The Bioscan 19(2): S.I (1), 230-233, 2024 www.thebioscan.com
AN INTERNATIONAL QUARTERLY JOURNAL OF LIFE SCIENCES

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

P GOWTHAMI DEVI [1], K BANGARU LAKSHMI [2], G MADHURI [3]

1 Assistant Professor, Kakaraparti Bhavanarayana (KBN) College, gowthamidevi@kbncollege.ac.in

2 Assistant professor, Kakaraparti Bhavanarayana (KBN) College, lakshmi089@kbncollege.ac.in

3 Assistant professor, Kakaraparti Bhavanarayana (KBN) College, madhurigongati@kbncollege.ac.in

DOI: https://doi.org/10.63001/tbs.2024.v19.i02.S.I(1).pp230-233

KEYWORDS
Sparrow search algorithm,
DL,
stock price prediction,
LSTM model,
sentiment analysis,
sentiment dictionary

Received on:

04-08-2024

Accepted on:

22-11-2024

ABSTRACT

Recent improvements in natural language generation offer a new means to influence public opinion on social media. Moreover, developments in language modelling have considerably augmented the generative capacities of deep neural models, equipping them with improved competencies for content creation. As a result, text-generative models have gained significant potency, enabling adversaries to leverage these capabilities to enhance social bots, facilitating the creation of authentic deepfake postings and swaying public conversation. In order to address this issue, trustworthy and efficient techniques for detecting social media deepfakes are essential. Automated text on Twitter has been identified in recent studies. To determine if tweets are generated by humans or bots, this study employs the publicly available Tweepfake dataset and a basic deep learning model that makes use of word embeddings. Using a CNN architecture, FastText word embeddings can detect deepfake tweets. This study utilised many machine learning models as baseline approaches to demonstrate the higher performance of the proposed approach. Some of the fundamental techniques used here were Term Frequency, Fast Text subword embeddings, FastText, and Term Frequency-Inverse Document Frequency. Additional comparisons with other DL models, such as CNN-LSTM and LSTM, demonstrate the effectiveness and benefits of the suggested approach in precisely resolving the issue. The experimental results demonstrate that Twitter data may be efficiently and accurately classified with a 93% accuracy rate using CNN architecture and FastText embeddings.

INTRODUCTION

Many other types of media, including text, images, audio, and video, can be shared on social media. A computer program likes, shares, and posts content on behalf of a fictitious social media account using deepfake technologies, search-and-replace, video editing, and gap-filling technologies. Using machine learning as an input, deep learning constructs feature representations. A combination of the words "deep learning" and "fake," the term "deepfake" describes misleading media created by artificial intelligence. The development and dissemination of deepfake multimedia on social media have already posed challenges in various domains, including politics, by misleading users into believing that they were produced by humans.

Social media facilitates the rapid dissemination of misinformation to manipulate public perceptions and beliefs, particularly to foster distrust in a democratic nation. For this, we utilise sockpuppet and cyborg accounts that exhibit varying degrees of human characteristics [6]. But social bots, which are entirely automated profiles on social media, act human.New breakthroughs in natural language generative models, such as Grover and GPT, plus the extensive use of bots make it easier for foes to propagate disinformation. The 2017 Net Neutrality case exemplifies this issue. millions of duplicates Comments played a crucial part in the Commission's decision to rescind [10]. The consequences of transformer-based models and incorrect beliefs might result from simple text manipulation. Some recent applications of GPT-2 [11] and GPT-3 [12] involve automating blog posts and generating tweets to demonstrate its generative capabilities. Answers to enquiries posted on /r/AskReddit were provided by the GPT-3-based bot "/u/thegentlemetre" [13]. The majority of the bot's remarks posed little threat. There may have been some misuse of GPT-3, which OpenAI should be worried about, even though no harm has been done. In order to safeguard social media democracy and genuine information, an automated system is required to detect deepfake text.

1. LITERATURE SURVEY

2.1 'Industrialized disinformation: 2020 global inventory of organized social media manipulation:

DemTech | Industrialized Disinformation: 2020 Global Inventory of Organized Social Media Manipulation (ox.ac.uk)

ABSTRACT: A danger to democracy arises whenever public opinion is swayed by social media. For the past 4 years, we have closely observed the social media campaigns launched by political parties and governments throughout the globe in an effort to sway public opinion, as well as their partnerships with commercial companies. We plan to conduct study in 2020 that looks at computational propaganda in 81 countries and how it's changing in terms of resources, tactics, tools, and capacities to influence public opinion.

2.2 Socialbots:Human like by means of human control?:

[1706.07624] Social Bots: Human-Like by Means of Human Control? (arxiv.org)

ABSTRACT: Public opinion is impacted by anonymous social bots. They may have used online and social media platforms to spread propaganda that affected the results of the last election. "Social Bot" is defined differently in several scientific domains, which is quite intriguing. The essay begins with a reasonable definition before moving on to explain Twitter social bots and the

limitations of their technology. Even if Big Data and Deep Learning have come a long way, bots still can't do everything. We go down the key points of effective human interactions and the steps to take to enhance and control bot capabilities.

