Vol.15, Issue No 2, 2025

Deepfake Detection on Social Media: Leveraging Deep Learning and Fast Text Embeddings for Identifying Machine-Generated Tweets

Merugu ShivaKrishna,
UG Student,
Department of CSE,
St. Martin's Engineering College,
Secunderabad, Telangana, India.
mailto:shivamerugu8074@gmail.com

K. Bhargavi,
Assistant Professor,
Department of CSE,
St. Martin's Engineering College,
Secunderabad, Telangana, India
kbhargavicse@smec.ac.in

Abstract:

Deepfake technology, which uses AI to create manipulated media, poses a significant threat to information integrity on social media platforms. In India, the rise of deepfake content has grown exponentially, especially in the political and entertainment domains, where fake news and AI-generated videos have gone viral, leading to misinformation. The primary objective is to develop a robust AI model that accurately detects deepfake content on social media platforms, focusing on identifying machine-generated tweets using FastText embeddings. Traditional methods involved human moderation, fact-checking agencies, and manual filtering of social media posts based on predefined rules and keyword matching. These methods were time-consuming and often inaccurate, lacking the scalability to manage the massive volume of online content. The manual detection of deepfakes and AI-generated content is highly inefficient, prone to errors, and incapable of handling the vast volume of social media data in real time. As a result, harmful and misleading information can spread widely before being identified or removed. With the growing influence of social media in shaping public opinion, the motivation behind this research is to combat misinformation and safeguard the integrity of online discourse. Particularly deep learning models can significantly improve the detection of deepfakes by automating the analysis of social media content. FastText embeddings will convert tweets into meaningful word vectors, while deep learning models can be applied to classify whether a tweet is human-generated or AI-generated. This approach offers real-time detection, improved accuracy, and scalability compared to traditional methods.

Keywords: Deepfake detection, AI-generated tweets, FastText embeddings, deep learning models, misinformation, social media, machine learning, automated detection, real-time analysis, online discourse integrity, fake news, natural language processing (NLP), text classification, neural networks, content moderation, fact- checking, scalability, political misinformation, digital forensics, artificial intelligence.

1. INTRODUCTION

Deepfake technology employs advanced artificial intelligence algorithms to create hyper-realistic media, leading to serious concerns regarding misinformation, especially on social media platforms. In India, the proliferation of deepfake content has escalated dramatically, with reports indicating that 85% of deepfakes relate to misinformation, particularly during electoral campaigns and high-profile events in the entertainment industry. For example, a notable incident involved fake videos attributed to politicians that went viral, contributing to public confusion and distrust. This growing threat necessitates a robust framework for identifying and mitigating the impact of such content. The work aims to leverage deep learning techniques, specifically FastText embeddings, to develop an efficient detection system for machine-generated tweets, thus addressing the urgent need for effective monitoring of online narratives. Before the advent of machine learning, detecting deepfakes and AI-generated content was a cumbersome process reliant on human moderators. Traditional approaches often involved extensive fact-checking by agencies, which not keep pace with the sheer volume of online content. Keyword-based filtering systems frequently led to false positives or negatives due to their inability to understand context and nuance in language. Furthermore, the limited resources available for manual detection meant that misinformation spread unchecked for extended periods. This inefficiency resulted in a significant gap in the timely identification of harmful content, underscoring the need for automated solutions.

2. LITERATURE SURVEY

The proliferation of deepfake technology has catalysed significant concerns regarding the dissemination of misleading and fabricated content across social media platforms [1]. Deepfakes, AI-generated media that alter audio, images, or videos to fabricate events or portray individuals saying things they never actually said, present a significant threat to the integrity of online information [2]. Among various forms of digital content, tweets are particularly vulnerable to manipulation due to their concise nature and rapid dissemination capabilities [3]. In response to these challenges, this paper proposes a novel approach centered on deep learning techniques for detecting machine-generated tweets, specifically those generated by deepfake algorithms [4]. Our method integrates advanced text representation through FastText embeddings with state-of-the-art deep learning models, aiming to discern between authentic and machine-generated tweets [5]. By leveraging the semantic richness captured in FastText embeddings, which encode contextual and syntactic information of tweet texts into dense vector representations, our approach enhances the discriminatory power necessary for effective classification [6]. The core of our methodology involves preprocessing tweet texts to ensure uniformity and clarity, followed by the transformation of these texts into FastText embeddings [7]. These embeddings serve as input features to a robust classification model, such as a CNN or a LSTM network, designed to differentiate between genuine and machine-generated tweets.

