Enhancing Online Social Media Fake News Recognition Using Deep Learning Models

[1]Ms.Renuka B N [2]Rakshitha B V

[1]Assistant Professor, Department of MCA, BIET, Davanagere
[2] Student,4th Semester MCA, Department of MCA, BIET, Davanagere

ABSTRACT

New tools that can be used to sway public opinion on community media stages take lived made possible by recent developments in accepted linguistic generation. To improve the identification of false information on social media, a standard Convolutional Neural Network (CNN) architecture is suggested. Term Frequency (TF), term frequency-inverse document frequency (TF-IDF), Fast Text embeddings, and Fake News Acceptance features are some of the standard mechanism education models that are compared to the suggested method in order to show how effective it is. Furthermore, the suggested CNN model's performance is assessed in comparison to alternative deep learning methodologies, including hybrid CNN-LSTM architectures and Long Short-Term Memory (LSTM) networks. The grades reveal that the CNN-based method, mainly when combined beside Fake News Recognition features, achieves superior accuracy—reaching up to 93%—in classifying tweet data. This highlights the model's efficiency and effectiveness in detecting fake social media content.

Keywords: LSTM, CNN, TF-IDF

I. INTRODUCTION

Significant advances in text production consume been completed conceivable by the quick development of natural language processing (NLP) technologies, which allow machines to create content that is additional realistic and fluent than that of humans. While these innovations have numerous beneficial applications, they also pose serious challenges—particularly in the realm of online misinformation. Adversaries have turned societal broadcasting locations similar Twitter interested in their main targets, using generative models to produce convincing fake content and use social bots to sway public opinion. Because these bots are capable of producing false messages on their own, it becomes more challenging to discern between real and fake content. Strong detection methods are essential as the danger of AI-generated disinformation increases. Maintaining the integrity of online dissertation in social media environments requires the ability to recognize machine-generated text. This study investigates deep learning-based methods for identifying phony or bot-generated tweets in order to address this urgent problem. The study examines how well a Convolutional Neural Network (CNN) model enhanced with different textual features can categorize tweets equally moreover human- before bot-created using the publicly available twee fake dataset.

II. RELATED WORK

S. Agrawal and J. P. Verma, Fake news detection using deep learning initially gained prominence in the field of computer vision, where it focused on detecting manipulated visual content such as full-face synthesis, identity swaps, facial attribute changes, emotion alterations, and body reenactment.

Over time, these techniques extended to audio manipulation, enabling the generation of speech mimicking real voices after only a few seconds of exposure. Text synthesis soon followed, supported by significant breakthroughs in language modelling [1].

H. Siddiqui, E. Healy, and A. Olmsted, A pivotal advancement occurred in 2017 with the introduction of the self-attention mechanism and the Transformer architecture, which revolutionized language modelling. These models estimate the probability of word sequences through statistical and probabilistic techniques. Transformer-based models—such as GPT, BERT, GPT-2, and others—not only enhanced natural language generation but also improved various natural language understanding tasks. In 2019, Radford et al. introduced GPT-2, a powerful pre-trained language model capable of generating coherent, human-like paragraphs from minimal input [2].

M. Westerlund, That same year, GROVER was developed to generate structured documents such as news articles, while CTRL allowed for controlled text generation based on specific styles and tasks. Later, researchers introduced OPTIMUS, which integrated variational autoencoders into the text generation process. The GPT-2 team also conducted internal studies to explore detection strategies. These included a traditional machine learning classifier using logistic regression on TF-IDF features and a simple zero-shot method that flagged a text as machine-generated if its likelihood score aligned more closely with known machine generated outputs than with human-written ones [3].

J. Tarnowski, J. Kalla, and P. M. Aronow, To assists in detection, the Giant Language Model Test Room (GLTR) stayed introduced. GLTR visualizes statistical irregularities in word prediction patterns— differences between human and machine writing—to help users identify AI-generated text. It highlights how generated words are selected from a model's next-token distribution, which often deviates from natural human language patterns [4].

S. Vosoughi, D. Roy, and S. Aral, The only dedicated study on identifying deepfake social media messages using GPT-2 involved detecting manipulated Amazon reviews. Multiple detectors were evaluated, including those based on Grover, GLTR, RoBERTa, and an ensemble classifier using logistic regression [5].

