*Journal of Science Engineering Technology and Management Science*
*Volume 02, Issue 04, April 2025* *ISSN: 3049-0952*
*www.jsetms.com* *DOI:10.63590/jsetms.2025.v02.i04.87-93*

# DETECTING DEEP FAKE TWEETS WITH FAST TEXT EMBEDDINGS AND CNN-DRIVEN RANDOM FORESTS

**Dr. P. Babu, A. Sruthi, D. Sumanvitha, J. Jasmitha**

*Department of Computer Science and Engineering (AI&ML), Geethanjali Institute of Science and Technology, Nellore, Andhra Pradesh, India.*

***To Cite this Article***

*Dr. P. Babu, A. Sruthi, D. Sumanvitha, J. Jasmitha,“ Detecting Deep Fake Tweets With Fast Text Embeddings And CNN-Driven Random Forests”, Journal of Science Engineering Technology and Management Science, Vol. 02, Issue 04, April 2025,pp: 87-93, DOI:* http://doi.org/10.63590/jsetms.2025.v02.i04.pp87-93

*Submitted: 09-03-2025* *Accepted: 18-04-2025* *Published: 26-04-2025*

## ABSTARCT

This research presents an advanced framework for detecting fake tweets using a dedicated Twitter fake tweet dataset. The goal is to improve detection accuracy by combining natural language processing (NLP) and machine learning techniques. Traditional manual detection methods face several limitations, including time-consuming analysis, susceptibility to subjective bias, and poor scalability in the face of the rapid growth of social media data. These methods often fail to identify subtle linguistic and contextual cues associated with misinformation. The proposed approach addresses these challenges through a multi-step pipeline. It begins with NLP preprocessing to clean and standardize the tweets, followed by FastText embeddings to transform textual data into numerical vectors for detailed analysis. The dataset is then split into training and testing sets to ensure robust evaluation. A deep learning convolutional neural network (DLCNN) is used for feature extraction, capturing complex patterns within the data. These features are then classified using a Random Forest classifier to determine whether tweets are real or fake. The model's performance is rigorously evaluated using appropriate metrics, demonstrating its effectiveness and reliability in real-world scenarios.

**Keywords**: Convolutional Neural Network, Random Forest Classifier, Tweet Classification, Fake tweet Detection, Machine Learning, Deep learning.

## 1.INTRODUCTION

Online Social Networks (OSNs) have revolutionized digital communication, providing a platform for instant information sharing, discussion, and interaction. Among these, Twitter stands out as one of the most influential social media platforms due to its vast user base and real-time content dissemination. However, the rapid growth of OSNs has also led to the increasing spread of misinformation and deep fake content, particularly in the form of machine-generated tweets. This issue has significantly impacted various sectors, including politics, education, and public trust. India, with one of the highest numbers of social media users, is particularly vulnerable to the negative consequences of deep fakes. Political misinformation, fake news campaigns, and manipulated narratives have influenced public opinion and even national elections. Students, who rely on OSNs for news, academic discussions, and social interactions, are especially susceptible to such deceptive content, leading to misinformation, academic dishonesty, and misinformed decision-making. Furthermore, the rampant spread of synthetic media has eroded public trust in digital communication, making it difficult for users to distinguish between real and fake content.

Given Twitter's dominance as a primary communication tool for news, politics, and social interactions, detecting machine-generated tweets is critical. Traditional moderation techniques and manual fact-checking are insufficient to tackle the scale and sophistication of AI-generated content. Therefore, leveraging deep learning models combined with Fast Text embeddings can significantly enhance the accuracy and efficiency of detecting synthetic tweets. Fast Text embeddings allow models to capture contextual and sub-word-level information, making them highly effective in identifying unnatural patterns in text. One of the biggest consequences of deep fake content is the erosion of public trust in OSNs. When users are repeatedly exposed to misleading information, fake news, and AI-generated tweets, their confidence in social media as a reliable source of information diminishes. This decline in trust is particularly concerning for governments, media organizations, and businesses that rely on OSNs to communicate with the public.

