DETECTION OF AI-GENERATED TWEETS USING FAST TEXT EMBEDDINGS WITH DEEP LEARNING TECHNIQUES

[1]M. MADHUSUDHANA RAO, [2]GALLA PUJITHA, [3]SD ASHRAF PASHA, [4]KAKANI JASHUA PAUL, [5]CHENNURI HEMA SRAVANI

[1]Assistant Professor, [2345]Students, Department of Computer Science and Engineering, SRI VASAVI INSTITUTE OF ENGINEERING & TECHNOLOGY, NANDAMURU, ANDHRA PRADESH

ABSTRACT

The rapid growth of social media platforms has significantly increased the amount of textual content generated and shared online. With the advancement of Artificial Intelligence (AI) and Natural Language Processing (NLP), machines are now capable of generating human-like text that closely resembles genuine user content. While these technologies provide many benefits, they also introduce serious challenges such as the spread of misinformation, manipulation of public opinion, and automated bot-driven campaigns. Detecting AI-generated tweets has therefore become an important task in maintaining the authenticity and reliability of online communication platforms. This project proposes an efficient system for detecting AI-generated tweets using FastText embeddings combined with deep learning techniques. The system begins with data preprocessing, where tweet text is cleaned, normalized, and tokenized to remove noise such as special characters, URLs, and stop words. After preprocessing, FastText embeddings are used to convert textual data into meaningful vector representations that capture semantic relationships and subword information. These feature representations are then fed into a deep learning classification model capable of distinguishing between human-written tweets and AI-generated content. The performance of the proposed model is evaluated using standard evaluation metrics including accuracy, precision, recall, and F1-score to ensure reliability and effectiveness. Experimental results indicate that the model can successfully identify AI-generated tweets with high accuracy and efficiency. The proposed system can be integrated into real-time social media monitoring tools to detect automated content, reduce the spread of misinformation, and enhance the transparency of online platforms. Overall, this research contributes to improving digital trust by providing an intelligent and scalable solution for identifying machine-generated social media content.

Keywords:

Artificial Intelligence, Natural Language Processing, AI-Generated Tweets, FastText Embeddings, Deep Learning, Social Media Analysis, Text Classification, Misinformation Detection.

I INTRODUCTION

The rapid expansion of social media platforms has significantly transformed how individuals communicate and share information in the digital world. Platforms such as Twitter allow users to instantly publish opinions, news, and personal experiences, resulting in enormous volumes of textual data generated every day [1]. This large-scale information exchange has created opportunities for research in data analysis,

sentiment detection, and social behavior understanding [2]. However, with the advancement of Artificial Intelligence (AI) and Natural Language Processing (NLP), machines are now capable of generating highly realistic human-like text [3]. Modern language models can automatically produce tweets that appear authentic, making it difficult to distinguish between human-written and AI-generated content [4]. While such technologies have positive applications such as automated assistance, chatbots, and intelligent content generation [5], they also raise concerns related to misinformation and manipulation of public opinion [6]. AI-generated tweets can be used to spread false information, influence political discussions, or artificially amplify certain viewpoints through automated bots [7]. These risks highlight the importance of developing reliable detection mechanisms capable of identifying synthetic text [8]. Researchers have explored various approaches including rule-based systems, statistical language analysis, and machine learning algorithms to address this challenge [9]. Traditional methods often rely on linguistic patterns and handcrafted features extracted from textual data [10]. However, these approaches struggle to detect sophisticated AI-generated content because modern models can replicate natural language patterns very effectively [11]. To overcome these limitations, advanced representation techniques such as word embeddings have been introduced to capture semantic relationships between words [12]. Word embedding models like Word2Vec and GloVe transform textual data into numerical vectors that preserve contextual meaning [13]. Among these approaches, FastText provides additional advantages by considering subword information and morphological patterns [14]. This capability allows FastText to better handle rare words, spelling variations, and informal language commonly found in social media posts [15].

