Fake Profile Detection Using Deep Learning Algorithm

D. Guna Shekar, K. Siva Likith, G. Lakshmi Priya, Ms. Suvitha S,

Student, Department of AI&DS, Muthayammal Engineering College, Rasipuram, Tamil Nadu, India.
Student, Department of AI&DS, Muthayammal Engineering College, Rasipuram, Tamil Nadu, India.
Student, Department of AI&DS, Muthayammal Engineering College, Rasipuram, Tamil Nadu, India.
Assistant Professor, Department of AI&DS, Muthayammal Engineering College, Rasipuram, Tamil Nadu, India.

---***---

Abstract - *The rise in e-scams, which can be attributed to approximately 30% of fake social media accounts, has emphasized the pressing need to identify these fraudulent profiles. Due to the limitations of the current model in handling multi-model networks, efforts have been made to address real-time issues. This research introduces an advanced deep-transfer learning model that enhances the detection of fake profiles by conducting a comprehensive analysis of diverse social media data samples. Our model collects a wide range of data from various social media platforms, including posts, likes, comments, multimedia content, user activity, login behaviors, and more. Each type of data is individually processed to identify suspicious patterns that are indicative of fake accounts. For example, discrepancies like male profiles predominantly posting about or using images of females. Similarly, audio signals undergo transformations such as 1D Fourier, Cosine, Convolutional, Gabor, and Wavelet Transforms. On the other hand, image and video data are processed using their 2D counterparts. Text data is transformed using Word2Vec, which assists our binary Convolutional Neural Network (bCNN) in distinguishing between genuine and fake profiles. Feature optimization is carried out using the Grey Wolf Optimizer (GWO) for 2D data and the Elephant Herding Optimizer (EHO) for 1D data, ensuring minimal redundancy in features. Subsequently, separate 1D CNN classifiers are employed to classify the refined features and identify fake profiles. The results from these classifiers are combined through a boosting mechanism. Our findings demonstrate an increase in accuracy by 8.3%, precision by 5.9%, and recall by 6.5% compared to traditional methods.*

***Key Words*: Social Media, Fake, Profile GWO, EHO, CNN, Multimodal, Cold Start, Issues.**

1.INTRODUCTION

The rise of big data platforms, particularly social media, has brought about a new set of challenges, including identity theft, due to their widespread popularity. Unfortunately, spammers and con artists have taken advantage of these platforms, leading to an increase in cybercrime activities such as spamming. The impact of these threats on Social Media Platforms (SMPs) varies, but their existence cannot be denied.

Understanding human interpersonal relationships can provide valuable insights into user behavior, preferences, and conversational tendencies. While these insights can enhance the quality of products and services, they can also be used against unsuspecting users. For example, online discourse can be manipulated by actors who are unknown to other participants, altering the conversation's direction.

The anonymity provided by SMPs is their main attraction, but it can also be their downfall. Users can create fake profiles, causing distress to unsuspecting victims. Activities such as rumor-mongering, cyberbullying, and the dissemination of false information are examples of the negative use of these platforms. The competitive nature of SMPs, as demonstrated by metrics like "likes" or "followings," exacerbates the issue, prompting users to use both overt and covert tactics to gain an advantage.

1.1 Motivation and Contributions

The digital revolution has brought about a new era of global connectivity and information sharing. However, alongside these positive developments, there is a darker side that cannot be ignored. Recent estimates indicate that nearly one-third of all social media accounts are fake, highlighting the growing threat of e-scams and of news, communication, and business interactions for many individuals. The presence of fraudulent entities on social media undermines trust and sabotages genuine interactions, creating an environment that is ripe for cybercrimes.

Unfortunately, the existing models designed to address this issue often fall short. They are either ineffective, unable to handle multimodal data, or too complex to be applied in real-time situations. As social media continues to play an increasingly significant role in personal, professional, and societal spheres, there is an urgent need for an advanced solution that prioritizes user safety, trust, and platform integrity.