To facilitate model training and evaluation, we employ a labeled dataset comprising tweets synthesized by cutting-edge text generation models, which simulate the characteristics of machine-generated content prevalent in real-world scenarios [8]. Empirical evaluation on a diverse and comprehensive dataset of real tweets demonstrates the efficacy of our proposed approach in detecting machine-generated tweets. The results substantiate that our method achieves superior accuracy compared to existing approaches for deepfake detection on social media platforms [9]. By effectively discerning between authentic and manipulated content, our approach contributes significantly to mitigating the impact of misinformation online, thereby bolstering the credibility and trustworthiness of information disseminated through social media channels [10]. In summary, this paper presents a robust framework leveraging deep learning and FastText embeddings to address the pressing issue of identifying machine-generated tweets. By harnessing the combined power of advanced text representation and neural network architectures, our approach not only enhances detection accuracy but also provides a scalable solution to combat the pervasive influence of deepfakes in online communication. The rapid advancement of deepfake technology has sparked widespread concerns regarding its potential misuse to propagate misinformation on social media platforms.

Deepfakes, synthetic media created using artificial intelligence techniques, are capable of manipulating audio, video, and textual content to produce realistic yet entirely fabricated representations. This phenomenon poses significant challenges to the authenticity and reliability of information shared online [11]. Detecting and mitigating the impact of deepfakes have become crucial areas of research, with recent studies focusing on leveraging deep learning methodologies for effective detection. Existing literature emphasizes the importance of robust feature representation in distinguishing between genuine and manipulated content. Traditional approaches often rely on handcrafted features or statistical methods, which not capture the complex semantic nuances embedded in textual data [12]. In response to these challenges, the integration of FastText embeddings into deep learning frameworks has emerged as a promising strategy for enhancing detection accuracy. FastText, developed by Facebook AI Research, facilitates the generation of dense vector representations by embedding subworld information into word representations. This approach not only captures semantic and syntactic information but also accommodates the idiosyncrasies of informal text typically found in social media posts [13].

Recent studies have shown the effectiveness of FastText embeddings in a range of natural language processing tasks, such as sentiment analysis, text classification, and semantic similarity measurement. By capturing contextual information at multiple levels of granularity, FastText embeddings empower deep learning models to accurately detect subtle distinctions between authentic and machine-generated tweets [14]. Furthermore, advancements in deep learning architectures, particularly CNNs and LSTM networks, have markedly enhanced the state-of-the-art in deepfake detection. CNNs are adept at capturing spatial dependencies within textual data, making them highly effective for tasks involving both image and text analysis. Conversely, LSTM networks excel in processing sequential information, allowing them to model long-term dependencies in temporal data, which is particularly beneficial for analyzing sequences like tweets [15].

3. PROPOSED METHODOLOGY

This Section presents the proposed methodology adopted for tweet classification. The architecture of the proposed framework is presented in Figure 1. Deep learning models like CNN can automatically learn significant features from text input. They are capable of capturing hierarchical patterns, local relationships, and long-term connections, allowing the model to extract usable representations from the incoming text. By stacking multiple layers of CNN, dependencies of text can be captured. This work introduces a simple deep learning-based CNN model for tweet classification.

