Enhanced Deepfake Detection on Social Media: Applying CountVectorizer and Random Forest for Optimal Tweet Classification

[1] Dr. Chinnem Rama Mohan, [2] Palicharla Joshuva Prabhakar, [3] Narra Karthik, [4] Pandi Manoj Kumar, [5] Komarala Rithwik, [6] Mamidi Masthan Sai Prabhas

[1] *Associate Professor, Department of CSE, Narayana Engineering College, Nellore, Andhra Pradesh, India*
[2 , 3 ,4 ,5 & 6] *UG Scholars, Narayana Engineering College, Nellore, Andhra Pradesh, India*

Abstract: *Social media stages were made for individuals to interface and share their conclusions and thoughts through writings, pictures, sound, and recordings. Moreover, text-generative models have ended up progressively effective permitting the foes to utilize these momentous capacities to boost social bots, permitting them to produce practical deepfake posts and impact the talk among the common open. Addressing this issue requires the development of reliable and accurate strategies for detecting deepfake messages on social media. Beneath this thought, current inquire about addresses the recognizable proof of machine-generated content on social systems like Twitter. In this ponder, a straightforward Machine learning demonstrate in combination with CountVectorizer (CV) is utilized for the classification of tweets as human-generated or bot-generated utilizing a freely accessible Tweepfake dataset. Additionally, the execution of the proposed strategy which is Random Forest Classifier is moreover compared against other machine learning models such as MLP, LR, SVM, DT, KNN, AC, SGC, GBM and NB showing the adequacy and highlighting its preferences in precisely tending to the errand at hand. Test comes about show that the plan of the Random Forest Classifier is reasonable for productive and successful classification of the tweet information with a predominant 94% accuracy.*

keywords *– Text classification, Machine Learning, RF, SVM, DT.*

I. INTRODUCTION

Social media stages have revolutionized human communication, advertising unparalleled network and the capacity to share conclusions, thoughts, and data immediately through writings, pictures, sound, and recordings. In any case, this ease of communication has moreover encouraged the quick spread of deception, especially in majority rule social orders where it can dissolve believe in teach and impact open conclusion. Different sorts of accounts, extending from human-operated to robotized bots, contribute to this control. Eminently, in the 2017 Net Nonpartisanship case, millions of copied comments, numerous produced by bots, played a critical part in the Government Communications Commission's choice to revoke net nonpartisanship directions. This occasion underscored the impact of bot-generated deception on imperative approach decisions.

Further outlining the affect of progressed innovation, in 2019, the discharge of GPT- 2 showcased improved text-generating capacities that were about undefined from human-written substance. This raised concerns almost the potential abuse of such innovation. For illustration, a bot utilizing GPT-3 was able to connected with clients on Reddit, posting comments that showed up veritable. In spite of the fact that these intuitive were for the most part safe, they highlighted the ease with which such innovation can spread wrong data. The challenge of identifying machine-generated content, or deepfake content, is especially articulated with progressed models like GPT-2 and GPT-3. These models can create content that is for all intents and purposes unclear from human-written substance, posturing a noteworthy challenge for existing discovery strategies.

This issue is compounded by the predominance of short-form substance on social media, which can be indeed more challenging to analyze.

Therefore, creating successful strategies to distinguish and relieve the spread of deepfake content is significant for keeping up believe and unwavering quality in social media communications. Tending to this challenge includes making advanced discovery frameworks able of recognizing between honest to goodness and machine-generated content, eventually securing the astuteness of data shared on social media stages.

II. MOTIVATION

Existing investigate has made critical strides in recognizing deepfake substance, but numerous considers have centered on longer writings, such as news articles. These strategies regularly drop brief when connected to the shorter, more changed substance ordinarily found on social media. The energetic nature of social media stages, characterized by quick substance turnover and a different client base, presents interesting challenges that current discovery strategies do not completely address. Particular challenges in existing deepfake location strategies have provoked this venture. Conventional location procedures battle with the brevity and inconstancy of social media writings. The brief length of tweets, for case, makes it troublesome to apply strategies planned for longer, organized substance. Furthermore, the quick advancement of generative models implies location strategies must persistently adjust to unused, advanced fake substance. Moreover, the sheer volume of information on social media requires profoundly effective and versatile discovery solutions.

