for jobs combining both picture and text analysis because they are skilled at identifying spatial connections in textual data. On the other hand, LSTM networks are very good at processing sequential data, which enables them to represent long-term dependencies in temporal data. This is especially useful for analysing sequences, such as tweets [15].

3. METHODOLOGY

i) Proposed Work:

The suggested approach significantly enhances the accuracy of deepfake text identification compared to prior methods. This study's technique presents clear advantages over intricate transfer learning models like RoBERTa and BERT. The employment of a basic CNN model architecture offers numerous advantages. Firstly, it eliminates the necessity for considerable training duration and computational resources usually necessary for fine-tuning transfer learning models. This renders the suggested methodology more accessible and efficient, particularly for academics and practitioners with constrained resources. The proposed text identification method demonstrates that high-level performance is feasible even without labour-intensive transfer learning methods. This research contributes to the field of deepfake

<https://doi.org/10.36893/iej.2025.v54i05.05> identification and offers useful data for studies and applications in the future.

ii) System Architecture:

This article offers a new method based on deep learning technologies for identifying machine-generated tweets, particularly those produced by deep fake algorithms, in reaction to these difficulties. Aiming to distinguish between real and machine-generated tweets, our approach combines cutting-edge deep learning models with sophisticated text representation via Fast Text embeddings. Our method improves the discriminating strength required for efficient classification by using the semantic richness recorded in Fast Text embeddings, which encapsulate contextual and grammatical information of tweet sentences into dense vector representations. Our approach centres on preparing twitter messages to guarantee consistency and clarity; then, we convert these texts into FastText embeddings [7]. Designed to distinguish between real and machine-generated tweets, these embeddings are input characteristics to a strong classification model like a CNN or an LSTM network..

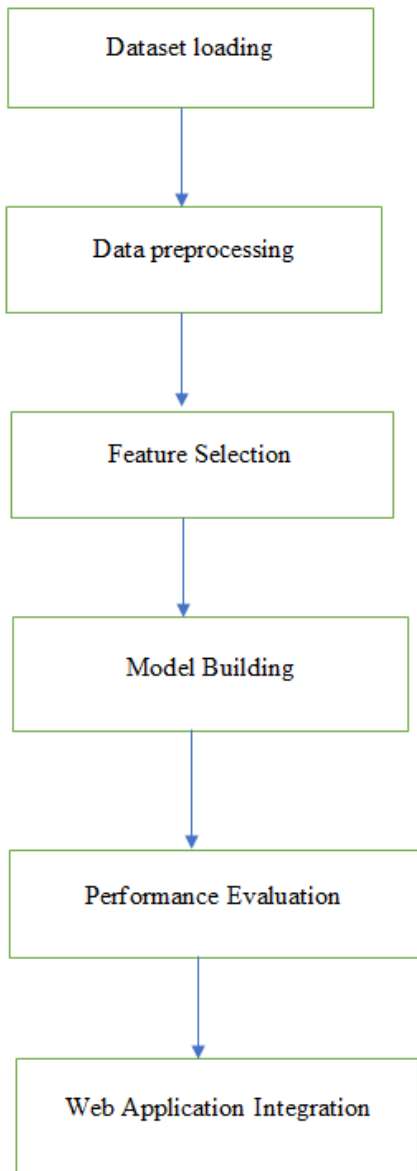


Fig 1 Proposed architecture

iii) Dataset collection:

FAKE TWEETS DATASET

With the help of TweepFake, 25,572 tweets were used in this investigation. There are 17 tweets from real people and 23 from bots in the set. Identified are all machines and humans. Humans (17 accounts, 12,786 tweets), GPT-2 (11 accounts, 3,861 tweets), RNN (7 accounts, 4,181 tweets), or Others (5

<https://doi.org/10.36893/iej.2025.v54i05.05>

accounts, 4,876 tweets) can be the text creation process.

So, these are the top 5 rows of the dataset



Fig 2 tweets dataset

iv) Data Processing:

Unstructured or semi-structured datasets include extraneous information. Training the model takes longer with this extraneous data, which could lead to worse results. It is necessary to pre-process data in order to optimise computational resources and the efficacy of machine learning models. In order for the model to make good predictions, text preparation is essential. Tokenisation, case normalisation, stopword removal, and numeral removal are all part of the pre-processing. Because of case sensitivity, ML models will recognise "MACHINE" and "machine" as separate words. Lowercase data must be preprocessed.

v) Feature selection:

In order to build a trustworthy model, it is necessary to select features that are important, non-redundant, and of high reliability. With the proliferation of both large and diverse datasets, it is crucial to systematically reduce their dimensions. Enhancing a predictive model's efficacy while decreasing computing costs associated with modelling is the primary objective of feature selection. One of the most important parts of feature engineering is feature selection, which involves finding the best features to feed into ML algorithms. In order to train a machine learning model with a smaller set of input variables, feature selection algorithms are used to filter out irrelevant features and duplicates. Feature selection in advance has several advantages over letting the machine learning model determine which features are most important on their own.

