AuthentiCheck: Combating Fake Twitter Profiles with Advanced Machine Learning Techniques

Pravin Patil, Priyanshu Chaudhari, Nilesh Wankhede, Aditya Khandelwal and Tejas Sonar
Department of Information Technology, SVKM's Institute of Technology Dhule, India

Keywords: Neural Networks, NTScraper, Random Forest Algorithm, SVM (Support Vector Machines), Snscrape, Stream Lit, Twitter API, Web Scraping.

Abstract: The increasing prevalence of fake profiles on social media platforms, particularly Twitter, has become a pressing issue, impacting user trust, security, and the integrity of online interactions. These fake accounts—comprising bots, impostors, and malicious entities—are frequently used for phishing, identity theft, spreading misinformation, and influencing public opinion. Despite efforts by social media platforms to tackle this problem, the evolving tactics of cybercriminals demand more advanced and effective solutions.To address this, we developed AuthentiCheck, a browser plug-in that leverages cutting-edge machine learning technologies to detect and classify fake Twitter profiles. AuthentiCheck analyzes key behavioral attributes, such as tweeting frequency, follower growth, and engagement patterns, to distinguish between legitimate and fraudulent accounts. Using a dataset of over 9,600 manually labeled profiles, we trained and evaluated several state-of-the-art machine learning models, selecting the most accurate and salable model for deployment.AuthentiCheck offers users a seamless and user-friendly way to verify profile authenticity with a simple right-click on a Twitter ID. The plug-in provides real-time feedback on a profile's legitimacy, empowering users to navigate social media more securely. This work demonstrates how advanced machine learning techniques can combat the challenges posed by fake profiles and lays the foundation for future improvements in real-time detection systems

1 INTRODUCTION

Social media sites such as Twitter have revolutionary transformed modes of communication and socialization among individuals as well as information sharing among the population. These modes of interaction have become inescapable tools used in personal interaction, business communication, and news communication. It is within these parameters that Twitter has received widespread use because it allowed its users to post shorter messages, follow other people, and engage in very rapid conversations. As of 2018, Twitter had over 336 million active monthly users, making it one of the most widely used social networking sites.

However, with such widespread usage comes the advent of significant challenges, particularly the rise in fraudulent accounts. These may include bots, impersonators, and other malicious actors that pose a significant threat to users and the online environment at large. Fraudulent accounts are used for various illicit activities, including phishing schemes, identity theft, and the spread of false information. Further, they compromise the integrity of social media platforms by distorting public perception, exaggerating follower numbers, and participating in synchronized digital assaults.

As fake accounts are becoming more common, it is impossible not to create efficient detection and remediation techniques for the same. Although the social media websites have implemented deletion of such fraudulent accounts, cyber crooks' strategies are advancing so fast that keeping up with the same has become challenging. Therefore, a need for such advanced technologies that can detect fake profiles in no time.

The current paper presents a browser extension called AuthentiCheck, which identifies false accounts on Twitter with the help of machine learning. Using critical behavioral features from user profiles, such as follower increase, tweet frequency, and statistics for engagement, the paper examines an account as valid or malicious. This tool provides a simple method of assessing and minimizing fraudulent risks associated

Patil, P., Chaudhari, P., Wankhede, N., Khandelwal, A. and Sonar, T.
AuthentiCheck: Combating Fake Twitter Profiles with Advanced Machine Learning Techniques.
DOI: 10.5220/0013621500004664

In *Proceedings of the 3rd International Conference on Futuristic Technology (INCOFT 2025) - Volume 3*, pages 455-462
ISBN: 978-989-758-763-4

with profiles in Twitter by checking the genuineness of individual accounts. The paper reviews how the AuthentiCheck tool was developed, from data collection to model training, analysis of its effectiveness in deciding fake profiles. Finally, this research contributes to the effort to make social media safer and to improve users' trust in them.

