FAKE SOCIAL MEDIA PROFILE DETECTION USING ENSEMBLE LEARNING

M.Sowmya Devi
Department of Computer Science and Engineering(AIML)
Dhanekula Institute of Engineering and Technology
Vijayawada,Andhra Pradesh, India s

Dr.M.Vinod Kumar
Faculty of Computer Science and Engineering(AIML)
Dhanekula Institute of Engineering and Technology
Vijayawada,Andhra Pradesh, India

B.Pranab Karthik
Department of Computer Science and Engineering(AIML)
Dhanekula Institute of Engineering and Technology
Vijayawada,Andhra Pradesh, India

R.Sai Vamsi Krishna
Department of Computer Science and Engineering(AIML)
Dhanekula Institute of Engineering and Technology
Vijayawada,Andhra Pradesh, India

Y.Sandhya
Department of Computer Science and Engineering(AIML)
Dhanekula Institute of Engineering and Technology
Vijayawada,Andhra Pradesh, India

***Abstract*— Social media platforms like Twitter, Facebook, and Instagram have transformed communication, allowing users to exchange information, interact with communities, and share opinions. However, the increasing presence of fake profiles has introduced challenges such as misinformation, spam, and fraudulent activities. Detecting these deceptive accounts is essential for preserving social media integrity, yet conventional detection methods struggle to keep pace with evolving fraudulent techniques. This research presents an approach to identifying fake profiles on Twitter by leveraging machine learning and deep learning models. The methods employed include Decision Tree, Support Vector Machine (SVM), Random Forest, XG Boost, Naïve Bayes, Logistic Regression. The dataset consists of multiple text-based attributes, such as URLs, account usernames, favorite counts, language, and timestamps. To improve detection accuracy, feature selection and preprocessing techniques are applied to optimize model efficiency. Each algorithm contributes to recognizing distinct patterns that differentiate genuine accounts from fake ones. this study enhances social media security by introducing a reliable and scalable detection framework. The proposed method helps combat misinformation and fraudulent activities, offering a practical solution for maintaining trust and authenticity in online interaction**

Keywords*— *Fake Profile Detection, Machine Learning, Social Media Security, Twitter Data Analysis, Ensemble Learning.

I. INTRODUCTION

Social media has become an integral part of modern communication, allowing users to connect, share information, and express their thoughts globally. Platforms like Twitter, Facebook, and Instagram enable rapid content dissemination, but they also face a growing issue—the presence of fake profiles. These fraudulent accounts are often created to manipulate public opinion, spread false information, conduct scams, and generate spam. Such activities not only mislead users but also pose security risks, making fake profile detection a crucial challenge for maintaining trust and authenticity in digital spaces.

Identifying fake profiles is complex, as fraudsters constantly develop new strategies to bypass security measures. Traditional detection methods, such as manual verification and rule-based filters, are insufficient for handling the increasing volume and sophistication of these deceptive accounts. Therefore, automated approaches using machine learning (ML) have gained prominence in detecting fake profiles by analyzing patterns and behaviors that differentiate fraudulent accounts from legitimate ones. this study presents a detection framework employing multiple ML models, including Decision Tree, Support Vector Machine (SVM), Random Forest, XG Boost, Naïve Bayes, Logistic Regression. The dataset used in this research consists of various text-based attributes, such as URLs, usernames, favorite counts, language, and timestamps. Feature selection and preprocessing techniques are applied to enhance the model's accuracy, ensuring more precise classification of real and fake profiles. the primary goal of this research is to develop a scalable and efficient fake profile detection system that improves security on social media platforms. By leveraging multiple learning algorithms, this approach enhances detection accuracy and adaptability, helping to mitigate misinformation, prevent cyber fraud, and create a more secure online environment.

II. LITERATURE REVIEW

The rapid expansion of social media has led to an increase in fake profiles, which are often created for malicious purposes such as misinformation spread, phishing, and fraudulent activities. Traditional detection methods primarily rely on user reports and manual moderation, which are inefficient due to delays and the ever-evolving nature of deceptive techniques. To address these challenges, researchers have explored various automated approaches to enhance detection accuracy and efficiency. One widely studied aspect of fake profile detection is the identification of bots, human-controlled fake accounts, and hybrid accounts known as cyborgs. These accounts exhibit unique behavioural patterns, such as unnatural posting frequency, automated responses, and engagement with misleading content. Researchers have classified detection techniques into profile-based analysis, which focuses on account attributes such as profile pictures, names, and descriptions, and activity-based analysis, which examines interaction patterns, posting frequency, and engagement metrics.