2. METHODOLOGY

The model in this study utilized various techniques such as XG Boost, a random forest method, and a profile-focused multi-layered neural network. These techniques were employed to extract observable features from the data, which were then saved in a CSV file for easy reading by the model. The authenticity of a profile is determined through the training, testing, and analysis of the model. To build the models, researchers opted for Google Colab due to its free GPU utilization. The Google Colab NVIDIA Tesla K80 GPU, with a capacity of 12 gigabytes (GB), can run continuously for 12 hours. This approach has proven effective in identifying fake profiles, potentially surpassing the accuracy of previous similar studies. Additionally, the design of the model emphasizes a visually appealing framework. A visual representation of the system architecture can be seen in Figure below.

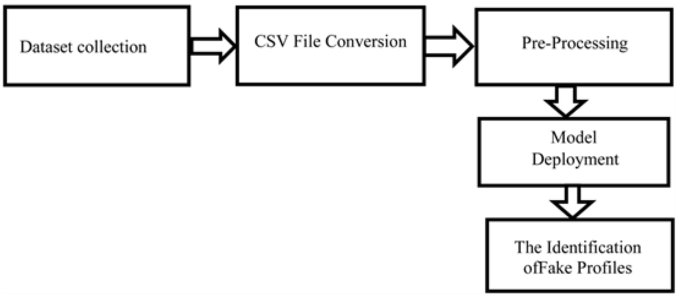

Fig -1:Methodology

3. EXISTING SYSTEM

In existing system implement support vector machine, random forest and other machine learning algorithms are used to detect the fake accounts. And also implementing a k-mean clustering algorithm on vector set to increase efficiency Supervised machine learning models require a label included in the corpus to predict the expected outcome. The existing systems use very fewer factors to decide whether an account is fake or not.

4. PROPOSED SYSTEM

In today's online social networks there have been a lot of problems like fake profiles, online impersonation, etc. To date, no one has come up with a feasible solution to these problems. In this project, intend to give a framework with which the automatic detection of fake profiles can be done so that the social life of people become secured and by using this automatic detection technique we can make it easier for the sites to manage the huge number of profiles using Multi-layer perceptron algorithm. Reduce the time complexity. Improve the accuracy in fake profile detection.

5. DATASET COLLECTION

According to the author [20], the MIB dataset was employed, consisting of 3474 genuine profiles and 3351 fake profiles. The dataset was composed of E13 and TFP for legitimate accounts, and TWT, INT, and FSF for fraudulent ones. The data was saved in CSV format for machine extraction. The characteristics used to identify fake profiles were displayed on the x-axis of Figure , while the number of entries for each feature in the dataset was shown on the y-axis after being selected during preprocessing.

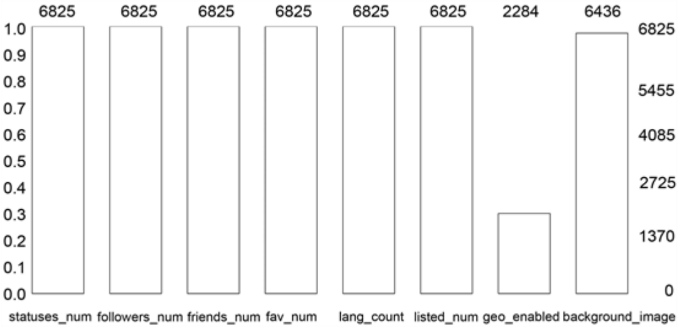

Chart -1: Collection

5.1 Model Development

In this section, the author has presented a proposed solution to address the challenge of identifying fake accounts by focusing on the characteristics of such situations. Initially, a calculation was performed to obtain the adjacency matrix of the social network's graph. Subsequently, another calculation was conducted to determine the degree of similarity among nodes (users of the social network) based on their network connections. Following this, similarity matrices were constructed for each of the mentioned metrics, which included similarity based on common friends, Jaccard similarity, cosine similarity, and other relevant measures. At this stage, multiple matrices were displayed to demonstrate the level of similarity between nodes.

All the data was categorized as normal because, in these circumstances, the data is imbalanced, with approximately 98-99 percent belonging to the majority class (normal users). This makes it challenging to accurately assess both the minority class (fake subscribers) and the overall classification accuracy. To address this issue, SMOTE was utilized to balance the statistics and tackle the problem effectively.