2 LITERATURE SURVEY

The proliferation of social media, such as Twitter, brings along with it an extensive avenue for information exchange and personal interaction; however, the proliferation has led to the establishment of fake profiles that provide avenues for the spread of false and malicious activities. Platform-based identification of fake profiles on the Twitter platform calls for an analysis of complex patterns in behavior and text using advanced techniques like machine learning, feature engineering, and NLP (K. M. Manojkumar, G. Gudikoti, J. Naveen, S. B. Devamane and G. C. Lakshmikantha,2023) (K. Shreya, A. Kothapelly, D. V and H. Shanmugasundaram,2022). Although these methodologies are intrinsically very effective, each has its own limitations related to scale, precision, and adaptability that keep prodding further research into adaptation for current detection methods for real-time applications.

The research by Manojkumar and Shreya was based on the feature-based machine learning approach in which certain characteristics like the follower-to-following ratio, tweet frequency, and engagement metrics were used to train classifiers (K. M. Manojkumar, G. Gudikoti, J. Naveen, S. B. Devamane and G. C. Lakshmikantha,2023). Studies as mentioned above have depicted how the supervised machine learning techniques classified under Random Forest and Support Vector Machines SVM can be present, even for identifying fake profiles. However, such methods require a lot of training data, which is hard to be obtained and to be maintained over dynamic social media scenarios.

Linguistic-based approaches employ NLP for textual feature analysis, which involves language pattern, sentiment, and syntactic structure. (Bhatia et al. ,2023) presented a technique that used NLP and deep learning to identify sources of fake news on Twitter, applying it to demonstrate the utility of the recurrent neural network in understanding textual cues that differentiate between authentic and fake profiles (P. Harris, J. Gojal, R. Chitra and S. Anithra,2021). These methods demonstrate the strength of NLP tools, including sentiment analysis and topic modelling, in detecting manipulated content. Still, issues regarding the management of several languages and dialects will hinder it from wider usage (Madhura Vyawahare and Sharvari Govilkar , 2022).

(Narayanan et al,2018), employ so-called hybrid models, where both feature engineering and machine learning algorithms are used together to maximize the accuracy of detection (T. Bhatia, B. Manaskasemsak and A. Rungsawang,2023). For instance, in the case of Narayanan's Iron Sense system, its ML algorithms were combined with feature-based methods. The way this was done is through the study of user activity along with metadata from the account and network connection behaviour.

Bio-inspired algorithms, in which recent examples would include the work of (Mahammed et al, 2022), can be used in evolutionary techniques to improve feature selection mechanisms when detecting fake profiles (L. P, S. V, V. Sasikala, J. Arunarasi, A. R. Rajini and N. Nithiya,2022). Despite being considered to be new and hence innovative, it is significantly restricted due to the computational requirements implicated in scalability. At any rate, this remains an area of inspiration derived from the changing dynamic and adaptive behaviour of users, as well as the ever-changing patterns developed on social media.

Research performed by (Harris,2021) and (Vyawahare&Govilkar,2022) analyzed the identification of spammers on platforms other than Twitter, including Instagram, where different behaviour of its users and unique feature spaces posed distinctive challenges (Sonowal, G., Balaji, V. & Kumar, N ,2024). Cross-platform studies enabled cross-platform comparisons and opened up the potential discussion for transferring knowledge learned on one platform to improve performance on another. Although theoretical, in practice, issues arise because of differences in data structures and user behaviour from one network to another, thus necessitating a customized model for each social network.

3 METHODOLOGY

3.1 Pre-processing

The web application includes the authentication of users before allowing them to access for the proposed model.

