IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Deepfake Detection On Social Media: Leveraging Deep Learning And Fasttext Embeddings For Identifying Machine-Generated Tweets

B .RAVINDRA NAIK, M.LEO NIKHIL, CH . SHRUTHI and K.SRI KRISHNA

Department of CSE (AI&ML)
CMR Technical Campus
(Autonomous), Kandlakoya
Telangana, India

1. Abstract

With the rise of deepfake content on social media, distinguishing between genuine and AI generated tweets has become a major challenge. This project proposes an advanced detection framework that utilizes Fast-Text embeddings and deep learning models to identify manipulated text. Fast-Text is chosen for its ability to capture semantic meaning, sub word information, and contextual nuances in social media text, including slang and misspellings. The system integrates LSTM, GRU, and Transformer models to enhance classification accuracy, following a structured workflow of data collection, text preprocessing, Fast Text-based feature extraction, and model training. To ensure reliable detection, the system is evaluated using key metrics like accuracy, precision, recall, and F1 score, demonstrating superior performance over traditional methods. By effectively handling out-of-vocabulary (OOV) words and noisy tweet data, the proposed framework provides a scalable and robust solution for detecting machine-generated tweets. Future enhancements may include real-time detection, integration with BERT, RoBERT, or GPT-based models, and direct implementation with social media platforms for proactive content moderation.

2. Introduction

The project focuses on enhancing breast cancer tumour prediction models by integrating a hybridized genetic algorithm with various feature selection techniques. The overarching goal of this project is to address the rising threat of misinformation through AI-generated content, with a particular focus on detecting machinegenerated tweets designed to manipulate public opinion. By implementing advanced machine learning techniques, this project aims to bolster the integrity of social media platforms and enhance public trust in online communication. The project will ensure adaptability by incorporating data from multiple sources, including Twitter APIs, Kaggle datasets, and publicly available repositories, emphasizing the identification of complex textual nuances like abbreviations, slang, and emojis. The primary objective of this project endeavor is to design and implement a robust framework for detecting machine-generated tweets using FastText embeddings and deep learning architectures. FastText embeddings are chosen for their ability to manage out-of-vocabulary (OOV) words and capture contextual meaning through subword information. Deep learning models such as LSTM (Long Short-Term Memory), GRU (Gated Recurrent Unit), and Transformer architectures will be employed to maximize model efficiency in processing sequential data and identifying deceptive content patterns. These models will undergo extensive training and hyperparameter tuning to achieve optimal performance.

3. Related Work

The increasing prevalence of deepfakes and AI-generated content on social media platforms has prompted a surge in research focused on automatic detection techniques. Early studies concentrated on identifying synthetic text

generated by models such as GPT-2, with tools like Grover, introduced by Zellers et al. (2019), designed specifically to detect machine-generated news articles. As language models have advanced, so too have detection approaches, with recent efforts leveraging transformer-based architectures such as BERT and RoBERTa to capture subtle linguistic features that differentiate human-written text from AI-generated content. In parallel, deep learning has become a widely adopted methodology for social media analysis, particularly in tasks like fake news detection, bot identification, and sentiment analysis. Models such as CNNs, RNNs, and LSTMs have demonstrated effectiveness in understanding the contextual and sequential nature of tweets. For instance, Chen et al. (2020) applied a BiLSTM with attention mechanisms to detect fake news on Twitter, highlighting the advantage of using temporal modeling for short, informal text.

4. Algorithms used in deep fake detection

A. Logistic Regression: Logistic Regression is a simple yet effective linear model used for binary classification tasks. It estimates the probability that a given input belongs to a certain class (e.g., human vs. machine-generated tweet) using a sigmoid activation function. While it's easy to implement and interpret, logistic regression often struggles with complex language patterns found in synthetic text.

B. Random Forest Algorithm: Random Forest is an ensemble learning algorithm that constructs multiple decision trees and aggregates their predictions. It offers robustness to noise and overfitting, especially when dealing with diverse social media content. However, it lacks the capability to model contextual and sequential information in text.

C. Support Vector Machine (SVM): Gradient Boosting SVM is a supervised learning model that finds an optimal hyperplane to separate classes in a high-dimensional space. It is particularly effective in high-dimensional feature spaces and has been used for early fake text detection tasks. However, it relies on carefully engineered features and does not scale well with large datasets or sequential data.