Y. Zhang, Y. Su, L. Weigang, and H. Liu, However, existing detection techniques have notable limitations. Most focus on long-form news content rather than short social media posts, and typically rely on a single adversarial model like GPT-2 or GROVER—an unrealistic assumption in real world settings where various generative models may be employed. Current approaches to fake text detection include graph-based methods, feature-engineered models, and deep learning architectures such as BiLSTM and RoBERTa [6].

Y. Bataller and M. ousel, A system for comprehensive survey highlighted ongoing trends, challenges, and limitations in fake news detection research. Datasets like PAN and Cresci have also been used, focusing on user profile features and behavioural patterns to identify bots [7].

T. Oladipupo, However, these datasets and methods often lack relevance to short-form social media content. To address this, the TweepFake dataset was introduced. It provides a rich collection of fake tweets generated by a variety of text generation models, supporting the development of more generalized and robust detection systems tailored to the unique challenges of social media environments [8].

III. METHODOLOGY

The proposed procedure focuses on detecting manipulated visual content in the form of machine-generated transcript on social media platforms. The process consists of structured steps including dataset selection, data preprocessing, algorithm implementation, and model optimization. A Convolutional Neural Network (CNN) typical is used as the primary deep learning architecture,

coupled with effective feature extraction techniques to enhance classification accuracy. The overall workflow is designed to identify whether tweets are authored by human users or generated by bots using advanced language models.

3.1 Dataset used

The dataset used for this training is the deep Fake dataset, which is publicly available and designed to support research in fake content detection on social media platforms. It consists of a comprehensive collection of tweets generated both by humans and various AI based language models. This diversity enables robust training and evaluation of detection algorithms, as the dataset includes multiple styles and complexities of text that reflect real world social media posts. The dataset's relevance to short-form content, such as tweets, makes it particularly suitable for studying machine generated misinformation in the context of online discourse.

3.2 Data preprocessing

The textual content undergoes a number of preprocessing steps to ensure the dataset is clean and appropriate for modelling. These procedures include normalization methods like lowercasing and tokenization after the text has been cleaned up by eliminating special characters, punctuation, and extraneous whitespace. The input quality for the machine learning models is enhanced by this standardization. Following preprocessing, the dataset is divided in an 80:20 ratio into training and testing sets. This separation makes it possible to train the model while keeping some data for controlled, objective performance evaluation.

3.3 Algorithm used

The core algorithm employed in the planned system is a three-layer Convolutional Neural Network (CNN). This model is particularly well-suited for text cataloguing tasks as it can automatically extract evocative features from the raw input data. Unlike outmoded models that depend on heavily on handcrafted features, the CNN is capable of capturing both local and hierarchical patterns within short text sequences such as tweets. Additionally, its deep architecture enables it to learn complex relationships in the data, enhancing the accuracy of classification. The model architecture avoids reliance on fixed vocabularies, making it adaptable to the evolving language used in social media.

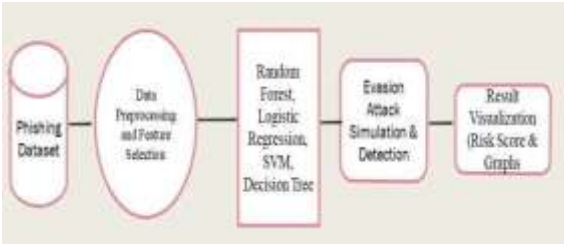

Figure 3.3.1 : System architecture

3.4 Techniques

To further enhance performance, the model incorporates Fake News Recognition features for text vectorization. These features provide semantically rich representations of the input data, improving the learning process. The study also benchmarks the proposed CNN model against several traditional and deep learning approaches. Baseline comparisons include traditional methods using Term Frequency (TF), TF-IDF, and Fast Text embeddings. Deep learning models like CNN-LSTM hybrids and Long Short-Term Memory (LSTM) models are also assessed. Accuracy, precision, recall, and F1-score are among the common metrics used to evaluate the efficacy of each model. Crucially, the CNN model demonstrates resilience to out-of-vocabulary (OOV) terms, a common issue in natural language processing, making it more robust and adaptable for real world applications.