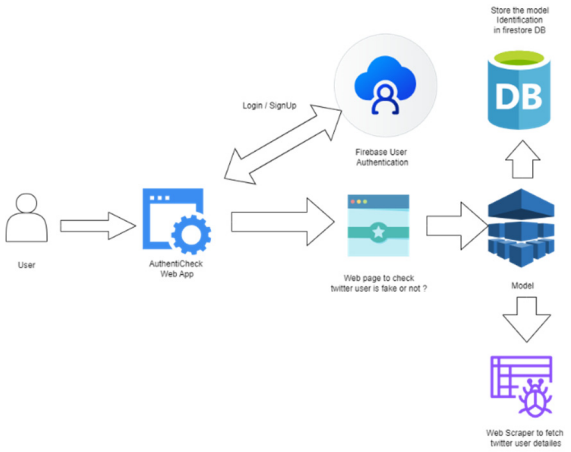

Figure 1: System Architecture

3.1.1 Data Retrieval

This ability of the system to acquire further data from web scraping tools such as NTScraper, snscrape or use the Twitter API to fetch the profile details like follower count, tweets, following count, and

3.1.2 User Interaction

A user inputs a Twitter profile username to verify authenticity using the front-end interface, that is, Stream lit application, HTML and CSS for designing the web application.

3.1.3 Twitter API Integration

An API, or Application Programming Interface, helps the connectivity between the twitter and application. In this case, it most likely helps the system to communicate with twitter, like fetching data or sending information to another module of the system.

3.1.4 Database Storage (Firebase Firestore)

The results of the analysis such as whether a profile is fake or real for reference later on to store on the Firebase database.

3.1.5 User History and Reporting

The system is capable of showing users what past profiles they have screened. This saved outcome means that they can monitor steadily and, therefore, observe patterns over time.

3.1.6 Display Outcomes

The user is presented with real-time feedback of the profile through the interface provided by Stream lit. In case the profile is a scam, then some additional information or risk factors like large follower-to-following ratio or other suspicious engagement-are also shown.

3.1.7 Visualization

The system provides different types of charts and graphs, to justify the outcomes made by system and gives recommendations or warnings to the user on the profile's authenticity.

3.2 Model Processing

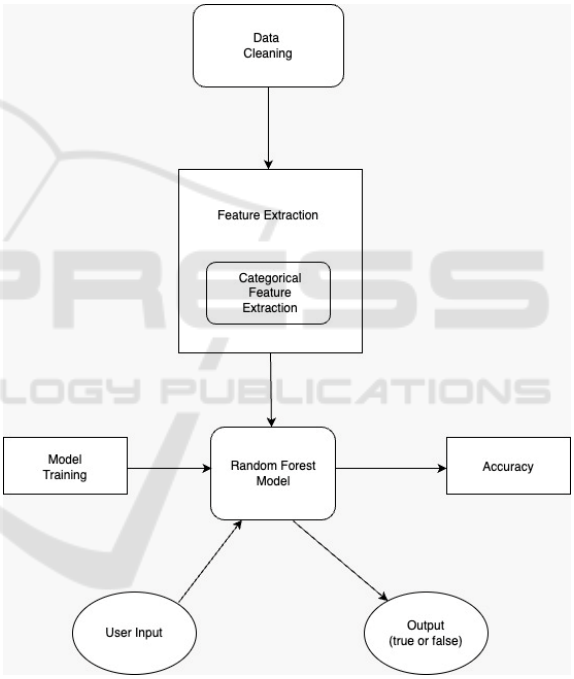

Figure 2: Model Processing

3.2.1 Data Cleaning

This includes cleaning of models by cleaning of data from which the retrieved data would then have its accuracy enhanced and reliability maintained. Handling profiles that did not have either a bio or a location ensured proper handling of this sort of profile. Numerical features such as followers count, following count and number of tweets were standardized.

3.2.2 Feature Selection

After pre-processing the data, following are some features related to identifying the fake profiles extracted.

- Account Age: Accounts which are too new are marked suspicious, using standard configurations, such as profile pictures or banner graphics, is considered to be very dangerous signs.
- Profile Completeness: A profile which does not possess important information, such as bio, location, or photo is more likely to be fake.
- Content Attributes: Repetitive or copied content helps detect patterns often associated with bots or fake accounts.