Several studies have proposed machine learning models for detecting fraudulent accounts. Supervised learning techniques, including decision trees, support vector machines, and ensemble methods, have been used to classify accounts based on labeled datasets. Some research also highlights the effectiveness of deep learning architectures, such as recurrent and convolutional neural networks, in analyzing text-based features and behavioral patterns to enhance detection accuracy. Another area of focus is the detection of fake profiles involved in misinformation dissemination. Researchers have investigated the characteristics of accounts that frequently share misleading content, emphasizing the need for real-time detection systems. The increasing presence of bots on social media platforms has raised concerns about cybersecurity threats, leading to the development of AI-driven detection mechanisms that analyze linguistic patterns, sentiment, and engagement metrics. The use of

hybrid models, combining multiple machine learning algorithms, has been explored to improve robustness and adaptability to new deceptive tactics. Feature engineering techniques, such as analzying metadata, social connections, and user activity logs, have also been effective in distinguishing fake accounts from genuine users.

Given the dynamic nature of social media fraud, continuous advancements in detection methods are necessary. Future research should focus on refining machine learning models, integrating real-time detection capabilities, and incorporating diverse data sources to create more reliable and scalable fake profile detection systems.

III. Proposed Fraud Detection Appraoch

The proposed system utilizes machine learning algorithms to accurately classify social media profiles as real or fake. Various models, including Decision Tree, Random Forest, XG Boost, Naïve Bayes, SVM, and Logistic Regression, are trained on a dataset containing both genuine and fraudulent profiles. These models analyze text-based attributes such as URLs, account names, engagement statistics, timestamps, and linguistic patterns to identify deceptive accounts. The system is designed to operate in real-time through a Flask-based web application, providing an intuitive and efficient method for fake profile detection.

A. *ARCHITECTURE*

The architecture of the proposed system is structured into multiple stages to ensure a systematic and efficient detection process. The key components include Data Preprocessing, Feature Extraction, Model Training, Prediction, and Deployment, each contributing to the overall effectiveness of the system.

1. Data Collection from Social Media

the first step involves collecting profile-related data from social media platforms. The dataset comprises various attributes, including Usernames & Account Names are Used to recognize naming patterns and inconsistencies. URLs & Links are Identifies potentially malicious or suspicious links often associated with fraudulent accounts. Timestamps are Assesses account creation time and activity patterns. Engagement Metrics are Examines user interactions such as likes, comments, and shares. Language & Text-Based Features Analyzes textual content for linguistic patterns and anomalies. The dataset can be sourced from publicly available repositories or manual labeled for effective training.

2. Data Preprocessing

Before utilizing the dataset in machine learning models, it undergoes necessary cleaning and transformation to improve consistency and quality. Key preprocessing steps includes Handling Missing Values Filling in or removing incomplete records. Removing Duplicates are Ensuring data uniqueness and preventing redundant records. Text Normalization are Standardizing text by converting it to lowercase, removing special characters, and eliminating unnecessary words. Encoding Categorical Data are Converting text-based attributes into numerical representations. Scaling Numerical Features are Normalizing numeric data, such as engagement metrics, for better model performance.

3. Feature Extraction

Feature extraction is a key step that transforms raw data into structured numerical representations, enhancing the model's accuracy. It consists of Profile-Based Features areAccount Age are Calculating the time difference between the account creation date and the present username Patterns are Identifying similarities or anomalies in usernames associated with fake accounts. Profile Completeness Checking if profile details, including bio, profile picture, and personal information, are provided. Activity-Based Features are Posting Frequency is Evaluating the frequency of content uploads. Engagement Metrics like Assessing likes, comments, and shares to identify abnormal. Extracting sentiment, word structures, and content similarity to detect patterns. Keyword Detection is used to Identifying specific words or links often associated with suspicious behaviour. Detecting irregular login behaviours and interaction patterns typical of bot activities. This process ensures that only the most informative and relevant attributes contribute to the model's learning process.

