International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

Machine-Generated Tweet Detection Using Deep Learning Techniques and FastText Representations

K. Pavani[1], K. Baby Ramya[2], Y. Meghana[3]
#1 Assistant Professor & Head of Department of MCA, SRK Institute of Technology, Vijayawada.

#2 Assistant Professor in the Department of MCA, SRK Institute of Technology, Vijayawada.

#3 Student in the Department of MCA, SRK Institute of Technology, Vijayawada.

Abstract: A new way to influence public opinion on social media has emerged thanks to recent advancements in natural language creation. The generative capabilities of deep neural models have been greatly enhanced by advancements in language modeling, giving them better skills for content production. Consequently, text-generative models have become quite powerful, which adversaries can use to their advantage to build social bots, which in turn makes it easier to create real deepfake posts and influence public discourse. As a solution, we must develop trustworthy algorithms to identify deepfake messages on social media. The detection of machine-generated content on social media platforms like Twitter is the current area of research interest. Using the open-source Tweepfake dataset, this research employs a simple deep learning model with word embeddings to distinguish between bot-generated and human-generated tweets. Using FastText word embeddings, a typical Convolutional Neural Network (CNN) architecture is built to detect deepfake tweets. In order to prove that the suggested method performed better than the baseline, this research used a large number of machine learning models. Term Frequency, FastText, FastText subword embeddings, Term Frequency-Inverse Document Frequency, and other properties were utilized by these core approaches. To further prove its effectiveness and highlight its benefits in solving the current challenge, the proposed method is compared to other deep learning models, such as CNN-LSTM and Long Short-Term Memory (LSTM). Results from experiments show that the CNN architecture configuration with FastText embeddings is suitable for effective and efficient classification of Twitter data, with an astounding 93% accuracy..

Index terms - *Deep learning, LSTM model, stock price prediction, sentiment analysis, sentiment dictionary, sparrow search algorithm.*

1. INTRODUCTION

1. INTRODUCTION

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

The purpose of creating social media platforms was to facilitate two-way communication and the sharing of ideas and opinions through the mediums of text, images, audio, and video. A social media bot is an automated computer that uses techniques like deepfake technology, video editing, search-and-replace, and gap-filling text to run a fake account. The content that the bot likes, shares, and posts may be real or phony. Feature representation can be learned from input data by means of deep learning, a subset of machine learning. Combining the terms "deep learning" with "fake," the term "deepfake" describes misleading multimedia created by artificial intelligence. By tricking viewers into thinking they were created by real people, deepfake multimedia has already caused problems in several fields, including politics, thanks to its proliferation on social media.

With the help of social media, false information can spread quickly, influencing public opinion and, in a democratic country, encouraging cynicism.Cyborg accounts and sockpuppets, which have varying degrees of human traits, are used to achieve this goal [6]. On the other hand, social bots—totally automated social media accounts—try to pass themselves off as human.New natural language generative models, such GPT and Grover, along with the widespread use of bots, give attackers a way to spread false information more convincingly. Case in point: the Net Neutrality lawsuit of 2017. In the end, the Commission decided to revoke due in large part to the millions of duplicate comments [10]. It is important to recognize that simple text manipulation techniques might lead to false assumptions and that more complex transformer-based models may have unintended consequences. There have been recent examples of using GPT-2 [11] and GPT-3 [12] to generate tweets for the purpose of testing their generating capabilities and to automatically generate blog posts. Under the handle "/u/thegentlemetre," a GPT-3 bot interacted with Reddit users by responding to questions posed on the /r/AskReddit subreddit [13]. Even if the bot mostly spoke harmless things. Despite the fact that no harm has come to pass thus far, OpenAI should be worried about the possible abuse of GPT-3 that could result from this incident. It is critical to set up an autonomous detection system for deepfake text and other forms of machine-generated text in order to protect genuine information and social media democracy..

2. LITERATURE SURVEY

2.1 'Big data analytics: Challenges and applica tions for text, audio, video, and social media data.