3.2.3 Feature Extraction

Categorical attributes, such as sidebar colours or the presence of a verified badge, are converted into numerical representations to make them more interpretable to the model.

Supplementary characteristics are derived from the primary data to enhance the model's ability for better prediction. The main features are made up of:

- Follower-to-Following Ratio: Values which are too high or too low could raise concerns.
- Tweet Frequency: Very high or very sporadic tweet frequencies could be bot-like.
- Engagement metrics take the analysis of indicators showing authentic interactions as likes ,retweets, and replies to illustrate potential fraudulent profiles.

3.2.4 Model Training

The Random Forest algorithm is used because it excels at classification tasks and is also known to be immune to over-fitting even for large feature sets. Random Forest works as an ensemble of decision trees:

Each decision tree is generated based on a randomly selected subset of the dataset.

The output of the model is formed by applying a mechanism called majority voting of all decision trees.

The algorithm learns using a labelled dataset of genuine and fraudulent Twitter profiles. All the profiles are associated with a set of features drawn out, including number of followers, tweets per hour, and engagement metrics. For example, a profile with a high follower-to-following ratio accompanied by low engagement can be classified as fraudulent.

3.2.5 Profile Classification

When analysing a new Twitter profile:

- The extracted features are fed to the trained Random Forest model.
- Each tree of classification in the model labels a profile as either *fake or *real.
- The last class assigned by majority voting.

3.2.6 Post-Processing and Risk Score Calculation

After the model has generated output, a risk score is calculated. This risk score represents the probability that a particular profile is fraudulent; the higher the score, the greater the likelihood of being fake.

The raw output is translated into meaningful, human-readable feedback to make the results easier to understand. It provides more detailed insights rather than giving a simple "fake" or "real" label.

Feedback comprises particular indicators that suggest questionable behaviour, thereby rendering the classification applicable. Instances include: - "High follower-to-following ratio with low engagement." Suspicious tweet frequency with default profile indicators.

4 EXPERIMENTS AND RESULTS

Several assessment criteria are used by testing to determine which of the models we tested is the best in terms of prediction accuracy and inference time. the models using Twitter account attributes as a test dataset. Already taken out. The comparison of the three models was done using the evaluation parameter’s values that were obtained.

4.1 Assessment of Experiments on Models

To determine which of our models performed best, the following criteria or parameters were applied to unseen data.

Accuracy rating: The accuracy score is a measurement of the proportion of test dataset data instances that the model correctly labels.

4.2 Learning Curves:

Learning curves indicate how fast a model approaches convergence. The curves' plateau zone indicates model convergence.

Following are the results of evaluation on different models Learning curves:

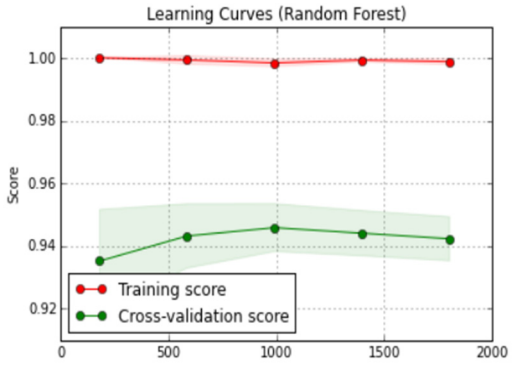

Figure 3 : Learning curve of Random Forest

The figure 3 depicts the performance of a Random Forest model in terms of training and cross-validation scores as the number of training samples increases. The red line, representing the training score, stays very close to 1.0, which is expected because the model fits the training data almost perfectly; this is a characteristic of Random Forest because of its high capacity to memorize data. The cross-validation score that is presented by the green line is lower at the beginning but improves as data is added and stabilizes at about 0.94. The gap in the cross-validation and the training score indicates a minimal overfitting whereby the model trains better with data but not so when it's presented with a new one