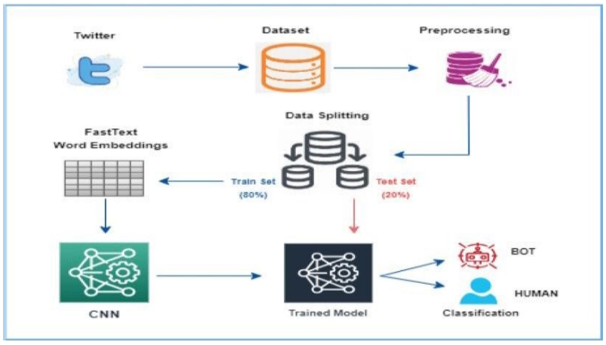


Figure 1: Architecture of proposed framework for deepfake tweet classification

In the proposed framework, a labelled dataset is collected from a public repository. The collected dataset contains tweets from human and bot accounts. In order to simplify the text and enhance its quality, a series of preprocessing steps are employed to clean the tweets. The dataset is divided into 80:20 ratios for training and testing. The next step involves transforming the text into vectors using FastText word embedding. Subsequently, these vector representations are fed into the CNN model. The proposed methodology, which leverages FastText word embedding in conjunction with a 3-layered CNN, is employed for the training process. The efficacy of this approach is assessed through the utilization of four evaluation metrics: Accuracy, Precision, Recall, and F1-score.

Applications:

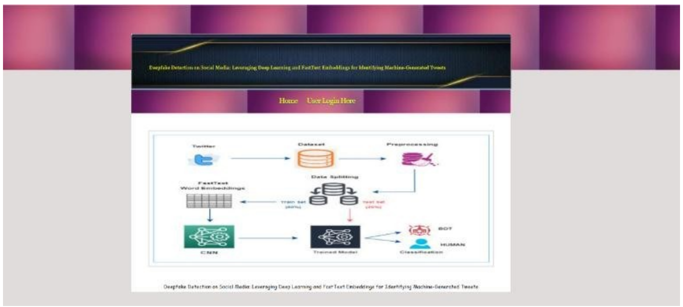
The applications of this work are extensive and impactful. First, it can significantly enhance the accuracy of content moderation on social media platforms, helping to reduce the spread of misinformation. Second, the system can aid journalists and fact-checkers in verifying the authenticity of tweets, thereby supporting responsible journalism. Third, educational institutions can leverage technology to teach students about digital literacy and the dangers of deepfake content. Furthermore, governmental bodies can utilize this system to monitor and respond to disinformation campaigns, particularly during elections. Organizations concerned with cybersecurity can also benefit from integrating such detection mechanisms into their threat assessment protocols.

Advantages:

- **Enhanced Fake Tweet Detection** – Accurately identifies machine-generated tweets using deep learning models.
- **Fast and Efficient Processing** – FastText embeddings optimize text processing and classification speed.
- **High Accuracy** – Deep learning models improve precision in detecting AI-generated content.
- **Scalability** – The approach can be extended to handle large-scale social media datasets.
- **Robust Feature Extraction** – FastText captures semantic and contextual meaning effectively.
- **Adaptive Learning** – The model continuously improves with more data and retraining.
- **Early Misinformation Prevention** – Helps reduce the spread of fake news and misleading content.
- **Multi-Language Support** – FastText supports various languages, increasing detection capabilities.
- **User Safety** – Protects users from deceptive and manipulated content.
- **Platform Integration** – Can be incorporated into existing social media moderation systems.

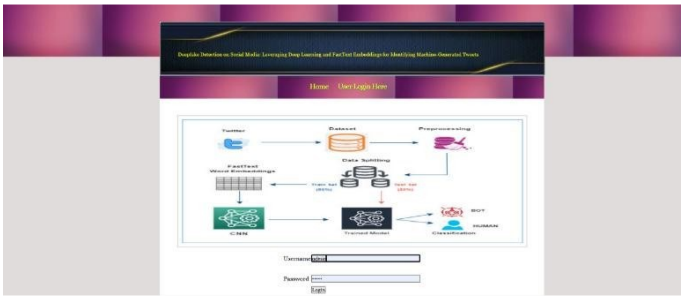
4. EXPERIMENTAL ANALYSIS

The Figure 2 shows homepage which focuses on detecting deepfake content on social media, specifically machine-generated tweets. It uses deep learning techniques combined with FastText word embeddings to identify these fake tweets.