(PDF) Big Data Analytics: Challenges And Applications For Text, Audio, Video, And Social Media Data (researchgate.net)

ABSTRACT: A recommendation system facilitates the comprehension of an individual's preferences and autonomously identifies fresh, appealing content based on the correlation between their likes and ratings of other objects. This paper presents a recommendation system for the extensive data available online, including ratings, reviews, opinions, complaints, notes, feedback, and comments regarding various items such as products, events, individuals, and services, utilising the Hadoop Framework. We have utilised Mahout Interfaces to analyse data from a movie review and rating platform.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

2.2 The emergence of deepfake technology: A review

(PDF) The Emergence of Deepfake Technology: A Review (researchgate.net)

ABSTRACT: Emerging digital technologies complicate the differentiation between authentic and counterfeit media. A new development exacerbating the issue is the appearance of deepfakes, which are hyper-realistic videos utilising artificial intelligence (AI) to portray individuals saying and doing things that never occurred. The extensive reach and rapid dissemination of social media enable convincing deepfakes to swiftly influence millions, resulting in detrimental effects on society. This study analyses 84 publicly accessible online news items to investigate the nature of deepfakes, their producers, the advantages and risks associated with deepfake technology, notable examples, and strategies for countering deepfakes. The findings indicate that deepfakes pose a considerable risk to society, the political framework, and commerce; however, they can be mitigated through legislation and regulation, corporate policies and voluntary measures, education and training, alongside advancements in technology for deepfake detection, content authentication, and prevention. The paper offers an extensive analysis of deepfakes and presents cybersecurity and AI entrepreneurs with potential prospects in combating media forgeries and misinformation.

3. METHODOLOGY

i) Proposed Work:

The purpose of creating social media platforms was to facilitate two-way communication and the sharing of ideas and opinions through the mediums of text, images, audio, and video. A social media bot is an automated computer that uses techniques like deepfake technology, video editing, search-and-replace, and gap-filling text to run a fake account. The content that the bot likes, shares, and posts may be real or phony. Feature representation can be learned from input data by means of deep learning, a subset of machine learning. Combining the terms "deep learning" with "fake," the term "deepfake" describes misleading multimedia created by artificial intelligence. By tricking viewers into thinking they were created by real people, deepfake multimedia has already caused problems in several fields, including politics, thanks to its proliferation on social media.

With the help of social media, false information can spread quickly, influencing public opinion and, in a democratic country, encouraging cynicism.Cyborg accounts and sockpuppets, which have varying degrees of human traits, are used to achieve this goal [6]. On the other hand, social bots—totally automated social media accounts—try to pass themselves off as human.New natural language generative models, such GPT and Grover, along with the widespread use of bots, give attackers a way to spread false information more convincingly. Case in point: the Net Neutrality lawsuit of 2017. In the end, the Commission decided to revoke due in large part to the millions of duplicate comments [10]. It is important to recognize that simple text manipulation techniques might lead to false assumptions and that more complex transformer-based models may have unintended consequences. There have been recent

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

examples of using GPT-2 [11] and GPT-3 [12] to generate tweets for the purpose of testing their generating capabilities and to automatically generate blog posts. Under the handle "/u/thegentlemetre," a GPT-3 bot interacted with Reddit users by responding to questions posed on the /r/AskReddit subreddit [13]. Even if the bot mostly spoke harmless things. Despite the fact that no harm has come to pass thus far, OpenAI should be worried about the possible abuse of GPT-3 that could result from this incident. It is critical to set up an autonomous detection system for deepfake text and other forms of machine-generated text in order to protect genuine information and social media democracy.

The proposed method outperforms previous approaches in terms of deepfake text identification accuracy. There are obvious benefits to the method used in this work compared to complex transfer learning models such as BERT and RoBERTa. Using a simple CNN model architecture has many benefits. To start with, it does away with the need to spend a lot of time and energy training transfer learning models to a finer degree, which is typically required. This makes the proposed approach easier to implement and more practical, especially for researchers and professionals working with limited resources.

The suggested method shows that advanced performance doesn't have to be achieved with complex and time-consuming transfer learning models, making it an appealing alternative for text identification tasks. This study's findings improve deepfake recognition and provide important information for both theoretical and practical applications in the field.

ii) Conceptual Framework:

Rapid disinformation propagated through social media may sway public opinion and, in a democratic nation in particular, plant the seeds of mistrust.We accomplish this by utilizing cyborg accounts and sockpuppets, which exhibit different levels of human-like characteristics [6]. Contrarily, social bots—completely automated identities on social media platforms—strive to appear human.As bots become more commonplace and new natural language generative models emerge, such as GPT [8] and Grover [9], attackers have a new tool to distribute misinformation with increased credibility. In the 2017 Net Neutrality case, for instance, millions of duplicate comments significantly influenced the Commission's decision to abolish. Keep in mind that more complicated models reliant on transformers may cause unforeseen effects, and that basic text manipulation techniques could cause erroneous assumptions. Recently, new applications of GPT-2 [11] and GPT-3 [12] have emerged, such as automating the production of blog posts and producing tweets to evaluate its generative powers. Under the handle "/u/thegentlemetre," a GPT-3 bot engaged in conversation with Reddit users, providing considerate responses to inquiries asked on the /r/AskReddit forum [13]. Regardless, the vast majority of bot comments were completely innocuous. Even if nothing bad has happened thus far, OpenAI still needs to be concerned about how this episode could lead to the abuse of GPT-3. In order to protect actual information and democracy on

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

social media, it is essential to set up an autonomous detection system for deepfake text, which is machine-generated text.

The proposed method outperforms previous approaches in terms of deepfake text identification accuracy. There are obvious benefits to the method used in this work compared to complex transfer learning models such as BERT and RoBERTa. Using a simple CNN model architecture has many benefits. To start with, it does away with the need to spend a lot of time and energy training transfer learning models to a finer degree, which is typically required. This makes the proposed approach easier to implement and more practical, especially for researchers and professionals working with limited resources.

The suggested method shows that advanced performance doesn't have to be achieved with complex and time-consuming transfer learning models, making it an appealing alternative for text identification tasks. This study's findings improve deepfake recognition and provide important information for both theoretical and practical applications in the field.

The proposed method outperforms previous approaches in terms of deepfake text identification accuracy. There are obvious benefits to the method used in this work compared to complex transfer learning models such as BERT and RoBERTa. Using a simple CNN model architecture has many benefits. To start with, it does away with the need to spend a lot of time and energy training transfer learning models to a finer degree, which is typically required. This makes the proposed approach easier to implement and more practical, especially for researchers and professionals working with limited resources.

The suggested method shows that advanced performance doesn't have to be achieved with complex and time-consuming transfer learning models, making it an appealing alternative for text identification tasks. This study's findings improve deepfake recognition and provide important information for both theoretical and practical applications in the field..

ii) System Architecture:

The quick spread of false information through social media platforms can influence public opinion and, in particular, sow seeds of distrust in a democratic country.To achieve this goal, we use cyborg accounts and sockpuppets, which display varying degrees of human-like traits [6]. In contrast, social bots—totally automated social media accounts—try to pass themselves off as human.New natural language generative models, like GPT [8] and Grover [9], along with the widespread use of bots, give attackers a way to spread false information more convincingly. An example of this is the 2017 Net Neutrality case, when the Commission's decision to abolish was heavily impacted by millions of duplicate comments. It is important to recognise that simple text manipulation techniques might lead to false assumptions and that more complex models based on transformers may have unintended consequences. New uses of GPT-2 [11] and GPT-3 [12] for generating tweets to test its generative powers and automating blog post production have surfaced recently. A GPT-3 bot interacted with Reddit users

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

under the handle "/u/thegentlemetre," responding to questions posed on the /r/AskReddit subreddit with thoughtful answers [13]. Despite the fact that most bot remarks were harmless. Despite the fact that no harm has come to pass thus far, OpenAI should be worried about the possible abuse of GPT-3 that could result from this incident. The establishment of an autonomous detection system for machine-generated texts, also known as deepfake text, is crucial for the safeguarding of real information and democracy on social media.

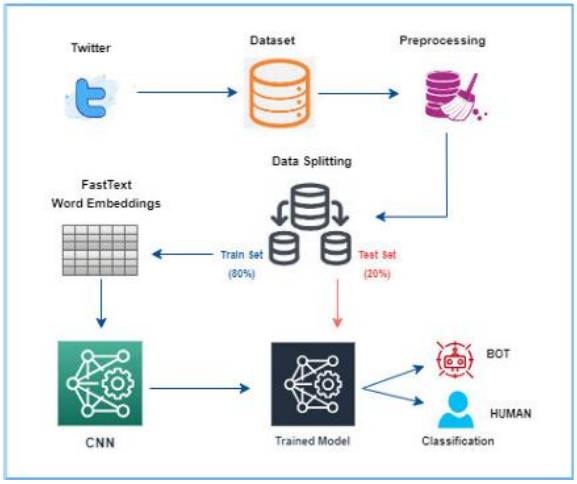

Fig 1 Proposed architecture

iii) Dataset collection:

FAKE TWEETS DATASET

This study employs the TweepFake dataset, which has a total of 25,572 tweets. The dataset consists of tweets from 17 human accounts and 23 bot accounts. Every instance of human and bot is accurately designated. The latter specifies the text generation mechanism employed, which may be human (17 accounts, 12,786 tweets), GPT-2 (11 accounts, 3,861 tweets), RNN (7 accounts, 4,181 tweets), or Others (5 accounts, 4,876 tweets).