4. Model Training

Once the features are extracted, different machine learning models are utilized to differentiate between genuine and fraudulent user accounts. The models employed in this research include:

After extracting the essential features, the dataset is divided into training and testing sets to ensure a fair evaluation of the models. During the training phase, machine learning algorithms analyze patterns within the labeled data to distinguish between real and fraudulent profiles. Each model is fine-tuned using hyperparameter optimization to enhance its predictive capability. Naïve Bayes is a probabilistic machine learning algorithm based on Bayes' Theorem, which assumes that all features are independent of each other. Decision Tree classifiers operate by making hierarchical splits based on the most relevant attributes, whereas Random Forest enhances this approach by combining multiple decision trees to improve accuracy and reduce the likelihood of overfitting. Boosting techniques such as XG Boost iteratively refine weak learners, adjusting their importance to minimize classification errors. The Support Vector Machine (SVM) algorithm maps data points into a higher-dimensional space and determines the optimal boundary that separates genuine and fake accounts. Logistic Regression, on the other hand, applies a probabilistic approach using a sigmoid function to determine the likelihood of a profile being fraudulent. To ensure the models generalize well to unseen data, validation techniques like k-fold cross validation are applied. Additionally, performance is assessed using metrics such as accuracy, precision, recall, and F1-score, allowing the selection of the most effective model for detecting fake profiles.

5. Performance Evaluation

To determine the efficiency of the trained models, several evaluation methods are applied. Accuracy Computes the percentage of correctly classified accounts out of the total predictions made. Precision indicates the proportion of correctly identified fake accounts among all those predicted as fake, while recall measures how well the model detects actual fraudulent accounts. F1-Score is a balanced metric that considers both precision and recall to provide a comprehensive evaluation of classification performance.

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

$$\text{Precision} = TP / (TP + FP)$$

$$\text{Recall} = TP / (TP + FN)$$

$$\text{F1-Score} = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$$

$$\text{Cross-Validation Mean Accuracy} = (1 / k) * \Sigma \text{ Accuracy}$$

6. Model Deployment

Once the best-performing model is selected, it is deployed to make real-time predictions on user accounts.

7. System Monitoring

Once deployed, continuous monitoring is essential to maintain system efficiency against evolving fraudulent tactics. This includes Periodic Model Updates, Real-Time Alerts , User Feedback Integration.

8. Predicting Fake or Real Accounts

The verification of online accounts involves a systematic process to determine whether an account is genuine or fraudulent. When an account is submitted for verification, the system begins by extracting key details such as username characteristics, activity patterns, interaction rates, timestamps, and the presence of external links. These attributes provide crucial insights into the nature of the account. Once the necessary features are extracted, they are processed through a machine learning model trained on a dataset containing both authentic and fake accounts. The model analyzes the provided data, identifies hidden patterns, and generates a classification result, labeling the account as either "fake" or "real." The decision is based on statistical correlations and learned trends from previous data. this automated verification process helps in efficiently identifying fake accounts that could be involved in spamming, misinformation, or fraudulent activities. Since the system is powered by machine learning, it continuously learns from new data, improving its accuracy over time. By integrating such models into account verification workflows, platforms can enhance security, protect users, and ensure a safer online environment.