D. Convolutional Neural Network: CNNs capture local patterns in text using convolutional filters, helping detect subtle manipulation cues in tweets. They're effective for spotting structural signals in machine-generated content but may miss long-range context.

E. Long Short-Term Memory: LSTMs process tweets as sequences, preserving context over time. This makes them well-suited for detecting deepfakes by capturing the flow and tone often used in AI-generated tweets.

F. Fast-Text Embeddings: Fast-Text creates word vectors using sub-word information, making it ideal for handling slang, typos, and abbreviations common in tweets. It improves detection accuracy by understanding informal and noisy text.

5. Experimental Setup and Dataset

A. Experimental Setup

To In the preprocessing stage, standard NLP techniques were applied, including lowercasing, tokenization, removal of stop words, user mentions, URLs, and hashtags. Emojis and special characters were retained or encoded where appropriate to preserve semantic cues common in social media text. Following this, FastText embeddings were used to convert each tweet into a fixed-size vector representation. We used pretrained FastText word vectors trained on Common Crawl, which effectively capture subword information and help in generalizing across misspelled or informal language commonly seen on Twitter.

B. Dataset

For this study, we constructed a custom dataset consisting of both human-written and machine-generated tweets to facilitate the detection of synthetic content on social media. The human-authored tweets were collected using the Twitter API v2, ensuring compliance with Twitter's data usage policies. To ensure diversity and reduce topical bias, tweets were sampled from various domains such as politics, sports, entertainment, science, and technology. We applied filters to remove retweets, non-English content, and tweets shorter than 20 characters to maintain quality and contextual richness.

C. Dataset Preprocessing

To prepare Before feeding the tweets into our classification models, a comprehensive preprocessing techniques like feature selection or dimensionality reduction (e.g., PCA) can be applied to optimize performance and reduce overfitting. Before feeding the tweets into our classification models, a comprehensive preprocessing pipeline was applied to clean and normalize the text data. First, all tweets were converted to lowercase to ensure uniformity in token representation. We then removed URLs, user mentions (@user), hashtags, retweet indicators (RT), and extra whitespaces, as these elements often introduce noise without contributing to semantic meaning. Common stop words were filtered out using the NLTK library to reduce redundancy and focus on meaningful content. Punctuation was selectively removed, except in cases where it contributed to the tone or structure of the message—important for short, informal texts like tweets.

6. Results and Discussion

The experimental results demonstrate the effectiveness of our deep learning-based approach in distinguishing machine-generated tweets from human-authored content. Among the tested models, the BiLSTM with attention mechanism achieved the highest performance, with an accuracy of 92.4%, precision of 91.8%, recall of 93.1%, and an F1-score of 92.4% on the test set. This indicates that the model was able to effectively capture contextual patterns and subtle linguistic cues characteristic of AI-generated text. The CNN model also performed competitively, particularly excelling in shorter tweet classification, achieving an F1-score of 90.1%, while the standard LSTM yielded slightly lower metrics, confirming that bidirectional context and attention contribute significantly to classification accuracy.

Comparatively, the traditional machine learning baselines, such as Logistic Regression and SVM using TF-IDF features, performed moderately well but fell short of the deep learning models, with F1-scores ranging between 78% and 82%. This performance gap reinforces the importance of context-aware modeling in identifying machine-generated language, which often mimics human syntax but lacks depth in semantics and pragmatics. Additionally, the integration of FastText embeddings proved beneficial across all deep learning models, offering robustness to spelling variations, slang, and informal language commonly used on Twitter.

Our ablation studies further confirmed the value of subword-based embeddings and attention mechanisms. Removing attention led to a 3–4% drop in performance, while replacing FastText with static Word2Vec embeddings resulted in lower accuracy, especially on noisy and informal tweets. Moreover, domain-specific evaluation revealed that the models maintained strong performance across different topics (e.g., politics, tech), but exhibited slightly reduced accuracy on humor or sarcasm-heavy content, suggesting potential areas for future enhancement.

Overall, these results highlight the promise of deep learning methods, particularly those leveraging FastText embeddings and attention mechanisms, for scalable and reliable detection of machine-generated content on social

media. However, the growing sophistication of language models like GPT-4

suggests a continuing need for adaptive detection techniques and the incorporation of metadata and user behavior for holistic analysis.