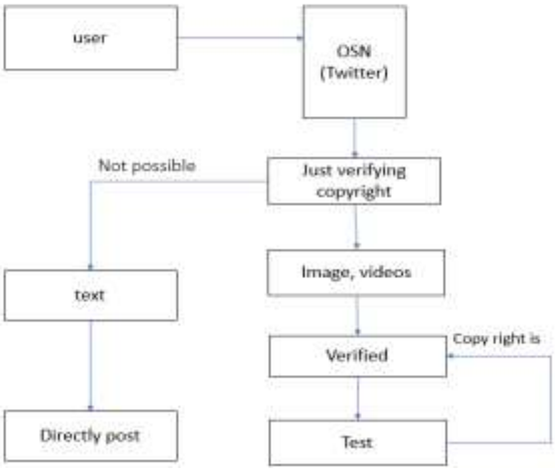


Fig. 1: Block diagram of fraud detection in twitter.

Even traditional fact-checking methods struggle to keep up with the volume of fake content being generated. This has led to growing concerns over the credibility of social media platforms and their ability to regulate misinformation effectively.

## 2. LITERATURE SURVEY

Khalid et al. [1] analyze AI-generated texts in the form of deepfake "tweets". These are evaluated using the logistic regression model, for which a maximum accuracy of 98.9% was achieved. Sentiment classification was validated through k-fold cross-validation [2]. The article by Al-Khazrajı et al. [3] also describes research on the impact of deepfakes on social networks. The discussions on this topic conclude by stating the necessity of collaboration between researchers and deeper education of the population. The analysis by Ahmed et al. [4] was conducted in eight countries. The factors analyzed were perception of accuracy, fear of missing out (FOMO), deficient self-regulation (DSR), and cognitive ability. The results showed that individuals with high FOMO or DSR are prone to sharing deepfakes, regardless of their cognitive level. Users with reduced cognitive abilities tend to distribute misinformation.

Mubarak et al. [5] classify the deepfakes into visual, audio, and textual deepfakes. Patel et al. [6] analyze different models and datasets used, implementation challenges, and certain future directions. These models are specific to image-based deepfakes [7]. One special category of deepfakes is audio deepfakes. Sunkari and Srinagesh [8] propose a model for combating audio deepfakes. The model uses a convolutional neural network (CNN) architecture to extract spectral features and a recurrent neural network (RNN) to analyze temporal dynamics.

Other models, such as large language models (LLMs), are used to identify text-based deepfakes. The article by De Angelis et al. [9] aims to study the Chat Generative Pre-Training Transformer (ChatGPT) phenomenon with respect to the spreading of deepfakes using LLMs. Koike et al. [10] proposed a model that identified text-based deepfakes. Following the research, the OUTFOX model achieved a performance of 96.9%. The article by Veerasamy and Pieterse [11] presents five areas of interest that highlight the major risks generated by deepfakes. The study suggests a holistic approach to stopping the danger posed by deepfakes. The paper by de Ruiter [12] emphasizes that deepfakes are not inherently dangerous but can become so when used by individuals with malicious intent toward society.

Hameleers et al. [13] compare classic disinformation and deepfake disinformation through various tests. The tests were conducted in the Netherlands, using textual manipulation and deepfakes, and considered the manner of appearance on social media (absent, supported, or discredited). The conclusion of the study highlights the danger posed by this technology. Twomey et al. [14] focus on the fact that deepfake technology has managed to confuse many people, to the point where the population can no longer distinguish between something real and something that is a deepfake. A total of 4869 tweets from the time of the Russia–Ukraine war were analyzed. The results showed that people interpreted most samples with real information as deepfakes. It was discovered that doubt about deepfakes sparked many conspiracies, which indicates massive ignorance among people regarding deepfakes. The research aimed to raise the alarm among governments, social platforms, and the media about educating the population.

The research by Vaccari and Chadwick [15] shows the degree of disinformation that deepfakes have, in contrast with news, on social media. Following the experiment conducted in the UK, it was concluded that people are more likely to be skeptical than deceived when dealing with a deepfake. Rafique et al. [16] highlight the danger brought by technological advancements through the creation of deepfakes. The authors propose a solution for classifying real information from deepfakes using deep learning and machine learning, achieving an accuracy of 89.5%.