2.3 Deepfake warnings for political videos increase disbelief but do not improve discernment: Evidence from two experiments

Deepfake Warnings for Political Videos Increase Disbelief but Do Not Improve Discernment: Evidence from Two Experiments (researchgate.net)

ABSTRACT: One product of recent machine learning advancements is the "deepfake," an impressively realistic CGI of a famous person making fraudulent statements. Despite policymakers' concerns that deepfakes could sway elections, studies have shown little impact. Voters may become suspicious of all political video content if they are continually warned about deepfakes, according to this essay's analysis of a downstream effect of these fake news stories. Respondents were unable to distinguish between genuine and fake films in our two online surveys. The ability of participants to detect manipulated footage was not enhanced by warnings concerning deepfakes. But the participants kept thinking the videos weren't real because of these cautions. Anyone watching any relevant video footage had their suspicions heightened by deepfakes. Deepfakes can be used by politicians and campaigns to reject and devalue real content, regardless of how convincing it is, according to our research.

METHODOLOGY

i) Proposed Work:

The suggested approach significantly enhances the accuracy of deepfake text identification compared to prior methods. This study's technique presents clear advantages over intricate transfer learning models like RoBERTa and BERT. The employment of a basic CNN model architecture offers numerous advantages. Firstly, it eliminates the necessity for considerable training duration and computational resources usually necessary for fine-tuning transfer learning models. This renders the suggested methodology more accessible and efficient, particularly for academics and practitioners with constrained resources.

The proposed text identification method demonstrates that high-level performance is feasible even without labour-intensive transfer learning methods. This research contributes to the field of deepfake identification and offers useful data for studies and applications in the future.

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 under the handle "/u/thegentlemetre," responding to questions posed on the /r/AskReddit subreddit with thoughtful answers [13]. Regardless, the vast majority of bot comments were harmless. Although no one has been injured, OpenAI should be concerned about the potential abuse of GPT-3 resulting from this incident. Automated deepfake text detection is necessary to safeguard real information and social media democracy.

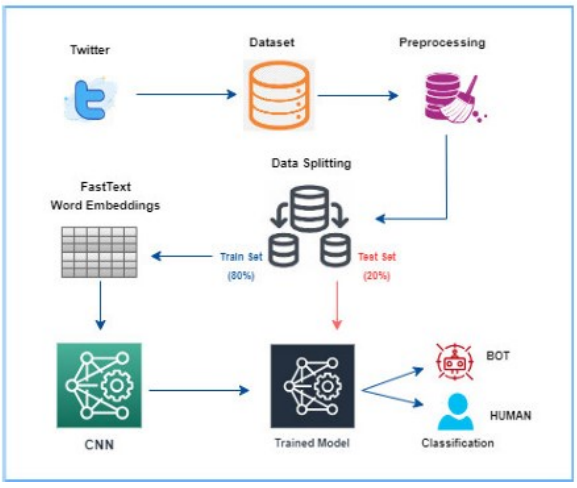

Fig 1 Proposed architecture

iii) Dataset collection:

FAKE TWEETS DATASET

With the help of TweepFake, 25,572 tweets were used in this investigation. There are 17 tweets from real people and 23 from bots in the set. Identified are all machines and humans. Humans (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) can be the text creation process.

So, these are the top 5 rows of the dataset

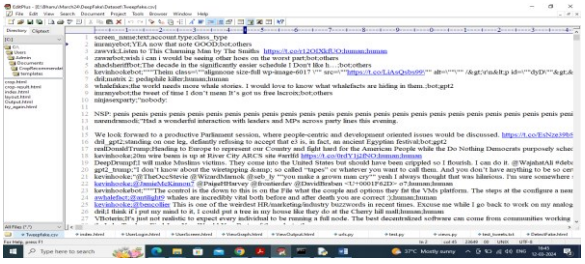

Fig 2 tweets dataset

iv) Data Processing:

Unstructured or semi-structured datasets include extraneous information. Training the model takes longer with this extraneous data, which could lead to worse results. It is necessary to pre-process data in order to optimise computational resources and the efficacy of machine learning models. In order for the model to make good predictions, text preparation is essential. Tokenisation, case normalisation, stopword removal, and numeral removal are all part of the pre-processing. Because of case sensitivity, ML models will recognise "MACHINE" and "machine" as separate words. Lowercase data must be preprocessed.

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 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.