4. EXPERIMENTAL RESULTS

Comparison Graph Screen Screen

Algorithm Name	Accuracy	Precision	Recall	F1SCORE
Naive Bayes	0.4	0.2500000000	0.2500000000	0.2500000000
Logistic Regression	0.0	0.0000000000	0.0000000000	0.0000000000
Decision Tree	0.0	0.0000000000	0.0000000000	0.0000000000
Random Forest	0.0	0.0000000000	0.0000000000	0.0000000000
Gradient Boosting	0.0000000000	0.0000000000	0.0000000000	0.0000000000
Support SVM	0.0000000000	0.0000000000	0.0000000000	0.0000000000
Ensemble Hybrid SVM	0.0000000000	0.0000000000	0.0000000000	0.0000000000

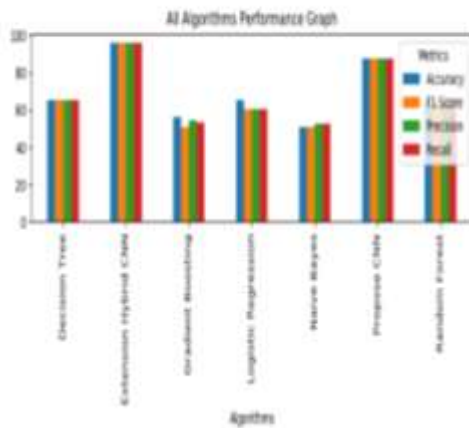


Fig 3: Running Machine Learning and Deep Learning Models

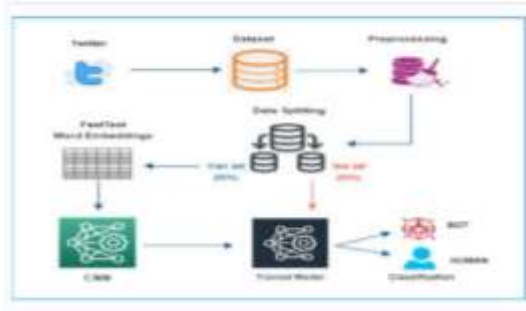


Fig 4: Predict Deep Fake Tweets

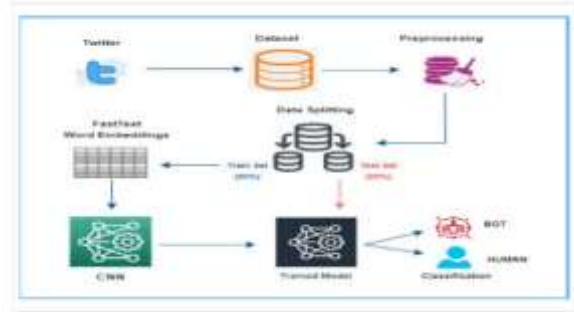


Fig 5: Predict Deep Fake Tweets (Human)

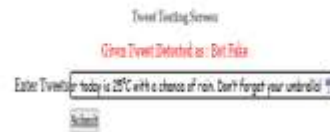
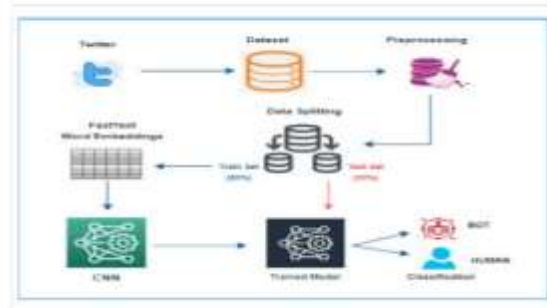


Fig 6: Predict Deep Fake Tweets (Bot)

The Django utility, machine learning, and text processing libraries are imported. While keras is utilised for deep learning models, libraries like pandas, numpy, and sklearn are used for data management and machine learning. Web requests and responses are



handled by Django tools like render, messages, and Http Response. recall, accuracy, precision, and f score: Lists that hold performance indicators for different models. prevent_words, lemmatizer, P.S. These are pre-processing methods for text that include lemmatisation, stemming, and stop word removal. This function pre-processes the text data by removing stop words and punctuation from a tweet and applying lemmatisation and stemming. Following reading, cleaning, and processing, the dataset (Tweepfake.csv) is converted into features (X) and labels (Y). The models and processed data are loaded from disc if they already exist; if not, they are processed from the beginning and stored. Additionally, altered input data is used to define and train a Convolutional Neural Network (CNN). If there are weights in the CNN model, they are loaded for predictions after being stored. For additional performance enhancement, the model additionally incorporates a hybrid strategy that combines CNN features with a Random Forest classifier.

5. CONCLUSION

Information integrity is being threatened by the growing amount of deepfake content on social media, particularly in delicate fields like politics and entertainment. This study uses deep learning and FastText embeddings to tackle the problem of identifying AI-generated content, especially machine-generated tweets. Compared to more conventional techniques like manual filtering and human moderation, which are frequently sluggish, prone to mistakes, and unable to scale with the enormous volume of online content, this method offers a significant advantage in the efficient and accurate detection of deepfakes.

In order to transform tweets into meaningful word vectors that deep learning models can process and categorise as either human-generated or AI-generated, FastText embeddings are essential. Real-

<https://doi.org/10.36893/iej.2025.v54i05.05>

time deepfake detection is made possible by this technique, which guarantees prompt identification and removal of deceptive content before it becomes extensively disseminated. The model outperforms rule-based systems that rely on predefined keywords and human input by combining deep learning with FastText, which increases accuracy and scalability.

To sum up, the suggested deep learning-based framework provides an automated and more dependable way to spot deepfake content on social media sites. In a time when digital content manipulation is becoming more complex, it promises to be essential in thwarting false information and maintaining the integrity of online discourse.

6. FUTURE SCOPE

We need better deepfake text detection technology to safeguard democratic processes and authentic information from the enormous impact of social media on public opinion. While [76] focusses on the challenges and opportunities of quantum machine learning, [77] describes the quantum method for detecting deepfake text. Improved techniques for detecting social media fraud and deception will be developed using quantum natural language processing and other cutting-edge methodologies.

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.