The expanding utilize of social bots and modern NLP models compounds these challenges. These advances can be utilized malevolently to spread deception rapidly and convincingly. This requires the advancement of vigorous location components that can keep pace with advancing dangers. By centering on recognizing deepfake content particularly in the setting of social media, this extends points to fill a basic crevice in current investigate.

III. LITERATURE SURVEY

Table 1: Literature Survey

Ref	Year	Methods	Dataset	Findings
[1]	2018	LSTM	Cresci and collaborators dataset	Researchers utilized tweet-level bot detection, leveraging textual features and metadata, and achieved a 96% AUC score.
[4]	2019	Deep Forest Algorithm	Dataset collected from the Twitter time-line (API)	The application of SMOTE in conjunction with the proposed model surpassed traditional machine learning models in performance.
[5]	2019	Fine-tuned BERT	Create their dataset using GPT-2 model	The study explored machine-generated text using discriminators on a balanced dataset.
[3]	2020	BERT-based	English tweets from the PAN competition dataset	A bot detection model was developed, achieving an F1 score of 83.86%.
[7]	2020	Bot-AHGCN	CTU-13 dataset and their own collected botnet dataset	A multi-attributed heterogeneous graph convolution approach was implemented for bot detection.
[8]	2020	Gaussian kernel Density peak clustering algo.	dataset consisting 1971 normal human accounts and 462 social bot accounts	Classification improvements were noted through the use of oversampling techniques.

[10]	2021	ROBERTa based detector	Tweepfake dataset	Discrimination between human-written and bot-generated text was achieved, and a deepfake tweet dataset was presented, alongside the application of 13 models for detection.
[9]	2022	BILSTM	A dataset consisting 30000 tweets from PAN-20	The role of social bots in disseminating fake news was examined.
[6]	2022	Google Bert	Cresci 2017 data set	A dataset was classified into bot or human-written text with an accuracy of 94%.
[11]	2022	GANBOT framework	Twitter social bot	The proposed model outperformed previous contextual LSTM methods for bot detection.
[12]	2022	Graph based approach	9 datasets including TwiBot-22	Graph-based detection on a large scale was addressed.
[13]	2022	Feature-based approach	5 Datasets (Cresci2015, Cresci2017, Lee2011, Varol2017 and Moghaddam2019)	Friendship preference features from profiles were employed for bot detection.
[14]	2023	Logistic Regression ROBERTa	Human ChatGPT Corpus (HC3)	Human analysis, linguistic evaluation, and bot-generated text detection were conducted, providing deep insights for future research.
[15]	2023	XGBoost	Human-written essays and ChatGPT generated essays	TF-IDF and hand-crafted features were used to detect ChatGPT-generated text.
[16]	2023	Transformer-based ML Model (DistilBERT)	ChatGPT query Dataset, ChatGPT rephrase Dataset	The difficulty of detecting rephrased ChatGPT text was examined, providing insights into its distinctive writing style.
[17]	2023	Fine-tune-based detection model (ROBERTa)	Human-written abstracts and AI- generated abstracts	The gap between machine-generated text and human-written text was investigated.

IV. PROPOSED SYSTEM

The proposed framework offers a few key functionalities and improvements over existing strategies, especially in taking care of the interesting challenges postured by brief social media writings. The extend includes collecting tweets from different sources to make a comprehensive dataset for preparing and testing the demonstrate. A user-friendly web application is created to encourage preprocessing, preparing, testing, and discovery of deepfake content in tweets. In differentiate to existing frameworks, which fundamentally center on longer writings, this venture leverages a novel highlight extraction strategy particularly suited for brief writings. The Number Vectorizer method is utilized for this reason, advertising a noteworthy advancement in preparing and analyzing tweet information.