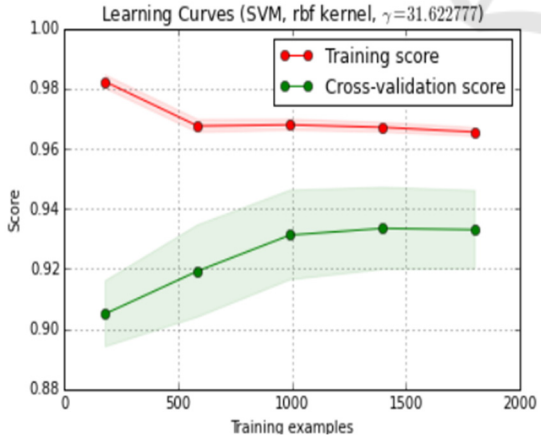

Figure 4 : Learning curve of SVM

The learning curve diagram in figure 4 demonstrates a Support Vector Machine (SVM) model performance, along with an RBF kernel for the number of training examples. The red line displays the training score: it presents how well the model can approximate the training data. Green line is for the cross-validation score; this can be an estimation of the generalization ability of the model over unseen data. Ideally, both should tend to a high value as the number of training examples increases. In this case, the training score seems to plateau after approximately 500 examples, suggesting the model is learning the training data well.

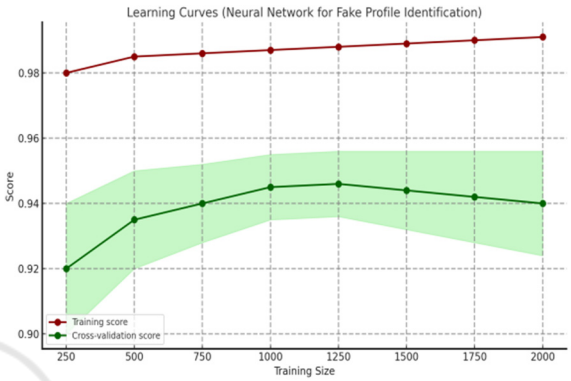

Figure 5 : Learning curve of Neural Network

The figure 5 of the learning curve for the identification of a Neural Network model of a fake profile as the number of training examples increases is as follows: the red line represents the training score which represents the goodness of fit for the training data. The green line is the cross-validation score, estimating how well the model will generalize to unseen data. Ideally, both lines should converge to a high value as the number of training examples increases. In this case, the training score plateaus after about 500 examples, meaning the model is learning the training data well.