The work by Uchendu et al. [17] analyzes the human capacity to detect deepfake texts. The research included experts and non-experts. These individuals attempted to detect deepfake text paragraphs. The research concluded that human expertise is the most important factor in detecting text-based deepfakes. The study [18] takes a perspective different from those of other researchers regarding deepfake analysis. Given the complexity of deepfake detection, the article adopts a human-centered perspective on the identification process. The study provides two results: the first is that people are unable to identify a deepfake correctly, and the second is that people overestimate their detection abilities. Their conclusion states the need for further research on the similarities between human and machine detection processes.

In contrast to these approaches, Kaur et al. [19] propose an AI model using a CNN architecture to identify deepfakes. The results showed a detection performance of 91%. The article by Sadiq et al. [20] also uses a CNN architecture, one based on FastText, to identify deepfakes on the Twitter platform. Following a comparison with other models, such as long short-term memory (LSTM) and CNN-LSTM, the proposed model achieved a detection performance of 93% for deepfake texts.

## 3.PROPOSED SYSTEM

The proposed system architecture as shown in Fig. 2. for AI-based deep fake detection follows a structured Django MVT framework, integrating user authentication, database management, and AI-driven text classification. Users can sign up and log in, with their credentials securely stored in an SQL database. Once authenticated, they can submit tweets for analysis, which undergo NLP preprocessing, FastText embedding conversion, deep feature extraction using CNNs, and final classification via a Random Forest model to determine if a tweet is human-written (Normal) or AI-generated (Fake). The results are stored in the database and displayed on a user-friendly HTML interface, showing the original tweet, classification label, and confidence score. The system ensures efficient data retrieval, history tracking, and real-time deepfake detection, helping users combat misinformation effectively while maintaining a seamless and secure experience. The system follows a structured algorithm to detect AI-generated (fake) tweets using FastText embeddings, Deep Learning CNN (DLCNN), and Random Forest Classifier (RFC) within a Django-based web application.

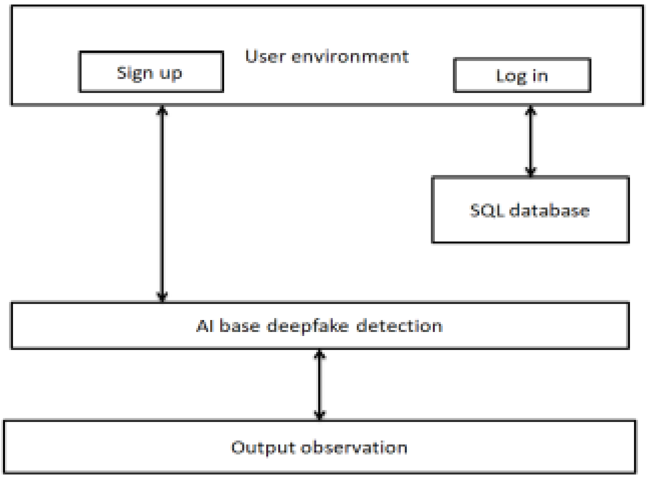


Fig. 2: Proposed Django architecture.

This research uses a hybrid model that combines a Convolutional Neural Network (CNN) with a Random Forest Classifier (RFC) to detect fake tweets. First, tweets are converted into numerical form using FastText embeddings, which help the model understand word meaning, context, and even misspelled or rare words. The CNN is then used to extract important features from the tweets. It performs a series of operations such as convolution, activation, pooling, and flattening to identify patterns in the text data as in Fig. 3. These features represent the core information of each tweet, making it easier to analyze and classify.