The framework employments a machine learning algorithm to improve discovery precision, accuracy: **Random Forest Classifier (RF)** These calculations are coordinates inside a Django system, utilizing a MySQL backend to oversee and store information, counting tweets, show correctness's, and client qualifications. The system's capabilities incorporate demonstrate preparing and testing, where directors can prepare and test numerous machine learning models to decide the most compelling approach for deepfake discovery. The framework permits admins to see the exactness of models, with comes about shown through different visualizations such as bar charts, pie charts, and line charts. Clients can enroll and log in to the framework, and get to highlights to foresee the genuineness of tweets. By tending to the confinements of existing frameworks, especially the trouble in recognizing deepfake substance in brief tweets, this venture points to give a comprehensive arrangement that moves forward precision and unwavering quality in real-world social media situations.

The implementation of the system has following modules:

LOADING THE DATASET:

- The dataset is loaded using Pandas, a powerful data manipulation library in Python. This step ensures that the data is in a structured format, making it easier to manipulate and analyze.

LABEL ENCODING:

- The labels are encoded into a binary format: "Real" is represented as 0 and "Fake" as 1.

The proposed system workflow shows in the below.

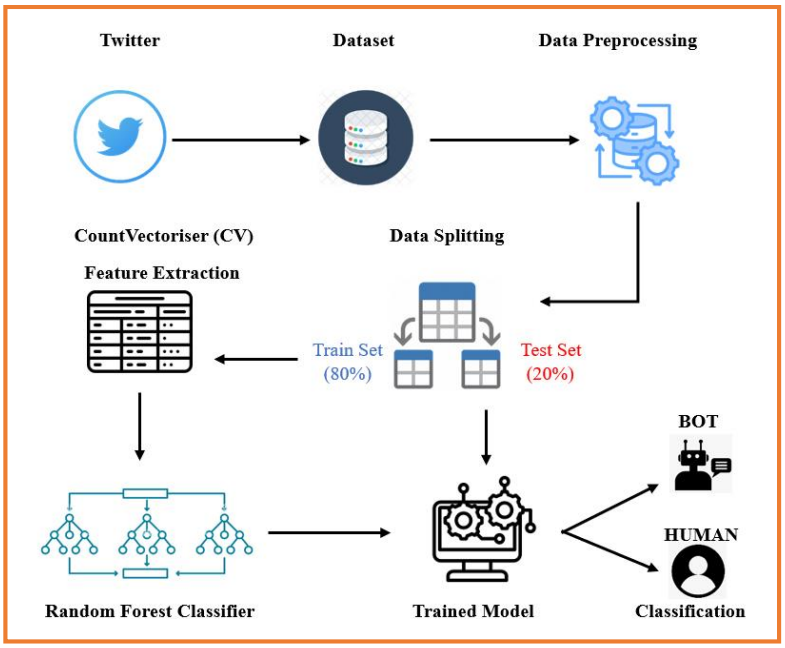

Fig. 1: Proposed System Overview

FEATURE EXTRACTION:

- The CountVectorizer (CV) from the scikit-learn library is used to convert the text data into numerical features. This method effectively transforms the tweet texts into a matrix of token counts, which serves as the input for our machine learning models.

SPLITTING THE DATA:

- The dataset is split into training and testing sets using an 80-20 split. This division allows us to train the models on one portion of the data and evaluate their performance on the other.

TRAINING MODEL:

The Random Forest Classifier builds many decision trees during the training phase. All the decision trees have been trained on a random subset of the dataset and used a random subset of features. Bootstrap aggregation or bagging is an ensemble learning technique that improves model's robustness and generalization capabilities. Data collection and preprocessing are first required in order to clean, tokenize and normalize the tweets for quality assurance purposes. At 80:20 ratio, dataset gets divided into training set and testing set which allows building models based on strong grounds and evaluating performance without biasing it using test data