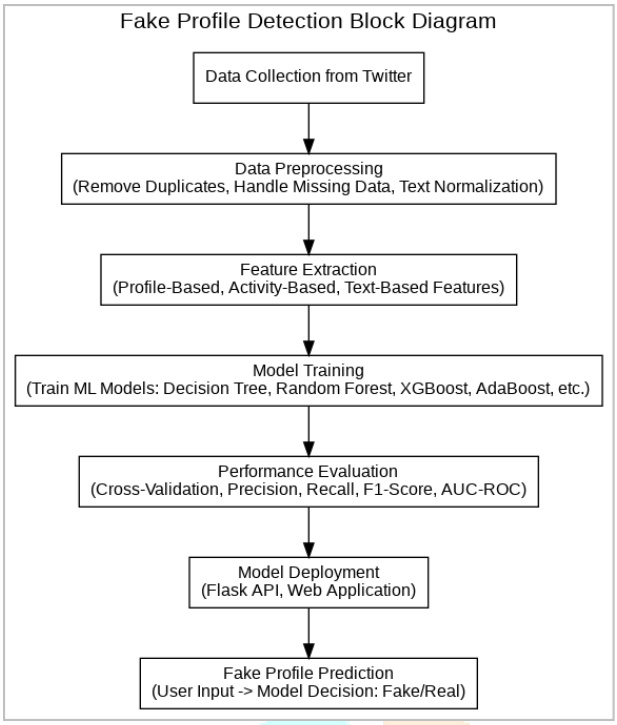

Fig. 1. Work flow of the Architecture

The proposed system architecture is designed to effectively detect fake profiles by analyzing various features extracted from user data. The process begins with collecting data from Twitter, which includes user details, tweets, and metadata. This raw data is then preprocessed to remove inconsistencies, handle missing values, and normalize textual information. Next, significant features are extracted, including profile attributes, user activity patterns, and textual characteristics, which help differentiate between genuine and fraudulent accounts. Machine learning models such as Decision Tree, Random Forest, XGBoost, AdaBoost, are trained on this processed data to recognize patterns associated with fake profiles. To ensure the reliability of predictions, the models are evaluated using performance metrics like cross-validation, precision, recall, F1-score, and support. The most accurate model is then deployed using a Flask API and integrated into a web-based system, allowing users to input profile details for classification. Based on the trained model's analysis, the system determines whether an account is real or fake, thereby enhancing security and trust in social media platforms.

B. WORKING WITH DATA SETS

To enhance the efficiency and accuracy of the fake profile detection model, unnecessary columns were eliminated from the dataset, retaining only the most relevant features for training and testing. The dataset contains various attributes related to Twitter profiles, such as the number of followers, friends, statuses, favorites, and other profile-related information. Figures **Fig. 2** and **Fig. 3** display the refined dataset used for model development. These selected attributes play a crucial role in distinguishing between real and fake accounts by analyzing user activity and engagement patterns. The refined dataset enables the model to focus on significant characteristics, improving its ability to make accurate predictions.

	favourites_count	followers_count	statuses_count	friends_count	default_profile
0	0	3	34	266	1.0
1	39	32	718	189	NaN
2	22588	17806	84194	16707	NaN
3	360	225	1202	441	NaN
4	381	692	4231	2001	NaN

Fig. 2. Training Data Set

default_profile_image	profile_use_background_image	utc_offset	listed_count	geo_enabled	lang_num
NaN	1.0	NaN	0	NaN	5
NaN	1.0	NaN	2	NaN	5
NaN	1.0	-10800.0	43	1.0	8
NaN	1.0	7200.0	1	1.0	9
NaN	1.0	19800.0	8	1.0	5

Fig. 3. Training Data Set

IV. PERFORMANCE EVALUATION

Evaluating model performance is crucial to ensuring the accuracy and reliability of predictions, particularly when differentiating between real and fake profiles. Various statistical metrics, including accuracy, precision, recall, F1-score, and support, confusion matrix provide insights into the effectiveness of trained models while also helping to mitigate the risks of overfitting. Logistic Regression, a probability-based classification algorithm, evaluates numerical attributes to determine whether an account is real or fake and achieved an accuracy of 97.22%, support value is of 1365, and precision is 0.98, recall is 0.96, and F1-score values is 0.97 each. making it a reasonable choice for binary classification problems. However, its accuracy was slightly lower than that of ensemble models. The Decision Tree model follows a structured, rule-based classification approach and demonstrated strong results with a testing accuracy of 98.68%, support value is of 1365, and precision is 0.99, recall is 0.99, and F1-score values is 0.99 each. This model exhibited high classification performance while maintaining stability. The Random Forest model further enhances prediction reliability by integrating multiple decision trees to improve accuracy and reduce overfitting, achieving a testing accuracy of 98.61%, support is 1365, while yielding a precision of 0.98, recall is 0.99 andF1-score is 0.99. These results highlight the robustness of ensemble learning methods. Naïve Bayes is a probabilistic machine learning algorithm based on Bayes' Theorem, which assumes that all features are independent of each other. It is efficient for classification tasks with testing accuracy of 94.51% %, support value of 1365, and precision is 0.97, recall is 0.92, and F1-score values is 0.94 each. XG Boost, a highly efficient boosting algorithm optimized for large datasets, employs gradient boosting techniques to achieve high classification performance. The model attained the highest testing accuracy of 98.83%, support value of 1365, and precision is 0.98, recall is 0.99, and F1-score values is 0.99 each indicating strong generalization and reduced risk of overfitting. Meanwhile, Support Vector Machine (SVM), a supervised learning algorithm that maximizes the margin between classes for optimal separation between real and fake profiles, achieved a training accuracy of 96.85%, support value is of 1365, and precision of 0.98, recall is 0.96, and F1-score values is 0.97 each, showcasing good classification performance though slightly behind ensemble techniques. Upon analyzing the above results, it is evident that ensemble models such as XG Boost, Random Forest, and Decision Tree consistently outperform other approaches due to their high accuracy, stability, and generalization ability. While Naïve Bayes, Logistic Regression, and SVM also delivered moderate results, they were slightly less effective in comparison to ensemble-based learning methods. These evaluation metrics emphasize that selecting the most suitable model depends on factors such as accuracy, computational efficiency, and robustness against overfitting, making ensemble models the optimal choice for fake profile detection.