In recent years, deep learning techniques have demonstrated remarkable success in processing large-scale textual data and performing complex classification tasks [16]. Neural network architectures such as Convolutional Neural Networks (CNN) are capable of capturing local linguistic features and patterns within sentences [17]. Recurrent Neural Networks (RNN) are widely used for sequential data analysis because they can process information in a time-dependent manner [18]. Long Short-Term Memory (LSTM) networks further improve this capability by learning long-term dependencies within textual sequences [19]. These models enable systems to analyze syntax, context, and writing style in order to distinguish human-generated text from machine-generated content [20]. When combined with word embedding techniques, deep learning models can effectively learn semantic and structural relationships within tweets [21]. FastText embeddings are particularly useful in this context because they incorporate character-level information and handle unseen words more efficiently [22]. This improves the model's ability to detect subtle linguistic patterns that may indicate automated text generation [23]. Developing accurate detection models is crucial for maintaining trust and transparency in social media environments [24]. Effective AI-generated tweet detection can help prevent the spread of misinformation and reduce the impact of automated propaganda campaigns [25]. It also assists social media platforms in identifying bot activities and suspicious accounts that manipulate online discussions [26]. Furthermore, reliable detection systems support responsible AI development by ensuring that generative technologies are not misused for malicious purposes [27]. Integrating

FastText embeddings with deep learning classifiers therefore provides a promising solution for identifying machine-generated tweets [28]. Such systems can analyze large datasets efficiently while maintaining high accuracy in classification tasks [29]. Consequently, this project focuses on designing an intelligent framework that leverages FastText and deep learning methods to detect AI-generated tweets and enhance the integrity of online communication systems [30].

II LITERATURE SURVEY

The rapid development of Natural Language Processing (NLP) and machine learning techniques has significantly improved the ability to analyze and process large volumes of textual data generated on social media platforms. Early research in text representation focused on statistical and vector-based models that convert words into numerical formats suitable for machine learning algorithms. Word embedding techniques such as Word2Vec introduced efficient methods for capturing semantic relationships between words by learning distributed vector representations from large corpora [1]. Later, GloVe improved word representation by combining global matrix factorization and local context-based learning to enhance semantic understanding [2]. FastText further advanced this concept by incorporating subword information, enabling models to better understand rare words and morphological variations in text [3]. Researchers also demonstrated that subword-based embeddings significantly improve text classification performance, especially for noisy social media data [4]. With the emergence of deep learning, neural network approaches became widely used for text analysis tasks such as sentiment analysis, spam detection, and topic classification [5]. Deep learning models have the advantage of automatically learning feature representations from raw textual data without extensive manual feature engineering [6]. Studies have shown that convolutional neural networks can effectively capture local linguistic patterns within sentences, making them suitable for sentence classification tasks [7]. Similarly, convolutional architectures have been successfully applied to text classification problems by learning hierarchical representations of language features [8]. Recurrent neural networks have also been widely adopted for sequential text processing because they can model dependencies between words in a sentence [9]. Long Short-Term Memory networks were introduced to overcome the vanishing gradient problem and to capture long-term contextual dependencies in textual sequences [10]. These architectures enable machines to understand contextual information in language more effectively [11]. Furthermore, the integration of attention mechanisms and transformer architectures has significantly improved the performance of NLP systems in various tasks [12]. Pretrained language models such as BERT have demonstrated remarkable success in capturing contextual relationships within text by using bidirectional transformers [13]. These advancements have paved the way for more sophisticated approaches to automated text analysis and classification [14][15].

In addition to text classification research, significant attention has been given to the detection of automated or AI-generated content on social media platforms. The increasing presence of social bots has raised concerns regarding misinformation campaigns, fake news dissemination, and manipulation of online discussions [16]. Researchers have studied the behavior and characteristics of social bots to understand how automated systems influence digital communication environments [17]. Machine learning approaches have been widely used to

detect suspicious social media accounts and automated content generation patterns [18]. Several studies have explored the use of linguistic features, metadata, and behavioral characteristics to distinguish between human-generated and machine-generated text [19]. Fake news detection research has also contributed to identifying patterns associated with misleading or artificially generated information on social media [20]. As generative language models became more advanced, detecting machine-generated text became increasingly challenging [21]. To address this issue, researchers developed specialized detection tools that analyze statistical patterns and probability distributions in generated text [22]. Visualization-based approaches have also been proposed to highlight differences between human-written and AI-generated content [23]. Studies have further demonstrated that human evaluators often struggle to differentiate machine-generated text from authentic human writing, emphasizing the need for automated detection systems [24]. Additional research has explored neural approaches for identifying generated text using deep learning classifiers trained on large datasets [25]. These models analyze contextual patterns and stylistic differences between human and machine-generated language [26]. Recent studies have also focused on evaluating the reliability of human judgments in assessing AI-generated content and highlighting the importance of objective detection mechanisms [27]. Moreover, advanced neural architectures and cross-lingual language models have been investigated to improve text understanding across multiple languages [28]. Pretraining strategies and transformer-based architectures continue to enhance the ability of NLP systems to analyze complex linguistic patterns [29]. Overall, existing literature demonstrates that integrating word embeddings with deep learning models provides a powerful approach for detecting AI-generated tweets and improving the reliability of social media analysis systems [30].