Feature extraction utilizes CountVectorizer (CV) to convert text data into numerical representation suitable for classifier During training, each decision tree is trained independently on different data subsets, promoting diversity among the trees The final prediction is made by aggregating the outputs of all individual trees, typically through a majority voting mechanism This reduces overfitting and increases the model's accuracy and robustness Random forest classifier has several advantages Its accuracy is very high as evidenced by our project where it achieved 94% percentage. Of note, the ensemble nature of this mitigates outliers and noise in the data. Due to its robustness, the model performs dependably over a broad variety of datasets with inherent variability. Moreover, Random Forest provides insights into important features; hence it helps interpret and understand how decisions are made by models. Such an ability enables a more detailed examination of the most valuable contributors for predictions as such promoting transparency and confidence in the classifier overall. In fact, such is a process that Random Forest classifier goes through together with its many benefits that enable it to be a very strong tool for accurate and reliable tweets classification.

V. EVALUATION METRICS

Each model is evaluated based on its accuracy, confusion matrix, and classification report.
Accuracy: The ratio of correctly predicted instances to the total instances.

$$\text{Accuracy} = \frac{(\boldsymbol{TP + TN})}{(\boldsymbol{TP + FP + TN + FN})}$$

Precision: The ratio of correctly predicted positive observations to the total predicted positives.

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

$$= \frac{\boldsymbol{True\ Positive}}{\boldsymbol{Total\ Predicted\ Positive}}$$

Recall (Sensitivity): The ratio of correctly predicted positive observations to all the observations.

$$\text{Recall} = \frac{\boldsymbol{True\ Positive}}{\boldsymbol{True\ Posititve + False\ Negative}}$$

$$= \frac{\boldsymbol{True\ Positive}}{\boldsymbol{Total\ Actual\ Positive}}$$

F1-Score: The weighted average of Precision and Recall.

$$\text{F1} = 2 \text{ x } \frac{Precision * Recall}{Precision + Recall}$$

Confusion Matrix: A summary of prediction results on the classification problem. The matrix includes True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) values.

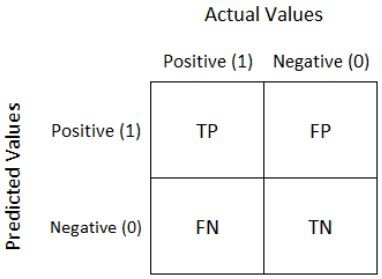

Fig 2: Confusion Matrix

VI. RESULTS

The proposed strategy incorporates a comprehensive approach to information dealing with, beginning with information collection and preprocessing. The utilize of instruments like CountVectorizer for highlight extraction and different classifiers such as Multi-Layer Perceptron (MLP), SVM, KNN, LR, DT, AC, GBM, SGC and Naïve Bayes. The usage in Django, with a MySQL backend, bolsters a user-friendly interface for both clients and admins, improving ease of use and administration capabilities. Several classifiers are trained on the training set.

COMPARISON OF CLASSIFIERS USING COUNTVECTORIZER (CV):

The comparison of classifiers in the below Table 2 using CountVectorizer (CV) highlights The Random Forest (RF) classifier outperformed all others with an accuracy of 94.1% and a recall rate of 1.00; however, some other classifiers such as MLP and Naïve Bayes also performed well with accuracies of 93.5% and 93.2%, respectively, but this was not enough to outshine RF's better metrics showing its ability to identify human tweets from bot-generated ones. However, there were other classifiers like SVM, DT, LR, AC, GBM that gave good results compared to RF in regard to precision and general effectiveness in the classification process. This confirms that the RF is unique among all these messages concerning deepfake tweet detection on social media sites demonstrating why it is the best available option in that aspect.

Tweet classification can be done through combining CountVectorizer (CV) with Random Forest (RF). CountVectorizer helps convert raw text data into structured numerical format to enable effective analysis of textual patterns. It turns text into matrix of token counts where CV captures word frequency which provides important features for the classification. For instance, the MLP and Naive Bayes achieved accuracies of 93.5% and 93.2% respectively which is a strong performance but not quite as good as that of RF. The same approach was implemented with other classifiers like MLP, SVM, and Naïve Bayes using features derived from CV but their results did not measure up to those of the RF classifier that came out on top with an accuracy rate of 94.1% and perfect recall rate of 1.00 in separating actual Tweets from fake ones.