5.2 Processing Of Text Posts

Initially, all text posts are transformed into 1D Feature Vectors using a Word2Vec-based model. This model helps in representing individual words as numerical sets. To accomplish this, a dictionary of positive and negative words is utilized to determine the polarity score of each word in the posted text. Equation 1 is used for this purpose, where Wi, Wn, and Wp represent input, negative, and positive words from the dictionary data samples, and Pout represents the output polarity levels. The total number of input words is denoted by N. Along with the input corpus, this polarity score is then given to a Word2Vec Model, which is illustrated in Figure 2. Word2Vec is a technique used for sentiment analysis and can be employed to create tools like a thesaurus, context detector, and continuous bag of words

(CBoW) engine. The document reveals various sections that explain this fact.

The program builds a vocabulary by identifying "active words" in a given text. The context creator utilizes the language and constructs a suitable environment. The context builder block provides word pairs to aid in finding word choices with related words. With access to these variations through the CBoW engine and skip gramme models, the feature extraction process can be carried out more effectively. These features are then passed to the first step of a two-stage neural network, which converts the input word pairs into emotions. As a result of the overall sentiment analysis, this context-aware model can be used to determine the sentiment associated with each word combination. While this engine is accurate in labeling emotions and events, training it for real-time sample sets requires significant effort.

5.3 Processing the Other 1D Features

After converting text features into numerical sets, multidomain features are extracted from 1D features such as login patterns, posting patterns, and page activity patterns. This is done to identify patterns in both fake and normal user behaviors. To achieve this, various types of features such as Fourier Features for frequency analysis, Cosine Features for entropy analysis, Gabor Patterns for spatial analysis, Wavelet Patterns for enhanced spatial analysis, and Convolutional Features for window-based analysis are extracted using equations 2 through 8. These equations involve different components such as dimensions for different window and stride components, activation scaling constants, and angular and wavelength components to extract Gabor features. The approximate and detailed wavelet components might contain redundancies, which are reduced through an EHO process discussed in the next sub-section of this text.

$$DFT_i = \sum_{j=1}^{N_f} x_j \times \left[coscos\left(\frac{2 \times \pi \times i \times j}{N_f}\right) - \sqrt{-1} \times sinsin\left(\frac{2 \times \pi \times i \times j}{N_f}\right)\right] \quad (2)$$

Fig -2:Formula

6 ARTIFICIAL NEURAL NETWORK

The neural networks in deep learning systems exhibit similar behavior to those found in the human brain, as stated in reference [22]. The model is constructed with an input layer, three hidden layers, and an output layer, each containing nodes or neurons. Keras' sequential was utilized by the author in building the model, with activation functions present in all layers except the output layer. The sigmoid function was used to activate the output layer. The model was optimized using the Adam optimizer and binary merge loss function. This architecture is an example of an ANN. The sigmoid function outputs a value between 0 and 1, indicating whether a given profile is real or fake based on its prediction.

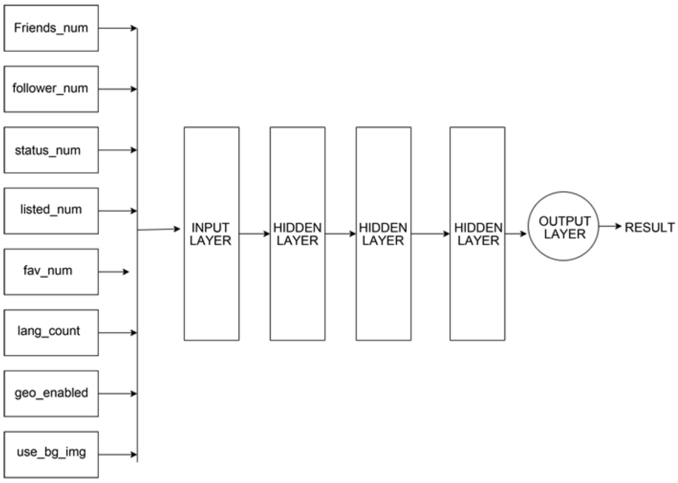

Chart -2: Working

7 THE EVALUATION STRATEGY UTILIZE

To establish a robust framework for identifying fake profiles on social media platforms, a meticulous approach was implemented to ensure the precision, accuracy, recall, and delay levels were rigorously evaluated in a transparent and replicable manner.