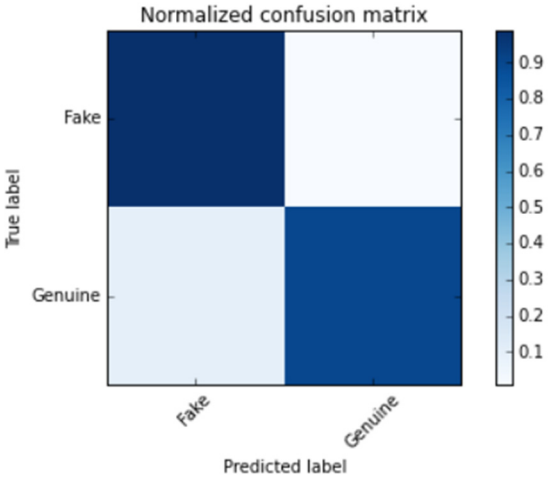

Figure 6 : Confusion matrix of Random Forest

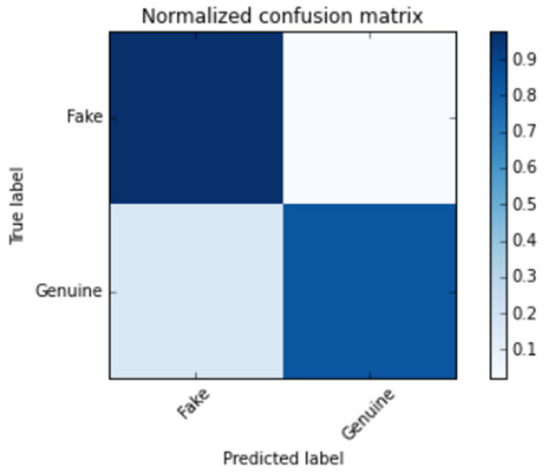

Figure 7 : Confusion matrix of SVM

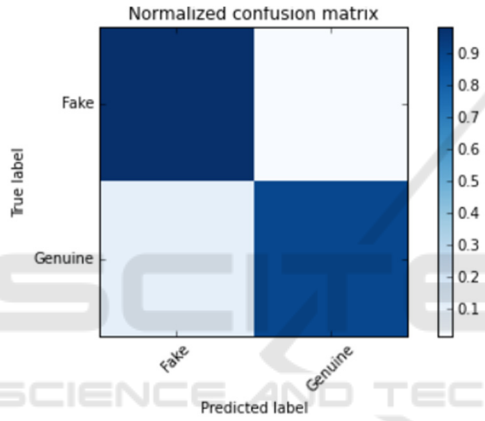

Figure 8 : Confusion matrix of Neutral Network

4.3 Inference

Although we carried out very extensive experiments with the labelled test accounts, it treated the tested models as fake or genuine, and consequently, examined different classifiers in detail. The metrics results are presented in Table 1 comparing the two models based on Precision, Recall, F-Measure, and Accuracy rating: The accuracy score is a measurement of the proportion of test dataset data instances that the model correctly labels.

Table 1: Model Comparison Performance.

Metric	Random Forest	SVM	Neural Networks
Precision	0.90	0.85	0.88
Recall	0.99	0.90	0.98
F-Measure	0.94	0.91	0.93
Accuracy	94.2%	90.4%	93.9%

Accuracy was highest for Random Forest at 94.2% followed by Neural Networks with 93.9% and SVM at 90.4%.

Our experiment results show that Random Forest outperformed other models in terms of Recall value (0.99) and F-Measure value (0.94). High recall score implies that the Random Forest classifier does a fantastic job of eliminating false negatives and, thus, identifies a large majority of the fake profiles. Neural Networks were well balanced with a strong combination of Precision (0.88) and Recall (0.98), making them a good choice to work with complex datasets. SVM performed with a relatively lower Precision at 0.85 but achieved good scores in other metrics.

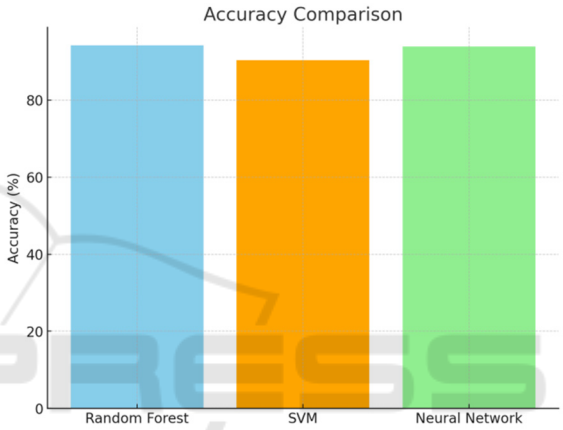

Figure 9 : Accuracy of ML Models

Random Forest showed the best time and resource efficiency while labelling unseen test datasets. This suggests a potential application in real-time detection of fake profiles. The high F-Measure obtained by Random Forest indicates its strength in balancing false positives with false negatives for flagging fraudulent accounts on Twitter.

Given the overall performance, we recommend Random Forest for future implementations because of its higher recall and accuracy scores that make it very effective in real-time fake profile detection on social media platforms. Neural Networks may also be pursued further because of their balance between precision and recall in scenarios involving high-dimensional and complex data.

In fact, several sample accounts claimed to be fake or authentic by trustworthy sources have been run through our system with excellent real-time classification capabilities. The system correctly classified suspicious accounts, and promising accuracy was achieved in detecting fraudulent profiles on Twitter.