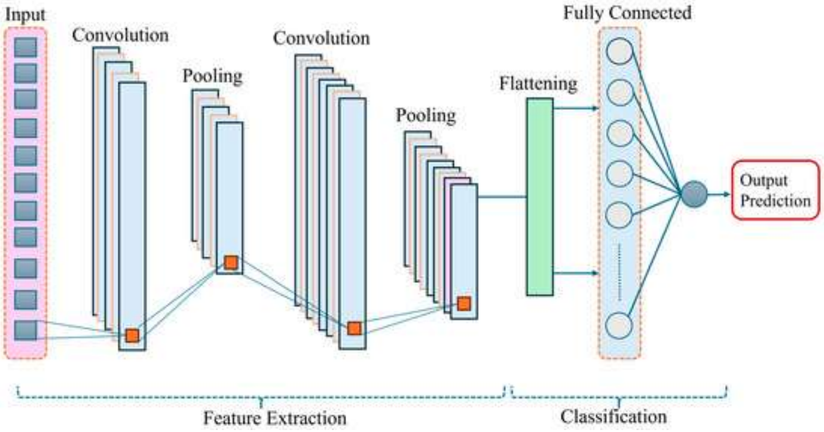
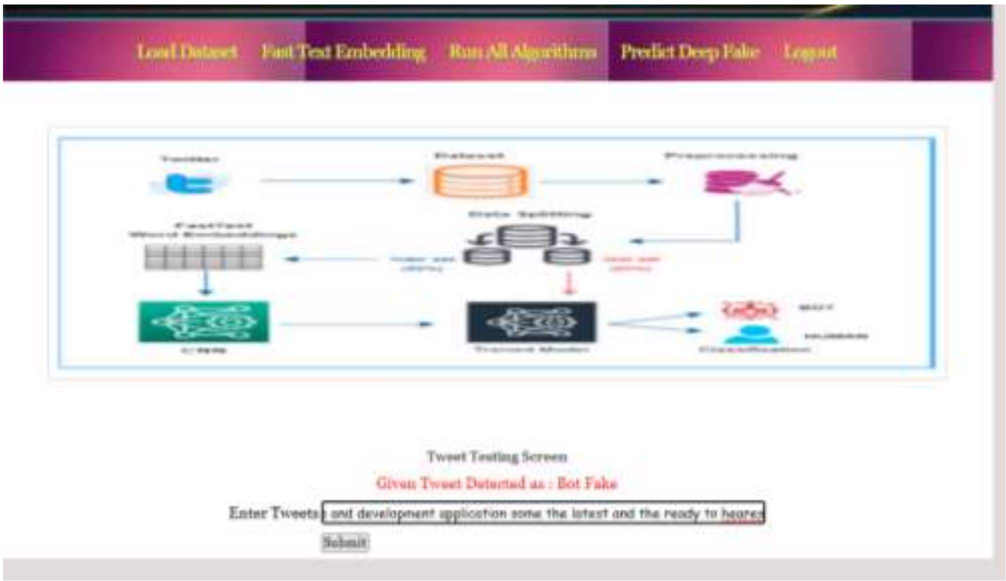


Fig. 3: Convolutional Neural Networks.

Instead of using the CNN alone for classification, the extracted features are passed into a Random Forest Classifier. RFC works by building many decision trees on different parts of the data and then combining their results through majority voting to make a final decision. This combination improves accuracy, reduces overfitting, and handles imbalanced data well—especially when fake tweets are much fewer than real ones. Overall, the CNN helps in deeply analyzing the text, while the RFC provides strong and reliable classification, making this hybrid model effective and efficient for fake tweet detection.

## 4. RESULTS AND DISCUSSION

Fig. 4 represents the Deepfake Detection System for social media, which uses deep learning and FastText embeddings to classify tweets as human-generated or bot- generated. The prediction interface allows users to input a tweet into the system, which then processes it through several stages: preprocessing, FastText word embeddings, and CNN-based classification. The trained model analyzes the input tweet and determines whether it originates from a bot or a human. This automated system ensures efficient deepfake detection by leveraging contextual word representations and neural network- based classification, helping mitigate misinformation on social media platforms.