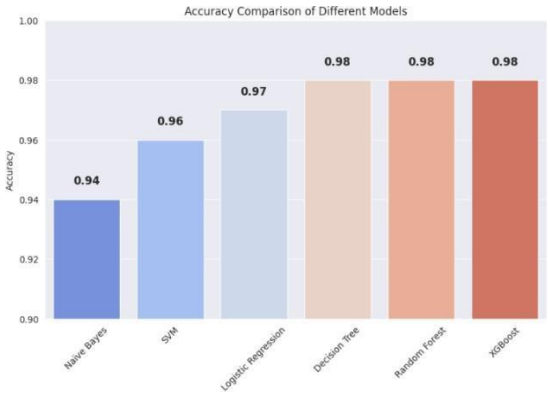

Fig. 4. Train Accuracy and Test Accuracy

	Model	Precision	Recall	F1-Score	Support
0	Naïve Bayes	0.97	0.92	0.94	1365
1	SVM	0.98	0.96	0.97	1365
2	Logistic Regression	0.98	0.96	0.97	1365
3	Decision Tree	0.99	0.99	0.99	1365
4	Random Forest	0.98	0.99	0.99	1365
5	XGBoost	0.98	0.99	0.99	1365

Fig. 5. Performance Scores

The evaluation results show that ensemble models such as AdaBoost, XGBoost, and next the Random Forest and Decision Tree achieved the highest classification accuracy. Traditional machine learning models like Logistic Regression, SVM,Naïve Bayes performed well but were outperformed by boosting techniques. Based on this analysis, the most accurate model was selected for deployment.

V. Conclusion

After thoroughly evaluating multiple machine learning models, I have selected **ensemble-based algorithms**, specifically **Random Forest, AdaBoost, and XGBoost**, for detecting fake profiles. These models consistently demonstrated superior performance in terms of **accuracy, precision, recall, and F1-score**, effectively capturing complex patterns within the data. Among them, **XGBoost emerged as the final model** due to its exceptional ability to handle imbalanced data, optimize computational efficiency, and enhance predictive accuracy. Its gradient boosting framework ensures improved generalization while minimizing overfitting. The results confirm that **XGBoost is the most reliable and scalable solution** for distinguishing between fake and real profiles, making it ideal for real-world deployment.

Fig. 6. Deployment fig

Fig. 7. Profile Detection

Fig. 8. Profile Detection

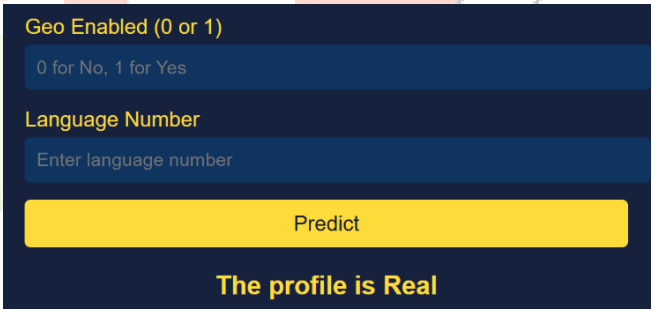

Fig. 9. Reporting the Profile

REFERENCES

[1] P. Chinnasamy, Ramesh Kumar Ayyasamy, Suthashini Subramaniam, Anbuselvan Sangodiah, Aqsa Iftikhar, and Ajmeera Kiran, "Fake Social Media Profile Identification Using Machine Learning," 2024 International Conference on Signal Processing,Electronics,Power,andTelecommunication(IConSCEP), TamilNadu,India,2024,pp.5,doi:10.1109/ICONSCEPT61884.2024.10627774.