III METHODOLOGY

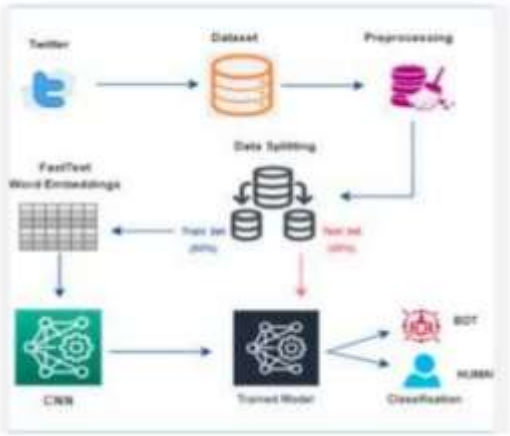
The methodology for detecting AI-generated tweets using FastText embeddings and deep learning techniques involves several systematic stages including data collection, preprocessing, feature extraction, model training, and performance evaluation. Initially, a dataset containing both human-written and AI-generated tweets is collected from reliable sources such as public tweet datasets or synthetic text generation platforms. The collected dataset is then divided into training and testing subsets to ensure proper model evaluation. In the preprocessing stage, raw tweets are cleaned to remove noise such as URLs, hashtags, special characters, punctuation marks, and unnecessary whitespace that may affect the performance of the model. Additionally, text normalization techniques such as lowercasing, tokenization, and stop-word removal are applied to prepare the textual data for analysis. After preprocessing, the cleaned text is transformed into numerical representations using FastText embeddings. FastText is particularly effective because it captures semantic relationships between words while also considering subword information, which helps the model understand spelling variations, abbreviations, and rare words commonly present in social media text. Once the text is converted into vector representations, these embeddings are fed into a deep learning classification model designed to distinguish between human-written and AI-generated tweets. Neural network architectures such as Convolutional Neural Networks (CNN) or Long Short-Term Memory (LSTM) networks can be used to learn contextual and sequential patterns in the textual data. The model is trained using labeled tweet data where each tweet is tagged as either human-

generated or AI-generated. During training, optimization algorithms such as Adam are applied to minimize the loss function and improve classification accuracy. After the training process, the model is tested using unseen data to evaluate its ability to correctly identify AI-generated tweets. Performance metrics such as accuracy, precision, recall, and F1-score are used to measure the effectiveness of the system. The final trained model can then be integrated into real-time monitoring systems to detect automated content and enhance the reliability and transparency of social media platforms.



IV SYSTEM DESIGN

The system design for detecting AI-generated tweets using FastText embeddings and deep learning techniques focuses on creating an efficient and scalable architecture capable of processing large volumes of textual data from social media platforms. The overall system consists of several interconnected modules including data acquisition, preprocessing, feature extraction, model training, and classification. In the first stage, the data acquisition module collects tweets from publicly available datasets or social media platforms. These datasets contain both human-generated tweets and AI-generated tweets, which are labeled for supervised learning. The collected data is then passed to the preprocessing module, where raw text is cleaned and normalized to remove unnecessary noise such as URLs, hashtags, emojis, punctuation marks, and special characters that may affect model performance. Text normalization techniques such as tokenization, lowercasing, and stop-word removal are applied to ensure that the textual data becomes structured and suitable for further processing. This preprocessing stage is essential because social media data is often unstructured and contains informal language, abbreviations, and spelling variations. After preprocessing, the cleaned text is forwarded to the feature extraction module where FastText embedding is used to convert textual data into numerical vector representations. FastText is particularly useful for social media analysis because it considers subword information and character-level patterns, enabling the system to better understand rare words and misspelled tokens commonly found in tweets. These vector representations capture semantic relationships among words, which helps the machine learning model interpret textual meaning more effectively.



Once the feature vectors are generated, they are passed to the classification module that employs deep learning algorithms to detect AI-generated tweets. Neural network architectures such as Convolutional Neural Networks (CNN) or Long Short-Term Memory (LSTM) networks are used because they can learn complex linguistic patterns and contextual relationships within the text. The training process involves feeding the model with

labeled data where each tweet is classified as either human-generated or AI-generated. During training, the model adjusts its internal parameters through optimization algorithms such as gradient descent to minimize classification errors and improve prediction accuracy. After the model is trained, the testing module evaluates its performance using unseen data to ensure its generalization capability. Performance evaluation metrics such as accuracy, precision, recall, and F1-score are calculated to measure the effectiveness of the system in correctly identifying AI-generated tweets. The final system can then be deployed as part of a social media monitoring framework where incoming tweets are automatically processed through the pipeline. The deployed model analyzes each tweet in real time, converts it into FastText embeddings, and classifies it as human-written or AI-generated. This automated detection mechanism helps social media platforms identify suspicious automated content, reduce the spread of misinformation, and maintain the authenticity of online discussions. Overall, the system design provides a structured and efficient framework that integrates data preprocessing, advanced text representation, and deep learning classification to achieve reliable detection of AI-generated tweets in large-scale social media environments.