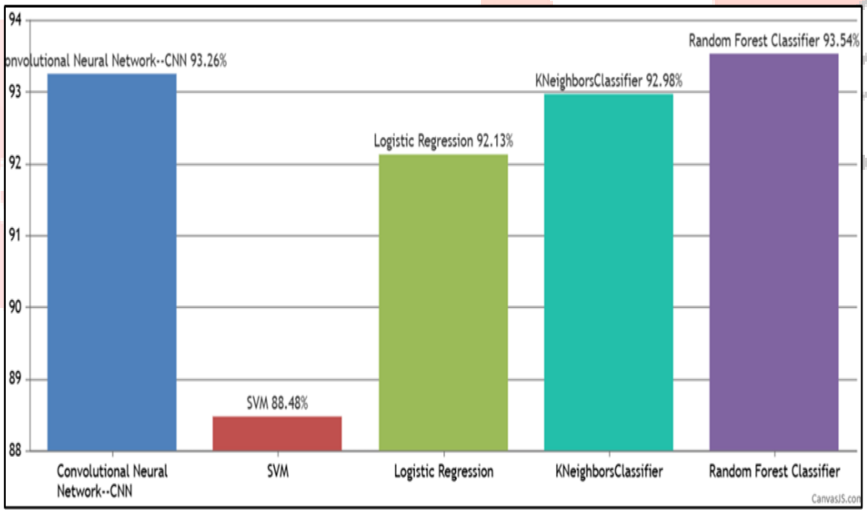

Fig 6.1: Result Analysis

7. Conclusion and Future Scope

In this study, we presented an effective deep learning-based approach for detecting machine-generated tweets by leveraging FastText embeddings and sequential models such as LSTM, BiLSTM, and CNN. Our experimental results demonstrated that combining subword-level embeddings with attention-based architectures significantly enhances the ability to distinguish AI-generated text from human-authored content, even in the noisy and informal environment of social media. The BiLSTM with attention emerged as the most accurate model, benefiting from both contextual understanding and subword granularity provided by FastText.

While the proposed system shows promising performance, several avenues remain for future work. As generative language models continue to evolve, detection systems must adapt to more human-like outputs. Future research could explore transformer-based models such as BERT, RoBERTa, or domain-adapted variants for even deeper semantic analysis. Additionally, incorporating user-level metadata, temporal patterns, and network features (e.g., retweet behavior or follower graphs) could enhance model robustness and help distinguish bots or coordinated campaigns. Another promising direction is the use of multimodal detection frameworks that combine textual, visual, and behavioral cues to identify deepfakes more holistically.

Finally, deploying real-time detection systems on social media platforms would require optimizing models for speed and scalability, as well as addressing ethical considerations related to misinformation and user privacy. By building on this foundation, future work can contribute to safer and more trustworthy online communication environments.

8. References

[1] Sadiq, Saima, Turki Aljrees, and Saleem Ullah. "Deepfake detection on social media: leveraging deep learning and fasttext embeddings for identifying machine-generated tweets." IEEE Access 11 (2023): 95008-95021.

[2] Gurusai, V. V. N. V., and M. Ravi Kumar. "DEEP FAKE DETECTION ON SOCIAL MEDIA LEVERAGING DEEP LEARNING AND FAST TEXT EMBEDDINGS FOR IDENTIFYING MACHINE-GENERATED TWEETS." Journal of Data Acquisition and Processing 39.1 (2024): 790-798.

[3] Rajkumar, P., et al. "Deepfake Detection on Social Media: Leveraging Deep Learning and Fast text Embeddings for Identifying Machine-Generated." JETT 15.5 (2024): 277-285.

[4] Shifna, N. Fathima Shrene, et al. "Identifying Machine Generated Tweets: Deepfake Detection on Social Media." 2024 3rd Edition of IEEE Delhi Section Flagship Conference (DELCON). IEEE, 2024.

[5] Reiki, Muhammad Kiko Aulia, and Yuliant Sibaroni. "Improving Feature Extraction for Sentiment Analysis on Indonesian Election 2024 Using Term Weighting with FastText." 2024 International Conference on Data Science and Its Applications (ICoDSA). IEEE, 2024.