So, these are the top 5 rows of the dataset

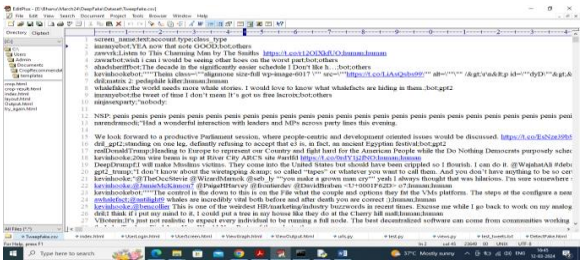

Fig 2 tweets dataset

iv) Data Processing:

Datasets contain superfluous information in unstructured or semi-structured formats. This superfluous data prolongs the model's training duration and may impair its performance. Pre-processing is essential for enhancing the effectiveness of machine learning models and optimising computational resources. Text preparation enhances the model's capacity to predict outcomes with precision. Pre-processing encompasses the following steps: tokenisation, case normalisation, stopword elimination, and numerical removal. Owing to the case sensitivity of machine learning models, the model will regard the keywords “MACHINE” and “machine” as distinct words. Consequently, the dataset must initially be transformed to lowercase during preprocessing.

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

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

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

Precision: A measure of accuracy, precision looks at the percentage of positive cases that were correctly identified. Therefore, the formula for calculating accuracy is expressed as:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Recall: Also referred to as sensitivity or true positive rate, this metric evaluates the model's capacity to accurately detect positive occurrences among all actual positive instances. It is particularly crucial in circumstances when the identification of positive cases is vital, such as in the diagnosis of illnesses.

$$Recall = \frac{TP}{TP + FN}$$

Accuracy: One way to measure how well a model performs in a classification task is by looking at its accuracy, which is the percentage of right predictions.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

F1 Score: The F1 Score is the harmonic mean of precision and recall, offering a balanced statistic that considers both false positives and false negatives, making it suitable for imbalanced datasets.

$$\text{F1 Score} = 2 * \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} * 100$$

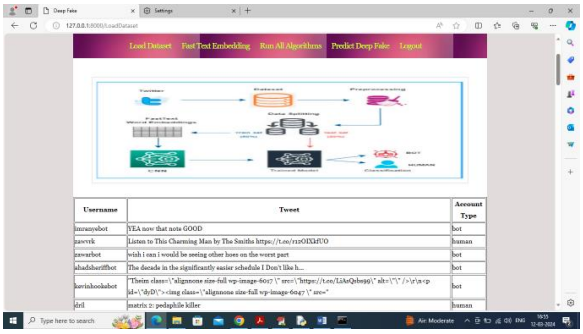

Fig 3 User input

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

Fig 4 Run all algorithms

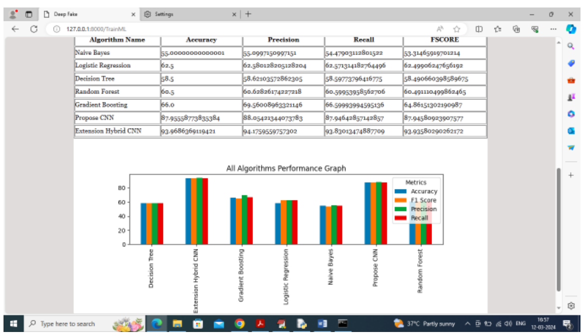

Fig 5 Graphs

5. CONCLUSION

The identification of deepfake text has become a critical and challenging task in the era of misinformation and manipulated digital content. This study addressed this issue by proposing an effective framework for deepfake text detection and evaluating its performance using a dataset containing tweets generated by both bots and human users. Several machine learning and deep learning models, along with feature engineering techniques, were employed for comprehensive analysis. Feature extraction methods such as Term Frequency (TF) and Term Frequency–Inverse Document Frequency (TF-IDF) were utilized, while word embedding techniques included FastText and FastText subword representations. The proposed approach, which integrates Convolutional Neural Networks (CNN) with FastText embeddings, achieved a significant accuracy of 93% in accurately identifying deepfake text content. Furthermore, the performance of the proposed model was compared with state-of-the-art transfer learning approaches presented in previous studies. The findings demonstrate that the CNN-based architecture provides advantages in terms of simplicity, computational efficiency, and effective handling of out-of-vocabulary words. The results contribute significantly to the field of deepfake detection and provide valuable insights for future research and real-world applications. As social media platforms continue to exert substantial influence on public opinion and information dissemination, the development of reliable and robust deepfake detection techniques remains essential for preserving information authenticity and safeguarding democratic processes.