Table. 2: Classification Result using CountVectorizer (CV)

Algorithm	Accuracy	Precision	Recall	F1-Score
MLP	93.5	0.94	0.99	0.97
SVM	88.7	0.94	0.94	0.94
LR	92.1	0.94	0.98	0.96
DT	88.4	0.94	0.93	0.94
KNN	93.5	0.94	0.99	0.97
RF	94.1	0.94	1.00	0.97
AC	92.4	0.95	0.98	0.96
SGC	89.3	0.94	0.95	0.94
GBM	92.6	0.94	0.98	0.96
NB	93.2	0.94	0.99	0.97

The output, displayed in the figure: 3 is prominently on the user interface, indicates that the tweet is classified as "Real." This result underscores the system's ability to accurately discern genuine tweets from potentially manipulated or machine-generated ones, leveraging the robust feature extraction capabilities of CountVectorizer and the high accuracy of the Random Forest algorithm.

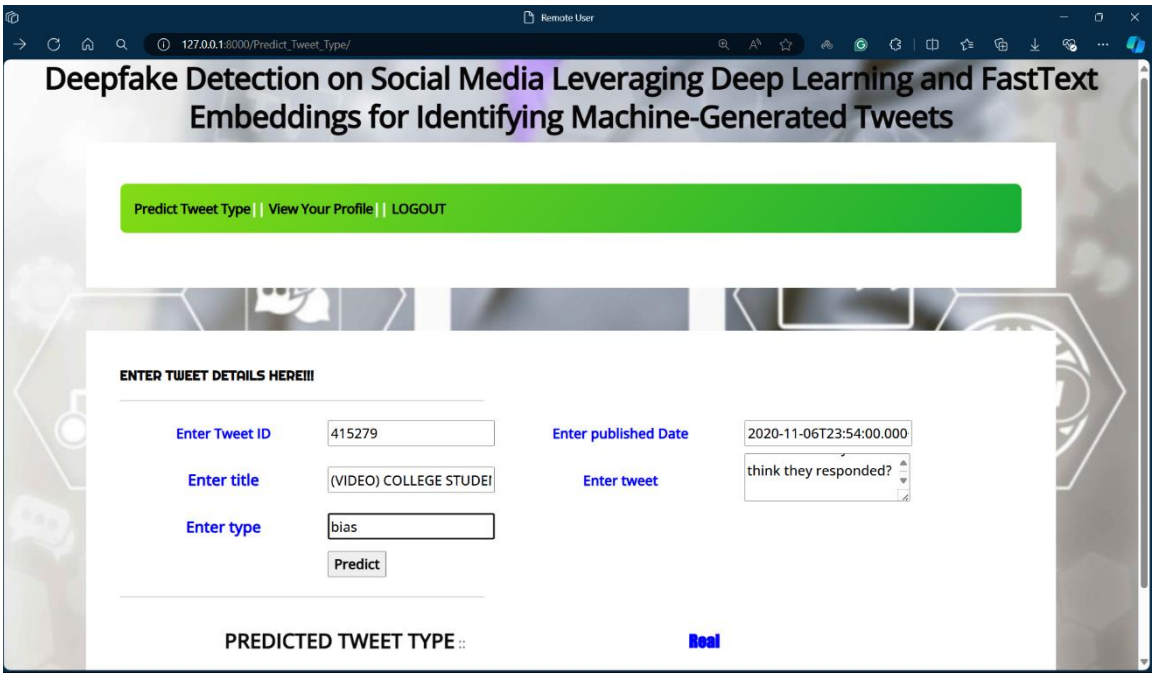

Fig. 3: Tested Result - Real

In this below figure. 4, the deepfake detection system, utilizing CountVectorizer for feature extraction and Random Forest for classification, has identified the tweet as "Deepfake." The interface displays the tweet details and the classification result, demonstrating the system's capability to accurately detect machine-generated content.