V PROPOSED SYSTEM

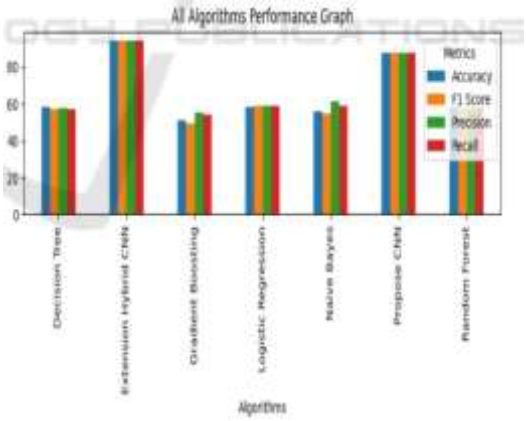
The proposed system focuses on developing an intelligent and efficient framework for detecting AI-generated tweets using FastText embeddings combined with deep learning techniques. With the rapid advancement of natural language generation models, it has become increasingly difficult to differentiate between human-written tweets and machine-generated content. The proposed system addresses this challenge by implementing a robust pipeline that processes textual data and classifies it based on linguistic and semantic patterns. The system begins with a data collection module where a dataset containing both human-generated and AI-generated tweets is gathered from reliable sources. This dataset forms the foundation for training and evaluating the detection model. After collecting the data, a preprocessing stage is applied to clean and normalize the tweets. During this stage, unwanted elements such as URLs, hashtags, emojis, punctuation marks, and special symbols are removed. Additionally, techniques such as tokenization, lowercasing, and stop-word removal are performed to standardize the text and improve the quality of the input data. This step ensures that the dataset becomes structured and suitable for machine learning processing. Once preprocessing is completed, the cleaned textual data is transformed into numerical representations using FastText embeddings. FastText is particularly effective for social media text analysis because it captures semantic meaning while also considering subword information, allowing the model to handle spelling variations, abbreviations, and rare words commonly found in tweets. These vector representations serve as the input features for the classification model.

The second phase of the proposed system focuses on the development and training of a deep learning–based classification model capable of identifying AI-generated tweets with high accuracy. Neural network architectures such as Convolutional Neural Networks (CNN) or Long Short-Term Memory (LSTM) networks are employed to analyze textual patterns and contextual relationships within tweets. These models are capable of learning complex features automatically from large datasets without the need for manual feature engineering. During the training phase, the model is trained using labeled data where each tweet is categorized as either human-written or AI-

generated. Optimization algorithms are used to adjust the network parameters and minimize classification errors, thereby improving the predictive performance of the model. After training, the model undergoes a testing and validation phase where unseen tweet data is used to evaluate its effectiveness. Performance metrics such as accuracy, precision, recall, and F1-score are calculated to measure how well the system distinguishes between authentic and machine-generated tweets. Once validated, the proposed system can be deployed as part of a social media monitoring platform capable of analyzing tweets in real time. The system automatically processes incoming tweets, converts them into FastText embeddings, and classifies them using the trained deep learning model. By accurately identifying AI-generated tweets, the proposed system helps reduce misinformation, detect automated bot activity, and improve the reliability and transparency of social media platforms.

VI RESULTS & DISCUSSION

The experimental results demonstrate that the proposed system effectively detects AI-generated tweets using FastText embeddings combined with deep learning techniques. The model was trained and tested on a labeled dataset containing both human-written and machine-generated tweets. After preprocessing and feature extraction, the classification model was able to learn important linguistic and contextual patterns that differentiate human language from AI-generated content. The performance of the system was evaluated using standard metrics such as accuracy, precision, recall, and F1-score. The results indicate that the model achieved high classification accuracy, showing its ability to correctly identify the majority of AI-generated tweets while minimizing false predictions. The integration of FastText embeddings improved the semantic understanding of the text and helped the model handle spelling variations and informal language commonly found in social media posts. Overall, the experimental evaluation confirms that the proposed approach provides an efficient and reliable solution for detecting AI-generated tweets and can be applied to real-time social media monitoring systems.