7.1 Data Preprocessing

To begin with, the datasets, which were a combination of various platforms, underwent a rigorous preprocessing regimen. Anomalous entries such as null or missing values were systematically eliminated. Additionally, outliers that had the potential to skew analytical outcomes were identified and appropriately managed. Given the inherent heterogeneity of data collected from different sources, normalization and standardization protocols were employed to ensure consistent data scaling. Categorical features were skillfully encoded, and textual data was tokenized to facilitate subsequent computational processes.

7.2 Model Training And Prediction

After refining the data, it was judiciously divided into training and test cohorts, with the former typically comprising either 70% or 80% of the total data volume. The cutting-edge deep-transfer learning model proposed in this research was rigorously trained using the training corpus. The model's predictive capabilities were then assessed on the test cohort following this intensive training regimen.

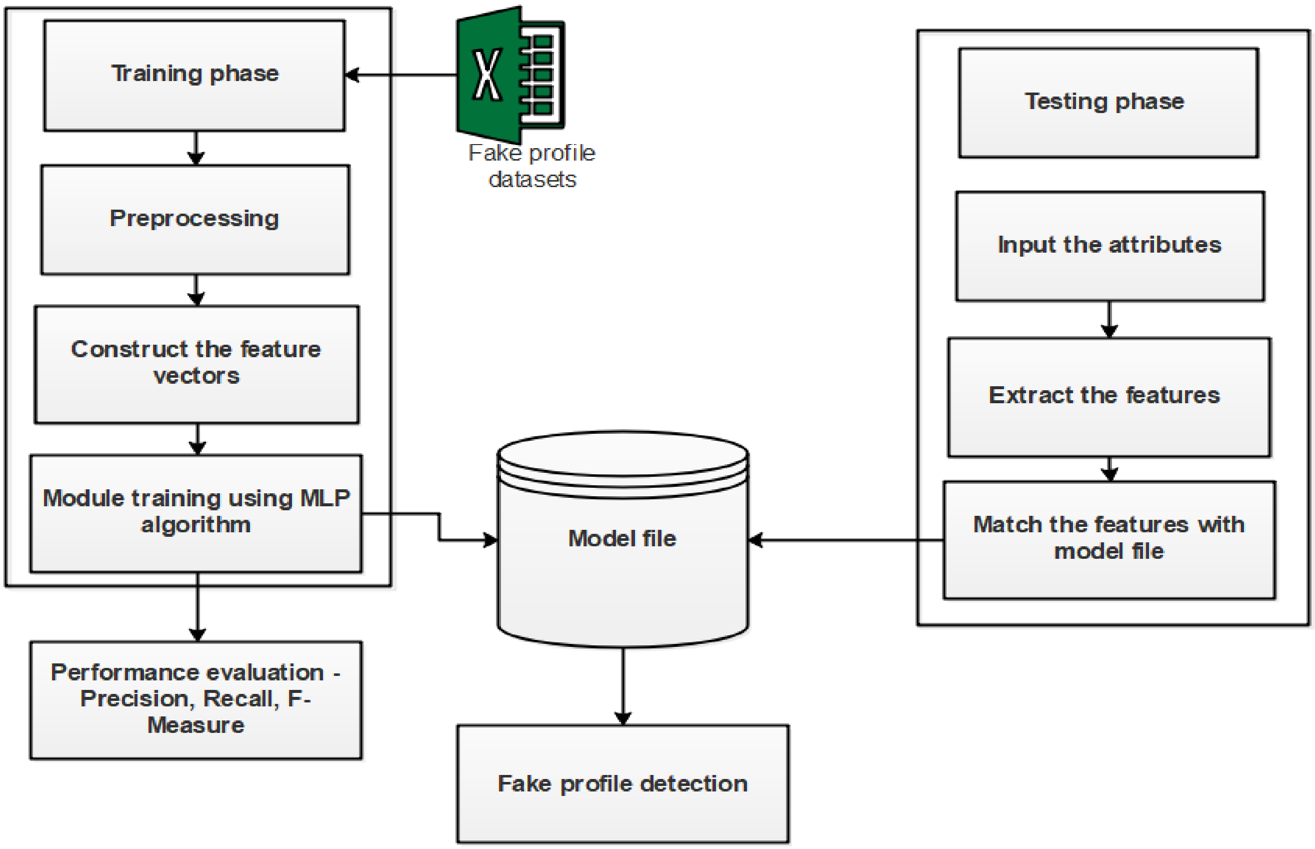
Training phase
Fake profile datasets
Preprocessing
Construct the feature vectors
Module training using MLP algorithm
Performance evaluation - Precision, Recall, F-Measure
Model file
Fake profile detection
Testing phase
Input the attributes
Extract the features
Match the features with model file