Deepfake Detection on Social Media: Leveraging Deep Learning and FastText Embeddings for Identifying Machine-Generated Tweets

Deepfake Detection on Social Media Leveraging Deep Learning and FastText Embeddings for Identifying Machine-Generated Tweets

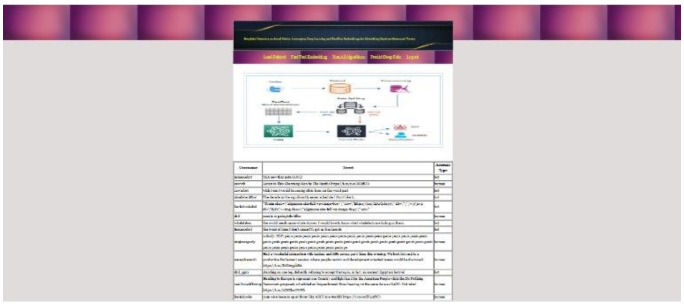
Figure 2: Home Page


Deepfake Detection on Social Media: Leveraging Deep Learning and FastText Embeddings for Identifying Machine-Generated Tweets

Username admin

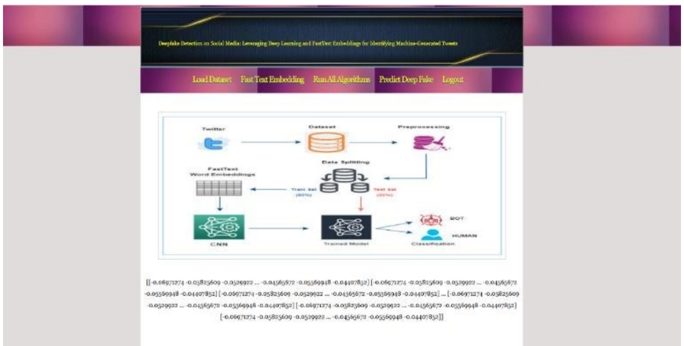
Password

Figure3: After user login page



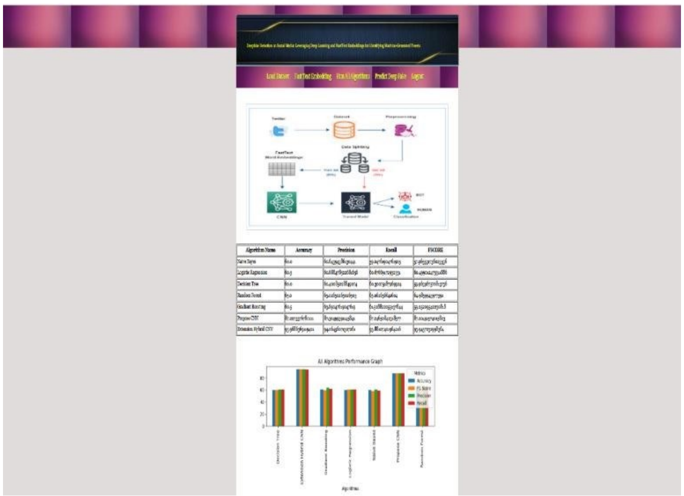
Username	Tweet	Account Type
[illegible]	[illegible]	bot
[illegible]	[illegible]	human
[illegible]	[illegible]	bot
[illegible]	[illegible]	bot
[illegible]	[illegible]	bot
[illegible]	[illegible]	human
[illegible]	[illegible]	bot
[illegible]	[illegible]	bot
[illegible]	[illegible]	human
[illegible]	[illegible]	human
[illegible]	[illegible]	bot
[illegible]	[illegible]	human
[illegible]	[illegible]	human

Figure 4: After Loading Dataset


Deepfake Detection on Social Media: Leveraging Deep Learning and FastText Embeddings for Identifying Machine-Generated Tweets