2. EXPERIMENTAL RESULTS

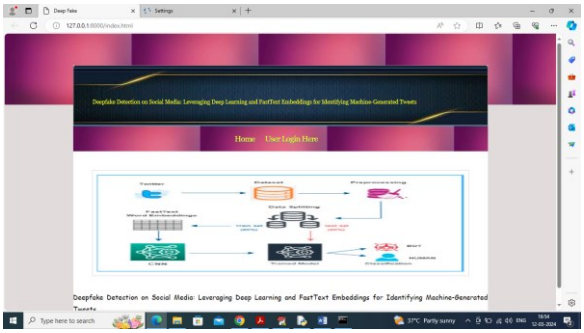

Fig 3 Home page

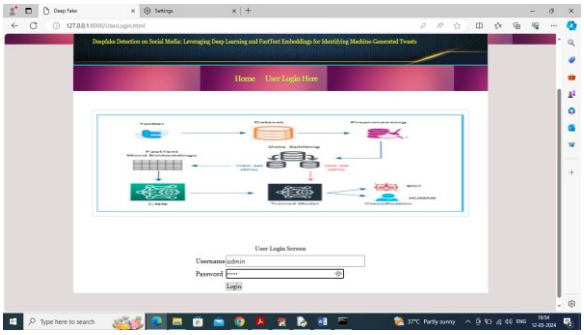

Fig 4 Signin page

Fig 5 Load Dataset link page

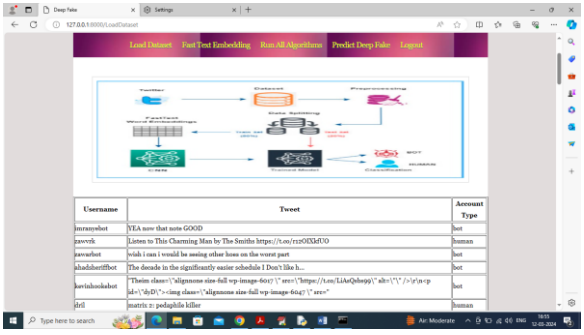

Fig 6 User input

Fig 7 Run all algorithms

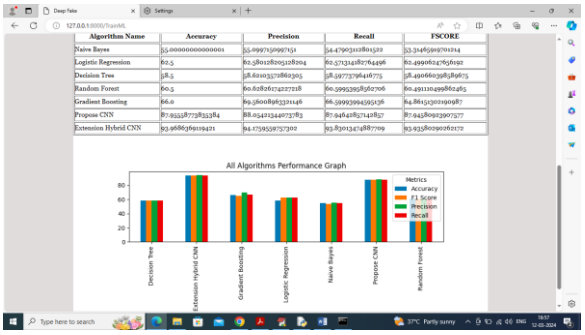

Fig 8 Graphs

CONCLUSION

With the prevalence of disinformation and altered content, the detection of deepfake texts has become an important yet challenging task. A deepfake text detection technique was introduced and evaluated in this research to tackle this problem. We use ML, DL, and feature engineering to sift through a dataset that includes both bot and human tweets. Some well-known approaches for word embedding include FastText and FastText subwords, whereas others, such Tf and TF-IDF, are used for feature extraction. The proposed strategy, utilising a combination of CNN and FastText approaches, achieved a commendable accuracy score of 0.93 in the precise identification of deepfake text. Additionally, the outcomes of the proposed methodology are juxtaposed with other cutting-edge transfer learning models from prior work. This study demonstrates that the implementation of a CNN model architecture excels in simplicity, computational efficiency, and the management of out-of-vocabulary phrases. The results of this study enhance the domain of deepfake identification and offer significant insights for future research and practical implementations. To this day, social media has an impact Real information and democratic processes can only be safeguarded by enhancing deepfake text detection. Both the challenges and the opportunities of quantum machine learning are discussed in [76], while the quantum method for detecting deepfake text is laid out in [77]. Improved techniques for detecting social media fraud and deception will be developed using quantum natural language processing and other cutting-edge methodologies.

3. FUTURE SCOPE

We need better deepfake text detection technology to safeguard democratic processes and authentic information from the enormous impact of social media on public opinion. While [76] focusses on the challenges and opportunities of quantum machine learning, [77] describes the quantum method for detecting deepfake text. Improved techniques for detecting social media fraud and deception will be developed using quantum natural language processing and other cutting-edge methodologies.

REFERENCES

- 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.
- 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.
- M. Westerlund, ''The emergence of deepfake technology: A review,'' Technol. Innov. Manage. Rev., vol. 9, no. 11, pp. 39–52, Jan. 2019.
- 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.
- 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.
- 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.
- C.Grimme,M.Preuss,L.Adam,andH.Trautmann,''Socialb ots:Human like by means of human control?'' Big Data, vol. 5, no. 4, pp. 279–293, Dec. 2017.
- X. Liu, Y. Zheng, Z. Du, M. Ding, Y. Qian, Z. Yang, and J. Tang, ''GPT understands, too,'' 2021, arXiv:2103.10385.
- 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.
- 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.
- J.-S. Lee and J. Hsiang, ''Patent claim generation by fine-tuning OpenAI GPT-2,'' World Pat. Inf., vol. 62, Sep. 2020, Art. no. 101983.
- R. Dale, ''GPT-3: What's it good for?'' Natural Lang. Eng., vol. 27, no. 1, pp. 113–118, 2021.
- W. D. Heaven, ''A GPT-3 bot posted comments on Reddit for a week and no one noticed,'' MIT Technol. Rev., Cambridge, MA, USA, Tech. Rep., Nov. 2020, p. 2020, vol. 24. [Online]. Available: www.technologyreview.com