Fig -2:Formula

7.3 Computational Metrics

The main focus of this methodology was to calculate accuracy, precision, and recall. Accuracy, which provides a comprehensive view, was determined by calculating the ratio of correctly predicted instances to the total predictions. Precision, which indicates the model's exactness, was derived from the ratio of genuine identifications to the total positive predictions. On the other hand, recall reflected the model's completeness by capturing the proportion of actual positives that were correctly identified.

7.4 ASSESSMENT OF DELAY LEVELS

In addition to the conventional metrics, the study focused on evaluating delay levels, which are crucial for real-time applications. This involved measuring the time interval between the model instance and the subsequent generation of an assimilated prediction based on an input. These intervals were meticulously captured through systematic timestamping. A comprehensive analysis compared delay levels with parameters such as dataset volume, profile intricacy, and computational load. This provided detailed insights into potential scenarios where the model may experience increased delays.

7.5 AGGREGATION AND COMPARATIVE ANALYSIS OF RESULTS

After the computation, individual metric values obtained from different datasets were aggregated, typically using mean or median values, to provide an overall performance view. The methodology culminated in a comparative analysis, where the metric outcomes of the proposed model were compared to those of existing models. This highlighted the relative improvements of the proposed model and identified areas for potential refinements.

Based on this approach, the accuracy of fake profile detection in relation to Total Test Entries (TTE) can be observed in Table 2. According to this evaluation and Figure 4, it is evident that the proposed model exhibited a 10.5% higher accuracy in fake profile detection compared to GM [6], 8.3% better accuracy than DCNN [9], and 4.5% higher accuracy than PSVM [21]. As a result, it proves to be advantageous for a wide range of real-time use cases.

This improved accuracy is attributed to the utilization of multidomain and multimodal features, their efficient selection, and the binary cascade 1D CNN classifier, which has been trained to optimize classification performance across various data types.

8 HARDWARE CONFIGURATION

Processor processor 2.6.0 GHZ	: Dual core
RAM	: 1GB
Hard disk	: 160 GB
Compact Disk	: 650 Mb
Keyboard	: Standard keyboard
Monitor	: 15 inch color monitor

9 SOFTWARE CONFIGURATION

Operating system	: Windows OS
Front End	:PYTHON
Back End	: MYSQL
Application	: Web application

10 CONCLUSION AND FUTURE SCOPE

In conclusion, the proposed model utilizes various social media inputs to identify behaviors associated with fake profiles. The collected components are represented as multidomain feature sets and processed separately to detect patterns. The model uses different transforms for audio, image, and video signals and employs a binary Convolutional Neural Network (bCNN) model with Word2Vec to distinguish between genuine and fake profiles. The Grey Wolf Optimizer (GWO) processes 2D features, while the Elephant Herding Optimizer (EHO) processes 1D features, including Word2Vec, to reduce feature redundancy. The model uses distinct 1D CNN classifiers to classify both feature sets, making it easier to detect fake profiles. In the future, this model can be further improved and expanded to include more social media interfaces and patterns.

Based on the accuracy evaluation, the proposed model exhibited a 10.5% higher fake profile detection accuracy compared to GM [6], an 8.3% higher accuracy compared to DCNN [9], and a 4.5% higher accuracy compared to PSVM [21]. Consequently, it offers significant advantages for a wide range of real-time use cases. This improved accuracy can be attributed to the incorporation of multidomain and multimodal features, their efficient selection, and the utilization of a binary cascade 1D CNN classifier. The classifier has been specifically trained to optimize classification performance across diverse data types. In terms of precision, the proposed model demonstrated a 10.2% higher fake profile detection precision than GM [6], a 6.5% higher precision than DCNN [9], and a 3.9% higher precision than PSVM [21]. This highlights its utility across various real-time use cases. The precision is enhanced by employing different feature representation models for numerical datasets, audio, image, and video samples. These models are trained to maximize precision performance when applied to multiple data types. The combination of these models with the binary cascaded approach contributes to improved real-time classification performance.