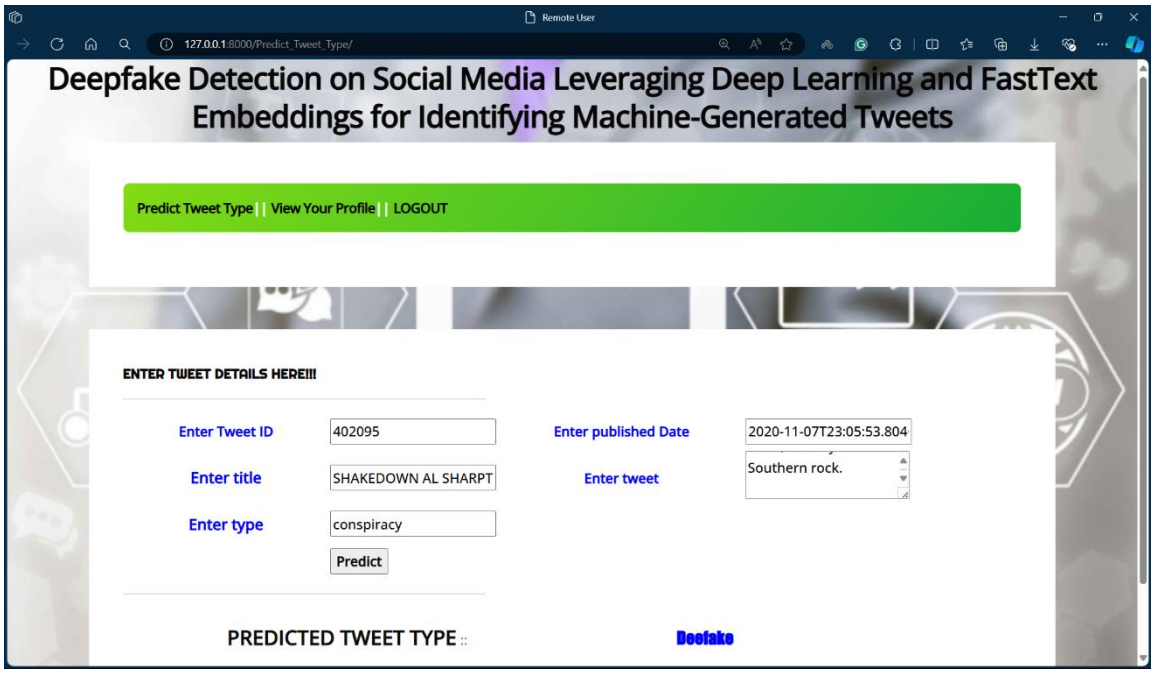

Fig. 4: Tested Result – Deepfake

VII. CONCLUSION

Deepfake content discovery is a basic and challenging errand in the time of deception and controlled substance. This think about pointed to address this challenge by proposing an approach for deepfake content discovery and assessing its adequacy. A dataset containing tweets of bots and people is utilized for investigation by applying a few machine learning with include building methods. Well-known include extraction method: CountVectorizer (CV) utilized. By leveraging a combination of strategies such as RF and CV the proposed approach illustrated promising comes about with a 0.94 accuracy score in precisely recognizing deepfake content. Moreover, the comes about of the proposed approach are compared with other state-of the- craftsmanship machine learning models from past literature. Overall, the selection of a RF show structure in this ponder appears its prevalence in terms of straightforwardness, computational productivity, and taking care of out-of-vocabulary terms.

VIII. FUTURE SCOPE

The extend shows considerable potential for future improvements and extensions, clearing the way for proceeded progressions in deepfake location and social media examination. One road for future investigation includes the refinement and enlargement of existing location calculations to reinforce exactness and effectiveness. Furthermore, coordination machine learning models with the framework seem upgrade its capability to adjust and advance in reaction to rising deepfake techniques.