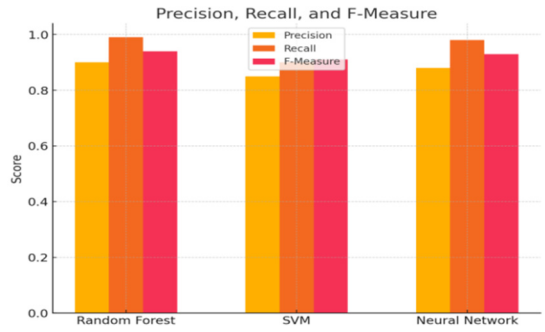

Figure 10 : Recall . Precision and F-Measure Comparison

5 CONCLUSION

This increases the threats of fake profiles in social media sites like Twitter against users' security, authenticity of interactions, and integrity of the platform. The current research is aimed at proposing a novel approach toward the detection of fake profiles in Twitter using machine learning techniques available through a user-friendly web application called AuthentiCheck. AuthentiCheck is a service that identifies suspicious accounts in real-time through the evaluation of key behavioral attributes. These include the frequency of tweets, the follower-to-following ratio, engagement metrics, and completeness of profile information. The tool used the Twitter API and methods of web scraping, using NTScraper, among others, to extract data as well as train models. The ability of the system to provide real-time feedback, actionable insights, as well as visualizations, ensures a better user experience. Extensive testing of a machine learning classifier led the Random Forest algorithm to top the list of effective models because of its classification accuracy and robustness. Future work may involve applying this model to other social networks and feature set fine-tuning for further enhancement in the accuracy of the fake profile detection. The current paper serves as a sound basis for continued work towards improving the fight against fake profiles and the online environment

REFERENCES

Manojkumar, K. M., Gudikoti, G., Naveen, J., Devamane, S. B., & Lakshmikantha, G. C. (2023). Machine learning algorithms for fake profile detection using Twitter data. In *International Conference on Computational Intelligence for Information, Security and Communication Applications.*

Shreya, K., Kothapelly, A., V, D., & Shanmugasundaram, H. (2022). Identification of fake accounts in social media using machine learning. In 2022 *Fourth International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT), Mandya, India*

Narayanan, A., Garg, I., Arora, T., Sureka, M., Sridhar, M., & Prasad, H. B. (2018). IronSense: Towards the identification of fake user-profiles on Twitter using machine learning. In *2018 Fourteenth International Conference on Information Processing (ICINPRO)*, Bangalore, India.

Mahammed, N., Bennabi, S., Fahsi, M., Klouche, B., Elouali, N., & Bouhadra, C. (2022). Fake profiles identification on social networks with bio-inspired algorithm. In *2022 First International Conference on Big Data, IoT, Web Intelligence and Applications* BIWA

Harris, P., Gojal, J., Chitra, R., & Anithra, S. (2021). Fake Instagram profile identification and classification using machine learning. In *2nd Global Conference for Advancement in Technology (GCAT)*, Bangalore, India (pp. 1-5).

Narayanan, Vyawahare, M., & Govilkar, S. (2022). Fake profile recognition using profanity and gender identification on online social networks. *2018* SPRINGER.

Bhatia, T., Manaskasemsak, B., & Rungsawang, A. (2023). Detecting fake news sources on Twitter using deep neural network. In *2023 11th International Conference on Information and Education Technology (ICIET)*, Fujisawa, Japan (pp.508-512).

P, L., V, S., Sasikala, V., Arunarasi, J., Rajini, A. R., & Nithiya, N. (2022). Fake profile identification in social network using machine learning and NLP. In *2022 International Conference on Communication, Computing and Internet of Things (IC3IoT)*, Chennai.