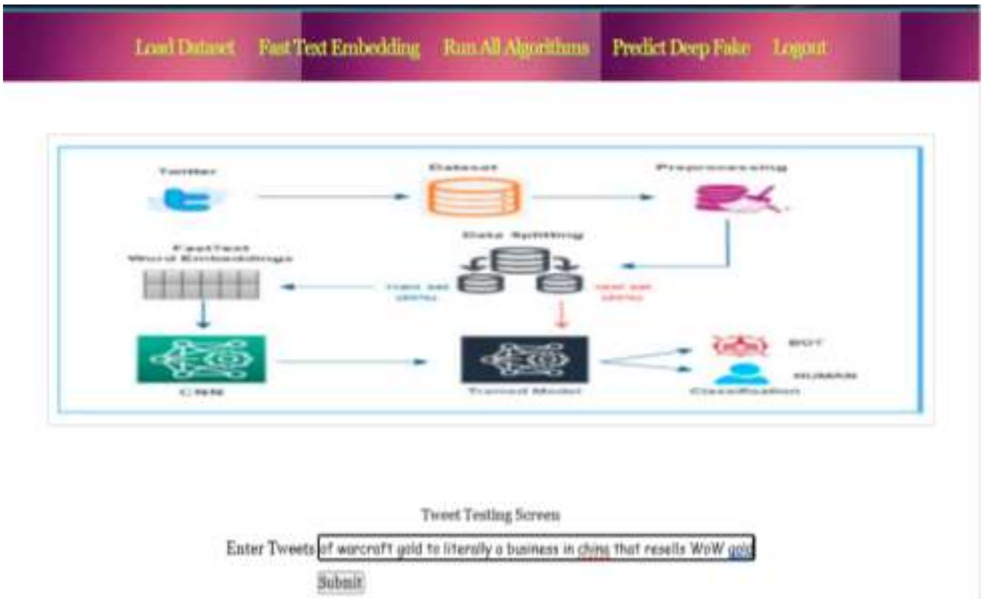
Tweet Testing Screen

Given Tweet Detected as : Bot Fake

Enter Tweets and development application some the latest and the ready to heared

Submit

Fig. 4: Prediction on test data-Deep Fake.

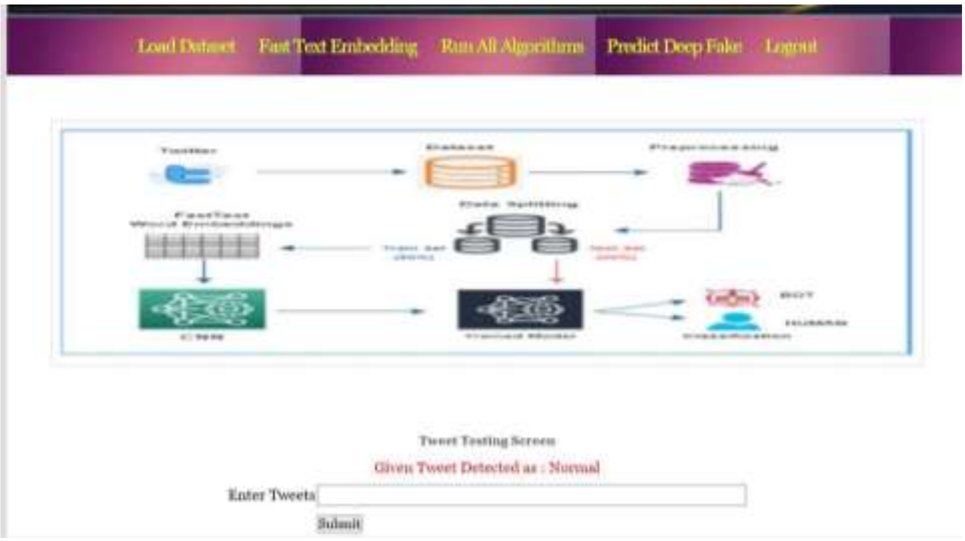


Tweet Testing Screen

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

Submit

Fig. 5 represents the Deepfake Detection System's output for a given tweet. The system takes an input tweet, processes it through FastText embeddings and a CNN-based classification model, and determines whether the tweet is generated by a bot or a human. In this specific case, the system has classified the given tweet as "Normal," indicating that it is likely human-generated rather than a deepfake or bot-generated content. The detection workflow involves data preprocessing, feature extraction using FastText, training with CNN, and final classification, ensuring an accurate prediction based on learned patterns.



Tweet Testing Screen

Given Tweet Detected as : Normal

Enter Tweets

Fig. 5: Output for the given label.

**REFERENCES**

[1] Khalid, M.; Raza, A.; Younas, F.; Rustam, F.; Villar, M.G.; Ashraf, I.; Akhtar, A. Novel Sentiment Majority Voting Classifier and Transfer Learning-Based Feature Engineering for Sentiment Analysis of Deepfake Tweets. IEEE Access 2024, 12, 67117–67129.

[2] Rosca, C.M.; Ariciu, A.V. Unlocking Customer Sentiment Insights with Azure Sentiment Analysis: A Comprehensive Review and Analysis. Rom. J. Pet. Gas Technol. 2023, 4, 173–182.

[3] Al-Khazraji, S.H.; Saleh, H.H.; Khalid, A.I.; Mishkhal, I.A. Impact of Deepfake Technology on Social Media: Detection, Misinformation and Societal Implications. Eurasia Proc. Sci. Technol. Eng. Math. 2023, 23, 429–441.