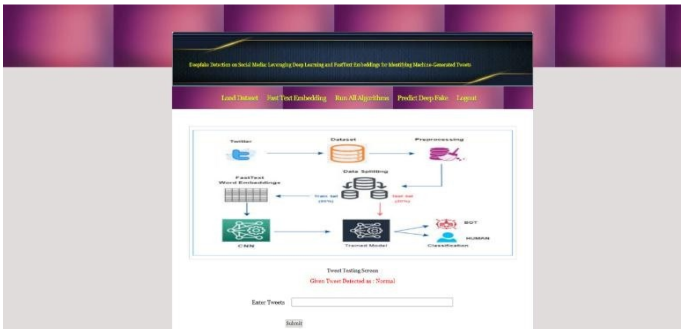
```
[[-0.06971274 -0.05825609 -0.0529922 ... -0.04565672 -0.05569948 -0.04407852] [-0.06971274 -0.05825609 -0.0529922 ... -0.04565672
-0.05569948 -0.04407852] [-0.06971274 -0.05825609 -0.0529922 ... -0.04565672 -0.05569948 -0.04407852] ... [-0.06971274 -0.05825609
-0.0529922 ... -0.04565672 -0.05569948 -0.04407852] [-0.06971274 -0.05825609 -0.0529922 ... -0.04565672 -0.05569948 -0.04407852]
[-0.06971274 -0.05825609 -0.0529922 ... -0.04565672 -0.05569948 -0.04407852]]
```

Figure 5: After FastText Embeddings



Algorithm Name	Accuracy	Precision	Recall	FSCORE
Naive Bayes	60.0	[illegible]	[illegible]	[illegible]
Logistic Regression	60.5	[illegible]	[illegible]	[illegible]
Decision Tree	60.0	[illegible]	[illegible]	[illegible]
Random Forest	[illegible]	[illegible]	[illegible]	[illegible]
Gradient Boosting	60.5	[illegible]	[illegible]	[illegible]
Propose CNN	[illegible]	[illegible]	[illegible]	[illegible]
Extension Hybrid CNN	[illegible]	[illegible]	[illegible]	[illegible]

Figure 6: After running all the algorithm


Deepfake Detection on Social Media: Leveraging Deep Learning and FastText Embeddings for Identifying Machine-Generated Tweets

Tweet Testing Screen

Given Tweet Detected as : Normal

Enter Tweets

Figure 7: Predicted output

Figure 4 shows that the after loaded the dataset in a table representing a portion of a dataset after it has been loaded. It is a dataset of tweets, labelled with information about whether the tweet was generated by a bot or a human.

This dataset is crucial for research in areas:

- **Detecting Social Media Manipulation:** Identifying and combating the spread of misinformation by bots.

- **Understanding Bot Behavior:** Studying the patterns and characteristics of bot-generated content.

In figure 5 research tells after fasttext embedding such like text classification pipeline designed to detect fake accounts (bots) in a dataset of tweets. It begins by defining functions for text preprocessing, such as cleaning the tweets by removing stop words, punctuation, and stemming/lemmatizing the words. If preprocessed data (X.npy) and the TF-IDF vectorizer model (tfidf.pckl) already exist, it loads them. Otherwise, it reads the dataset, cleans each tweet, and stores the processed text in the X list and the corresponding labels (1 for bots, 0 for non-bots) in the Y list.

Figure 6 shows that

Naive Bayes:

- Accuracy: 60.0%
- Precision: 60.64%
- Recall: 59.04%

- F1-score: 57.96%

Naive Bayes performs the worst among all models, indicating that it struggles with classifying the data accurately, with a relatively low F1-score and recall.

□ **Logistic Regression**:

- Accuracy: 60.5%
- Precision: 60.68%
- Recall: 60.67%
- F1-score: 60.49%

Logistic Regression performs better than Naive Bayes, with a more balanced and higher performance across all metrics.