[2] G. Bharath, K. J. Manikanta, G. B. Prakash, R. Sumathi and P. Chinnasamy, "Detecting Fake News Using Machine Learning Algorithms," *2021 International Conference on Computer Communication and Informatics (ICCCI)*, Coimbatore, India, 2021, pp. 1-5, doi: 10.1109/ICCCI50826.2021.9402470

[3] P. Chinnasamy, N. Kumaresan, R. Selvaraj, S. Dhanasekaran, K. Ramprathap and S. Boddu, "An Efficient Phishing Attack Detection using Machine Learning Algorithms," *2022 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC)*, Bhubaneswar, India,2022, pp.1-6,doi: 10.1109/ASSIC55218.2022.10088399.

[4] P. Chinnasamy, P. Krishnamoorthy, K. Alankruthi, T. Mohanraj, B. S. Kumar and L. Chandran, "AI Enhanced Phishing Detection System," *2024 Third International Conference on Intelligent*

Techniques in Control, Optimization and Signal Processing (INCOS), Krishnankoil, Virudhunagar district, Tamil Nadu, India, 2024, pp. 1-5, doi: 10.1109/INCOS59338.2024.10527485.

[5] U. D. Joshi, A. P. Singh, T. R. Pahuja, S. Naval and G. Singal, "Fake social media profile detection", Machine Learning Algorithms and Applications, pp. 193-209, 2021.

[6] H. M. F. Shehzad, A. Yasin, Z. K. Ansari, M. A. Khan and M. J. Awan, "Fake profile recognition using big data analytics in social media platforms", Int. J. Comput. Appl. Technol., vol. 68, no. 3, pp. 215, 2022

[7] K. Shu, A. Sliva, S. Wang, J. Tang and H. Liu, "Fake news detection on social media: A data mining perspective", ACM SIGKDD Explorations Newslett., vol. 19, no. 1, pp. 22-36, 2017.

[8] A. Bhattacharya, R. Bathla, A. Rana and G. Arora, "Application of Machine Learning Techniques in Detecting Fake Profiles on Social Media", the 9th International Conference on Reliability Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), pp. 1-8, 2021.

[9] A. Romanov, A. Semenov, O. Mazhelis and J.Veijalainen, "Detectionof Fake Profiles in Social Media - Literature Review", WEBIST 2017 - 13th International Conference on Web Information Systems and Technologies, vol. 1, pp. 363-369, 2017.

[10] N. Singh, T. Sharma, A. Thakral and T. Choudhury, "Detection of fake profile in online social networks using machine learning", Proc. IEEE Int.
Conf. Adv. Comput. Commun. Eng., pp. 231-234, 2018

[11] K. Krombholz, D. Merkl and E. Weippl, "Fake identities in social media: A case study on the sustainability of the facebook business model", J.
Service Sci. Res., vol. 4, no. 2, pp. 175-212, 2012

[12] R. P. Purba, D. Asirvatham and R. K. Murugesan, "Classification of instagram fake users using supervised machine learning algorithms", International Journal of Electrical and Computer Engineering (IJECE), vol. 10, no. 3, pp. 2763-2772, Juni 2020.

[13] G. Sansonetti, F. Gasparetti, G. D'aniello and A. Micarelli, "Unreliable users detection in social media: Deep learning techniques for automatic detection", IEEE Access, vol. 8, pp. 213154-213167, 2020.

[14] 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, pp. 20, Jan. 2023.

[15] T. K. Balaji, C. S. R. Annavarapu and A. Bablani, "Machine learning algorithms for social media analysis: A survey", Comput. Sci. Rev., vol. 40, May 2021.

[16] H. C. Soong, N. B. A. Jalil, R. K. Ayyasamy and R. Akbar, "The essential of sentiment analysis and opinion mining in social media: Introduction and survey of the recent approaches and techniques", 2019 IEEE 9th symposium on computer applications industrial electronics (ISCAIE), pp. 272-277, 2019, April

[17] P. Meel and D. K. Vishwakarma, "Fake news rumor information pollution in social media and web: A contemporary survey of state- of-the-arts challenges and opportunities", Expert Syst. Appl., vol. 153, Sep. 2020.

[18] P. K. Roy and S. Chahar, "Fake profile detection on social networking websites: A comprehensive review", IEEE Transactions on Artificial Intelligence, 2021.

[19] 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.

[20] F. Masood, G. Ammad, A. Almogren, A. Abbas, H. A. Khattak, I. U. Din, et al., "Spammer detection and fake user identification on social networks", IEEE Access, vol. 7, pp. 681