[4] Ahmed, S.; Ng, S.W.T.; Bee, A.W.T. Understanding the role of fear of missing out and deficient self-regulation in sharing of deepfakes on social media: Evidence from eight countries. Front. Psychol. 2023, 14, 1127507. [PubMed]

[5] Mubarak, R.; Alsboui, T.; Alshaikh, O.; Inuwa-Dutse, I.; Khan, S.; Parkinson, S. A Survey on the Detection and Impacts of Deepfakes in Visual, Audio, and Textual Formats. IEEE Access 2023, 11, 144497–144529.

[6] Patel, Y.; Tanwar, S.; Gupta, R.; Bhattacharya, P.; Davidson, I.E.; Nyameko, R.; Aluvala, S.; Vimal, V. Deepfake Generation and Detection: Case Study and Challenges. IEEE Access 2023, 11, 143296–143323.

[7] Rosca, C.M. Comparative Analysis of Object Classification Algorithms: Traditional Image Processing Versus Artificial Intelligence—Based Approach. Rom. J. Pet. Gas Technol. 2023, IV (LXXV), 169–180.

[8] Sunkari, V.; Srinagesh, A. Efficient Deepfake Audio Detection Using Spectro-Temporal Analysis and Deep Learning. J. Electr. Syst. 2024, 20, 10–18.

[9] De Angelis, L.; Baglivo, F.; Arzilli, G.; Privitera, G.P.; Ferragina, P.; Tozzi, A.E.; Rizzo, C. ChatGPT and the rise of large language models: The new AI-driven infodemic threat in public health. Front. Public Health 2023, 11, 1166120.

[10] Koike, R.; Kaneko, M.; Okazaki, N. OUTFOX: LLM-Generated Essay Detection Through In-Context Learning with Adversarially Generated Examples. Proc. AAAI Conf. Artif. Intell. 2024, 38, 21258–21266.

[11] Veerasamy, N.; Pieterse, H. Rising Above Misinformation and Deepfakes. Int. Conf. Cyber Warf. Secur. 2022, 17, 340–348.

[12] de Ruiter, A. The Distinct Wrong of Deepfakes. Philos. Technol. 2021, 34, 1311–1332.

[13] Hameleers, M.; Van Der Meer, T.G.L.A.; Dobber, T. You Won't Believe What They Just Said! The Effects of Political Deepfakes Embedded as Vox Populi on Social Media. Soc. Media + Soc. 2022, 8, 20563051221116346.

[14] Twomey, J.; Ching, D.; Aylett, M.P.; Quayle, M.; Linehan, C.; Murphy, G. Do deepfake videos undermine our epistemic trust? A thematic analysis of tweets that discuss deepfakes in the Russian invasion of Ukraine. PLoS ONE 2023, 18, e0291668.

[15] Vaccari, C.; Chadwick, A. Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. Soc. Media + Soc. 2020, 6, 205630512090340.

[16] Rafique, R.; Gantassi, R.; Amin, R.; Frnda, J.; Mustapha, A.; Alshehri, A.H. Deep fake detection and classification using error-level analysis and deep learning. Sci. Rep. 2023, 13, 7422.

[17] Uchendu, A.; Lee, J.; Shen, H.; Le, T.; Huang, T.-H.K.; Lee, D. Does Human Collaboration Enhance the Accuracy of Identifying LLM-Generated Deepfake Texts? Proc. AAAI Conf. Hum. Comput. Crowdsourcing 2023, 11, 163–174.

[18] Köbis, N.C.; Doležalová, B.; Soraperra, I. Fooled twice: People cannot detect deepfakes but think they can. iScience 2021, 24, 103364.

[19] Kaur, J.; Vinay, P.S.; Reddy, G.S.K.; Sai, P.V.S.; Reddy, V.M.M. Deepfake on Social Media: Harnessing Deep Learning for Identifying Falsified Tweet. Int. J. Sci. Res. Eng. Manag. 2024, 8, 1–8.

[20] Sadiq, S.; Aljrees, T.; Ullah, S. Deepfake Detection on Social Media: Leveraging Deep Learning and FastText Embeddings for Identifying Machine-Generated Tweets. IEEE Access 2023, 11, 95008–95021.