Future Scope

Future research can focus on integrating advanced technologies such as Quantum Machine Learning (QML) and Quantum Natural Language Processing (QNLP) to further enhance deepfake text detection capabilities. Recent studies have highlighted both the challenges and opportunities associated with quantum-based learning approaches in text analysis and misinformation detection. Additionally, the incorporation of advanced neural architectures, hybrid deep learning models, real-time monitoring systems, and explainable artificial intelligence techniques can improve detection accuracy and system transparency. These advancements may

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

contribute to developing more intelligent, scalable, and efficient detection systems capable of combating the increasing spread of misinformation and fraudulent content across social media platforms..

REFERENCES

[1] J. P. Verma and S. Agrawal, ‘‘Big data analytics: Challenges and applica tions for text, audio, video, and social media data,’’ Int. J. Soft Comput., Artif. Intell. Appl., vol. 5, no. 1, pp. 41–51, Feb. 2016.

[2] H. Siddiqui, E. Healy, and A. Olmsted, ‘‘Bot or not,’’ in Proc. 12th Int. Conf. Internet Technol. Secured Trans. (ICITST), Dec. 2017, pp. 462–463.

[3] M. Westerlund, ‘‘The emergence of deepfake technology: A review,’’ Technol. Innov. Manage. Rev., vol. 9, no. 11, pp. 39–52, Jan. 2019.

[4] J. Ternovski, J. Kalla, and P. M. Aronow, ‘‘Deepfake warnings for political videos increase disbelief but do not improve discernment: Evidence from two experiments,’’ Ph.D. dissertation, Dept. Political Sci., Yale Univ., 2021.

[5] S. Vosoughi, D. Roy, and S. Aral, ‘‘The spread of true and false news online,’’ Science, vol. 359, no. 6380, pp. 1146–1151, Mar. 2018.

[6] S. Bradshaw, H. Bailey, and P. N. Howard, ‘‘Industrialized disinformation: 2020 global inventory of organized social media manipulation,’’ Comput. Propaganda Project Oxford Internet Inst., Univ. Oxford, Oxford, U.K., Tech. Rep., 2021.

[7]C.Grimme,M.Preuss,L.Adam,andH.Trautmann,‘‘S ocialbots:Human like by means of human control?’’ Big Data, vol. 5, no. 4, pp. 279–293, Dec. 2017.

[8] X. Liu, Y. Zheng, Z. Du, M. Ding, Y. Qian, Z. Yang, and J. Tang, ‘‘GPT understands, too,’’ 2021, arXiv:2103.10385.

[9] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi, ‘‘Defending against neural fake news,’’ in Proc. 33rd Int. Conf. Neural Inf. Process. Syst. (NIPS), Dec. 2019, pp. 9054–9065, Art. no. 812.

[10] L. Beckman, ‘‘The inconsistent application of internet regulations and suggestions for the future,’’ Nova Law Rev., vol. 46, no. 2, p. 277, 2021, Art. no. 2.

Author Profiles

Ms. K. Pavani Working as Assistant & Head of Department of MCA, in SRK Institute of technology in Vijayawada. She has done MCA, M. Tech in Computer Science. Her area of interest includes Machine Learning with Python and DBMS.

Ms. K. Baby Ramya is working as an Assistant Professor in the Department of MCA at SRK Institute of Technology, Enikepadu, Vijayawada, NTR District. She completed her MCA from Krishna University. She has around 3 years of teaching experience at SRK Institute of Technology. Her areas of interest include Machine Learning, Data Science, and Computer Applications.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

Ms. Y. Meghana is an MCA Student in the Department of Computer Applications at SRK Institute of Technology, Enikepadu, Vijayawada, NTR District. She has Completed Degree in B.Sc. (Computer Science) from Maris Stella College,Vijayawada. Her area of interests are DBMS and Machine Learning with Python.