3.5 Flowchart

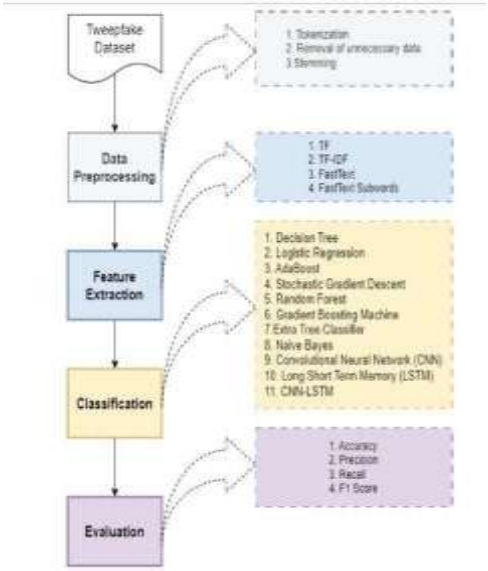

Figure 3.5.1: Flowchart

IV. RESULTS

4.1 Graphs

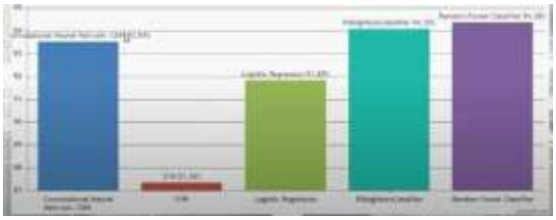

Figure 4.1.1 : Resulant graph

4.2 Screenshots

Figure 4.2.1 :Tweet Type ratio results in line chart

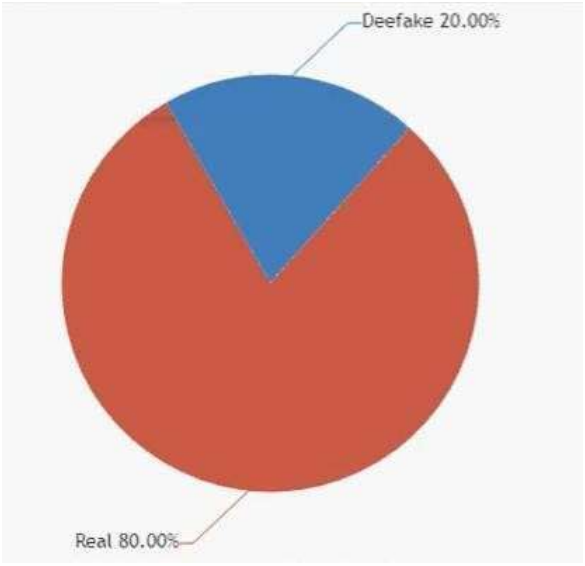

Figure 4.2.2 : Tweet Type ratio results in pie chart

4.3 Accuracy Result

The study suggests a deep learning method for identifying phony or artificially generated tweets on social media by employing a Convolutional Neural Network (CNN). The model successfully differentiates between content produced by humans and bots by utilizing the twee fake dataset and integrating

Fake News Recognition features. It demonstrated strong potential for combating false information online with a high accuracy of 93% when compared to other models such as logistic regression and LSTM.

V. CONCLUSION

In this study, a deep learning-based approach was proposed to effectively identify machine generated false content on social media platforms, particularly Twitter. By integrating on Convolutional Neuronal Network (CNN) with Fake News Recognition features, the model confirmed superior recital compared to traditional appliance knowledge in addition other subterranean knowledge techniques. The planned model achieved a high accuracy of 93%, showcasing its capability to accurately distinguish between human and bot-generated tweets. Its flexibility in handling out-ofvocabulary terms

further strengthens its applicability to real-world social media environments, where language usage is constantly evolving. The results validate the efficiency of the proposed framework and highlight its potential as a reliable tool for combating distortion and enhancing the honour of online treatise.

VI. REFERENCES

[1]. S. Agrawal and J. P. Verma, "Big data analytics: Challenges and applications for text, audio, video, and social media data," Int. J. Soft Compute, Artif. Inntel. 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. Inova. Manage. Rev., vol. 9, no. 11, pp. 39– 52, Jan. 2019.

[4]. J. Tarnowski, 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, and S. Aral, D. Roy, "The spread of true and false news online," Science, Vol. 359, no.6380, pp. 1146–1151, Mar. 2018.

[6]. Y. Zhang, Y. Su, L. Weigang, and H. Liu, "Rumour and commanding information propagation model considering super spreading in complex social networks," Physica A: Statistical Mechanics and its Applications, vol. 506, pp. 395-411, 2018.

[7]. Y. Bataller and M. Ousel, "Introduction to machine learning," Methods in Molecular Biology (Clifton N.J.), vol. 1107, pp. 105-128, 2014.

[8]. T. Oladipupo, "Types of Machine Learning Algorithms" in New Advance in Machine Learning, University of Portsmouth, United Kingdom, 2010.

*****