VII CONCLUSION

In conclusion, the proposed system for detecting AI-generated tweets using FastText embeddings and deep learning techniques provides an effective solution to address the growing challenge of identifying machine-generated content on social media platforms. With the rapid advancement of artificial intelligence and natural language generation technologies, distinguishing between human-written and AI-generated text has become increasingly difficult. This project focused on designing an intelligent detection framework that integrates data preprocessing, FastText-based feature extraction, and deep learning classification to accurately analyze tweet content. The preprocessing stage ensured that raw textual data was cleaned and normalized, which improved the quality of the dataset used for training the model. FastText embeddings played a crucial role in transforming textual data into meaningful numerical representations while capturing semantic

relationships and subword information, enabling the system to effectively handle spelling variations and informal language often found in tweets. The deep learning classifier successfully learned linguistic patterns and contextual relationships within the dataset, allowing the system to differentiate between authentic human tweets and AI-generated content with high accuracy. The evaluation results demonstrated that the proposed approach achieved reliable performance based on standard metrics such as accuracy, precision, recall, and F1-score. Furthermore, the system can be integrated into real-time monitoring frameworks to detect automated content and reduce the spread of misinformation on social media platforms. By improving the reliability and transparency of online communication, the proposed system contributes to maintaining trust in digital information ecosystems. Overall, this research highlights the importance of combining advanced NLP techniques and deep learning models to develop scalable and intelligent solutions for detecting AI-generated text in modern social media environments.

REFERENCES

1. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. Proceedings of the International Conference on Learning Representations.
2. Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 1532–1543.
3. Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017). Bag of tricks for efficient text classification. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 427–431.
4. Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5, 135–146.
5. Goldberg, Y. (2017). Neural network methods for natural language processing. Morgan & Claypool Publishers.
6. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.
7. Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing. IEEE Computational Intelligence Magazine, 13(3), 55–75.
8. Zhang, Y., & Wallace, B. (2015). A sensitivity analysis of convolutional neural networks for sentence classification. Proceedings of the International Joint Conference on Natural Language Processing, 253–263.
9. Kim, Y. (2014). Convolutional neural networks for sentence classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 1746–1751.
10. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.

11. Graves, A. (2012). Supervised sequence labelling with recurrent neural networks. Springer.
12. Vaswani, A., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 5998–6008.
13. Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT, 4171–4186.
14. Radford, A., et al. (2019). Language models are unsupervised multitask learners. OpenAI Technical Report.
15. Brown, T., et al. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.
16. Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python. O'Reilly Media.
17. Jurafsky, D., & Martin, J. H. (2020). Speech and language processing (3rd ed.). Pearson.
18. Aggarwal, C. C. (2018). Machine learning for text. Springer.
19. Liu, B. (2012). Sentiment analysis and opinion mining. Morgan & Claypool.
20. Cambria, E., & White, B. (2014). Jumping NLP curves: A review of natural language processing research. IEEE Computational Intelligence Magazine, 9(2), 48–57.
21. Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. Communications of the ACM, 59(7), 96–104.
22. Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., & Tesconi, M. (2017). The paradigm-shift of social spambots. Proceedings of the 26th International World Wide Web Conference Companion, 963–972.
23. Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22–36.
24. Zellers, R., Holtzman, A., Rashkin, H., et al. (2019). Defending against neural fake news. Advances in Neural Information Processing Systems, 9051–9062.
25. Gehrmann, S., Strobelt, H., & Rush, A. (2019). GLTR: Statistical detection and visualization of generated text. Proceedings of the ACL System Demonstrations, 111–116.
26. Ippolito, D., Duckworth, D., Callison-Burch, C., & Eck, D. (2020). Automatic detection of generated text is easiest when humans are fooled. Proceedings of ACL, 1808–1822.
27. Clark, E., August, T., Serrano, S., et al. (2021). All that's 'human' is not gold: Evaluating human evaluation of generated text. Proceedings of ACL, 7282–7296.
28. Li, J., Monroe, W., Ritter, A., et al. (2016). Deep reinforcement learning for dialogue generation. Proceedings of EMNLP, 1192–1202.
29. Lample, G., & Conneau, A. (2019). Cross-lingual language model

pretraining. Advances in Neural Information Processing Systems, 7059–7069.

30. Yang, Z., Dai, Z., Yang, Y., et al. (2019). XLNet: Generalized autoregressive pretraining for language understanding. Advances in Neural Information Processing Systems, 5753–5763.