- Refinement and increase of existing discovery calculations for moved forward precision and efficiency
- Integration of machine learning models to upgrade flexibility and responsiveness to developing deepfake techniques
- Expansion of stage usefulness to envelop a broader run of social media stages and substance types
- Collaboration with social media companies and cyber security specialists to create standardized conventions for deepfake location and mitigation
- Exploration of progressed advances like blockchain and decentralized systems to improve security and versatility against altering

IX. REFERENCES

[1] S. Kudugunta and E. Ferrara, ‘‘Deep neural networks for bot detection,’’ Inf. Sci., vol. 467, pp. 312–322, Oct. 2018.

[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] D. Dukić, D. Keča, and D. Stipić, ‘‘Are you human? Detecting bots on Twitter using BERT,’’ in Proc. IEEE 7th Int. Conf. Data Sci. Adv. Anal. (DSAA), Oct. 2020, pp. 631–636.

[4] K. E. Daouadi, R. Z. Rebaï, and I. Amous, ‘‘Bot detection on online social networks using deep forest,’’ in Proc. 8th Comput. Sci. On-Line Conf. Cham, Switzerland: Springer, 2019, vol. 2, no. 8, pp. 307–315.

[5] D. Ippolito, D. Duckworth, C. Callison-Burch, and D. Eck, ‘‘Automatic detection of generated text is easiest when humans are fooled,’’ 2019, arXiv:1911.00650.

[6] M. Heidari and J. H. Jones Jr., ‘‘BERT model for social media bot detection,’’ Mason Archival Repository Service, George Mason Univ. Libraries, Tech. Rep., 2022.

[7] J. Zhao, X. Liu, Q. Yan, B. Li, M. Shao, and H. Peng, ‘‘Multi-attributed heterogeneous graph convolutional network for bot detection,’’ Inf. Sci., vol. 537, pp. 380–393, Oct. 2020.

[8] B. Wu, L. Liu, Y. Yang, K. Zheng, and X. Wang, ‘‘Using improved conditional generative adversarial networks to detect social bots on Twitter,’’ IEEE Access, vol. 8, pp. 36664–36680, 2020.

[9] N.Hajli,U.Saeed,M.Tajvidi,andF.Shirazi,‘‘Socialbotsandthespreadof disinformation in social media: The challenges of artificial intelligence,’’ Brit. J. Manage., vol. 33, no. 3, pp. 1238–1253, Jul. 2022.

[10] T.Fagni,F.Falchi,M.Gambini,A.Martella,andM.Tesconi,‘‘TweepFake: About detecting deepfake tweets,’’ PLoS ONE, vol. 16, no. 5, May 2021, Art. no. e0251415.

[11] S. Najari, M. Salehi, and R. Farahbakhsh, ‘‘GANBOT: A GAN-based framework for social bot detection,’’ Social Netw. Anal. Mining, vol. 12, no. 1, pp. 1–11, Dec. 2022.

[12] S. Feng, ‘‘TwiBot-22: Towards graph-based Twitter bot detection,’’ 2022, arXiv:2206.04564.

[13] S. H. Moghaddam and M. Abbaspour, ‘‘Friendship preference: Scalable and robust category of features for social bot detection,’’ IEEE Trans. Dependable Secure Comput., vol. 20, no. 2, pp. 1516–1528, Mar. 2023.

[14] B. Guo, X. Zhang, Z. Wang, M. Jiang, J. Nie, Y. Ding, J. Yue, and Y. Wu, ‘‘How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection,’’ 2023, arXiv:2301.07597.

[15] R. Shijaku and E. Canhasi, ‘‘ChatGPT generated text detection,’’ Tech. Rep., 2023.

[16] S. Mitrović, D. Andreoletti, and O. Ayoub, ‘‘ChatGPT or human? Detect and explain. Explaining decisions of machine learning model for detecting short ChatGPT-generated text,’’ 2023, arXiv:2301.13852.

[17] Y. Ma, J. Liu, F. Yi, Q. Cheng, Y. Huang, W. Lu, and X. Liu, ‘‘AI vs. human—Differentiation analysis of scientific content generation,’’ 2023, arXiv:2301.10416.