The proposed model exhibited a 9.4% higher recall in detecting fake profiles compared to GM [6], a 6.1% better recall than DCNN [9], and a 4.5% higher recall than PSVM [21]. This higher recall demonstrates its effectiveness in a wide range of real-time use cases. The improvement in recall is achieved through the utilization of a 1D CNN classifier in binary cascade mode and multidomain feature sets, which enhance training for optimal recall performance across different data types.

When considering classification speed, the proposed model showed an 18.5% lower identification delay in detecting fake profiles compared to GM [6], a 19.5% lower identification delay compared to DCNN [9], and a 15.3% lower identification delay compared to PSVM [21]. This makes it highly beneficial for high-speed use cases. The reduction in identification delay is attributed to the use of EHO and GWO for feature selection, as well as a highly efficient 1D CNN classifier trained to maximize speed performance across various data types.

With these performance enhancements, the proposed model can now be effectively utilized in real-time social media fake profile detection scenarios. In the future, further improvements can be achieved by combining Generative Adversarial Networks (GANs), Auto Encoders (AEs), and Q-Learning, which will continuously enhance accuracy in different social media scenarios. Additionally, conducting an in-depth user profile analysis and incorporating explainable AI interfaces can enhance the analysis of fake-profile signatures, enabling real-time detection of fake profile sets.

REFERENCE

[1] P. Kantartopoulos, N. Pitropakis, A. Mylonas, and N. Kylilis, "Exploring adversarial attacks and defences for fake Twitter account detection," Technologies, vol. 8, no. 4, p. 64, Nov. 2020, doi: 10.3390/technologies8040064.

[2] K. Shahzad, S. A. Khan, S. Ahmad, and A. Iqbal, "A scoping review of the relationship of big data analytics with context-based fake news detection on digital media in data age," Sustainability, vol. 14, no. 21, p. 14365, Nov. 2022, doi: 10.3390/su142114365.

[3] H. Tuncer, "Interaction-based behavioral analysis of Twitter social network accounts," Appl. Sci., vol. 9, no. 20, p. 4448, Oct. 2019, doi: 10.3390/app9204448.

[4] M. A. Wani, N. Agarwal, and P. Bours, "Impact of unreliable content on social media users during COVID-19 and stance detection system," Electronics, vol. 10, no. 1, p. 5, Dec. 2020, doi: 10.3390/electronics10010005

[5] K. Machova, M. Mach, and M. Vasilko, "Comparison of machine learning and sentiment analysis in detection of suspicious online reviewers on different type of data," Sensors, vol. 22, no. 1, p. 155, Dec. 2021, doi: 10.3390/s22010155.

[6] H. Cai, J. Ren, J. Zhao, S. Yuan, and J. Meng, "KC-GCN: A semisupervised detection model against various group shilling attacks in recommender systems," Wireless Commun. Mobile Comput., vol. 2023, pp. 1–15, Feb. 2023, doi: 10.1155/2023/2854874.

[7] M. Mohammadrezaei, M. E. Shiri, and A. M. Rahmani, "Identifying fake accounts on social networks based on graph analysis and classification algorithms," Secur. Commun. Netw., vol. 2018, pp. 1–8, Aug. 2018, doi: 10.1155/2018/5923156.

[8] B. P. Kavin, S. Karki, S. Hemalatha, D. Singh, R. Vijayalakshmi, M. Thangamani, S. L. A. Haleem, D. Jose, V. Tirth, P. R. Kshirsagar, and A. G. Adigo, "Machine learning-based secure data acquisition for fake accounts detection in future mobile communication networks," Wireless Commun. Mobile Comput., vol. 2022, pp. 1–10, Jan. 2022, doi: 10.1155/2022/6356152.

[9] K. L. Arega and E. Shewa, "Social media fake account detection for Amharic language by using machine learning," Global Sci. J., vol. 8, no. 6, pp. 1–11, 2020.