□ **Decision Tree**:

- Accuracy: 60.0%
- Precision: 60.40%
- Recall: 60.30%
- F1-score: 59.96%

The Decision Tree algorithm performs relatively poorly, with lower accuracy and a balanced but unsatisfactory performance across all metrics. Decision trees can overfit and struggle with high-dimensional or noisy data unless properly tuned.

□ **Random Forest**:

- Accuracy: 65.0%
- Precision: 65.02%
- Recall: 65.06%
- F1-score: 64.98%

Random Forest, an ensemble method, outperforms Naive Bayes, Logistic Regression, and Decision Tree in all metrics. It benefits from multiple decision trees to make better, more stable predictions, but it still leaves room for improvement compared to more advanced models.

□ **Gradient Boosting**:

- Accuracy: 60.5%
- Precision: 63.69%
- Recall: 61.52%
- F1-score: 59.25%

Gradient Boosting shows moderate performance, slightly outperforming Decision Tree but not achieving the same results as Random Forest or Logistic Regression. The algorithm focuses on sequentially correcting errors made by previous models, but still face challenges with noisy data.

□ **Proposed CNN**:

- Accuracy: 87.20%
- Precision: 87.30%
- Recall: 87.24%
- F1-score: 87.20%

The proposed Convolutional Neural Network (CNN) significantly outperforms all traditional machine learning models. CNNs, typically used for image data, are also effective in learning patterns from structured data like text or tabular data, offering much better accuracy and balance in classification.

□ **Extension Hybrid CNN**:

- Accuracy: 93.98%
- Precision: 94.06%
- Recall: 93.88%
- F1-score: 93.94%

The Hybrid CNN model is the most powerful in this comparison, with the highest performance across all metrics. This model likely combines different architectures or data preprocessing methods, further improving its ability to classify data accurately.

The figure 7 shows that the machine Learning model Predicted output tweet is from the Normal.

5. CONCLUSION

The increasing prevalence of deepfake content on social media poses a serious threat to information integrity, especially in sensitive areas such as politics and entertainment. This research focuses on addressing the challenge of detecting AI-generated content, particularly machine- generated tweets, by leveraging deep learning and FastText embeddings. Through this approach, it is possible to efficiently and accurately detect deepfakes, offering a significant advantage over traditional methods like human moderation and manual filtering, which are often slow, prone to errors, and unable to scale with the vast amount of online content.

The use of FastText embeddings plays a crucial role in converting tweets into meaningful word vectors, which can then be processed by deep learning models to classify tweets as either human-generated or AI-generated. This method allows for the real-time detection of deepfakes, ensuring quicker identification and mitigation of misleading content before it spreads widely. By integrating deep learning with FastText, the model achieves higher accuracy and scalability, outperforming rule-based systems that depend on predefined keywords and manual input.

In conclusion, the proposed deep learning-based framework offers a more reliable, automated solution for identifying deepfake content on social media platforms. It promises to play a crucial role in combating misinformation and ensuring the integrity of online discourse in an era where digital manipulation of content is becoming increasingly sophisticated.

While this research demonstrates the potential of deep learning and FastText embeddings in detecting machine-generated tweets, there are several areas for future enhancement and exploration. One key direction for future work is improving the model's ability to handle various types of deepfake content, such as videos, images, and audio. Expanding the framework to analyze multimodal content help provide a more comprehensive solution for detecting deepfakes across different media formats.

Additionally, as AI models continue to evolve, so too will the techniques used to generate deepfakes. Future research can focus on making the model more adaptive to emerging AI-driven content generation methods. For example, incorporating adversarial training improve the model's resilience against new deepfake generation techniques that bypass current detection methods. Moreover, collaboration with social media platforms and other stakeholders such as fact-checking organizations lead to real-time implementation of deepfake detection systems. This would allow for the automated flagging and removal of harmful content, reducing the time it takes to address misinformation. Another potential area for improvement is enhancing the dataset used for training, ensuring it includes a diverse range of content to improve the model's accuracy in real-world scenarios.

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 ofFake 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 bySocial 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 UniversityPress, 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.