[10] P. Wanda and H. J. Jie, "DeepProfile: Finding fake profile in online social network using dynamic CNN," J. Inf. Secur. Appl., vol. 52, Jun. 2020, Art. no. 102465, doi: 10.1016/j.jisa.2020.102465.

[11] P. Chakraborty, M. M. Shazan, M. Nahid, M. K. Ahmed, and P. C. Talukder, "Fake profile detection using machine learning techniques," J. Comput. Commun., vol. 10, no. 10, pp. 74–87, 2022, doi: 10.4236/jcc.2022.1010006.

[12] J. B. Munga and P. Mohandas, "Feature selection for identification of fake profiles on Facebook," in Proc. 6th Kuala Lumpur Int. Conf. Biomed. Eng., vol. 86, J. Usman, Y. M. Liew, M. Y. Ahmad, and F. Ibrahim, Eds. Cham, Switzerland: Springer, doi: 10.1007/978-3-030-90724-2_53.

[13] M. Vyawahare and S. Govilkar, "Fake profile recognition using profanity and gender identification on online social networks," Social Netw. Anal. Mining, vol. 12, no. 1, Dec. 2022, doi: 10.1007/s13278-022-00997-3.

[14] E. P. Meshram, R. Bhambulkar, P. Pokale, K. Kharbikar, and A. Awachat, "Automatic detection of fake profile using machine learning on Instagram," Int. J. Sci. Res. Sci. Technol., pp. 117–127, May 2021, doi: 10.32628/IJSRST218330.

[15] F. Ajesh, S. U. Aswathy, F. M. Philip, and V. Jeyakrishnan, "A hybrid method for fake profile detection in social networkusing artificial intelligence," in Security Issues and Privacy Concerns in Industry 4.0 Applications. Hoboken, NJ, USA: Wiley, 2021, pp. 89–112, doi: 10.1002/9781119776529.ch5.

[16] K. Kaushik, A. Bhardwaj, M. Kumar, S. K. Gupta, and A. Gupta, "A novel machine learning-based framework for detecting fake Instagram profiles," Concurrency Comput., Pract. Exper., vol. 34, no. 28, p. e7349, Dec. 2022, doi: 10.1002/cpe.7349.

[17] S. Banerjee and A. Y. K. Chua, "Understanding online fake review production strategies," J. Bus. Res., vol. 156, Feb. 2023, Art. no. 113534, doi: 10.1016/j.jbusres.2022.113534.

[18] K. Balogun, A. B. J. Omar, M. Jabar, and M. M. Abdulmajid, "Spam detection issues and spam identification of fake profiles on social networks," J. Theor. Appl. Inf. Technol., vol. 95, pp. 5881–5895, Mar. 2017.

[19] P. K. Roy and S. Chahar, "Fake profile detection on social networking websites: A comprehensive review," IEEE Trans. Artif. Intell., vol. 1, no. 3, pp. 271–285, Dec. 2020, doi: 10.1109/TAI.2021.3064901.

[20] M. Aljabri, R. Zagrouba, A. Shaahid, F. Alnasser, A. Saleh, and D. M. Alomari, "Machine learning-based social media bot detection: A comprehensive literature review," Social Netw. Anal. Mining, vol. 13, no. 1, p. 20, Jan. 2023, doi: 10.1007/s13278-022-01020-5.

[21] S. Raza and C. Ding, "Fake news detection based on news content and social contexts: A transformer-based approach," Int. J. Data Sci. Analytics, vol. 13, no. 4, pp. 335–362, May 2022, doi: 10.1007/s41060 -021-00302-z

[22] A.Saravanan and V. Venugopal, "Detection and verification of cloned profiles in online social networks using MapReduce based clustering and classification,'' Int. J. Intell. Syst. Appl. Eng., vol. 11, no. 1, pp. 195–207, Jan. 2023.

[23] Y. Elyusufi, Z. Elyusufi, and M. A. Kbir, "Social networks fake profiles detection ed on account setting and activity," in Proc. 4th Int. Conf. Smart City Appl. New York, NY, USA: Association for Computing Machinery, Oct. 2019, pp. 1–5, doi: 10.1145/3368756.3369015.