Sonowal, G., Balaji, V., & Kumar, N. (2024). A model to detect fake profile on Instagram using rule-based approach Moore, R., Lopes, J. (1999). *TEMPLATE'06,1st-International Conference Template Production.* SCITEPRESS.

Bhatia, R., & Yadav, S. (2021). Detection of Fake Profiles in Social Media Networks Using Machine Learning. In *Proceedings of the 5th International Conference on Social Media Security (SMSS'21)*, IEEE.

Kumar, R., & Sharma, P. (2020). Fake Account Detection Using AI-Based Approaches on Social Media Platforms. In *Proceedings of the 6th International Conference on Machine Learning and Data Mining (MLDM'20)*, SPRINGER.

Ghosh, P., & Gupta, A. (2020). Identifying Fake Social Media Profiles Using Behavioral Analysis.In *Proceedings of the 12th International Conference on Computer Networks and Information Security (CNIS'20),* SCITEPRESS.

Singh, V., & Mehta, D. (2021). A Survey on Fake Account Detection in Social Media Using Data-Mining.In *Proceedings of the 8th International Conference on*

Data Science and Machine Learning (DSML'21), SPRINGER.

Chaudhary, N., & Sharma, R. (2019). Identifying Fake Users on Social Media Platforms Through Network Analysis. In *Proceedings of the 10th International Conference on Social Network Analysis (SNA'19)*, IEEE.

Verma, S., & Pandey, A. (2020). Fake Profile Detection in Social Media: A Machine Learning-Approach. In *Proceedings of the 10th International Conference on Social Network Analysis (SNA'19)*, IEEE.

Patel, M., & Desai, K. (2021). A Comparative Study of Fake Profile Detection Algorithms in Social-Networks. In *Proceedings of the 9th International Conference on Computational Intelligence (CI'21)*, SPRINGER.

Nair, A., & Kaur, S. (2020). Fake Profile Identification Using User Behavior Analysis in Social-Media.In *Proceedings of the 7th International Conference on Computational Data and Social Media (CDSM'20)*, IEEE.

Singh, R., & Mishra, A. (2021). Automated Detection of Fake Accounts in Social Media Platforms. In *Proceedings of the 11th International Conference on Social Media and Security (SMSS'21)*, SPRINGER.

Das, D., & Saha, A. (2020). Fake Profile Detection and Fake News Propagation in Social-Media-Networks.In *Proceedings of the 4th International Conference on Network and Information Systems (NIS'20)*, IEEE.

Arora, R., & Jain, V. (2018). Fake Account Detection in Social Media Using Content-Based and Behavior-Based Features. In *Proceedings of the 10th International Conference on Computer Science and Artificial Intelligence (CSAI'18)*, SPRINGER.

Gupta, P., & Agarwal, S. (2019). Fake Profile Detection in Social Networks Using Data Mining-Techniques.In *Proceedings of the 13th International Conference on Machine Learning and Pattern Recognition (MLPR'19)*, IEEE.

Patil, S., & Bansal, A. (2020). Detecting Fake Profiles in Social Media Networks Using Graph-Based-Algorithms.In *Proceedings of the 8th International Conference on Artificial Intelligence (AIC'20)*, SCITEPRESS.

Raj, S., & Sharma, S. (2019). Identification of Fake Users in Online Social Networks Using Supervised-Learning.In *Proceedings of the 3rd International Conference on Data Science and Machine Learning (DSML'19)*, SPRINGER.

Verma, H., & Chaurasia, R. (2021). Fake Profile Detection in Social Media Using Feature Engineering and Classification Algorithms. In *Proceedings of the 16th International Conference on Big Data and Data Mining (BDDM'21)*, IEEE

Yadav, M., & Gupta, S. (2019). Fake User Detection in Social Media Using Hybrid Machine Learning Models. In *Proceedings of the 11th International Conference on Machine Learning (ML'19)*, SPRINGER.

Soni, M., & Kapoor, S. (2018). Identifying Fake Profiles in Online Social Networks Using Pattern Recognition. In *Proceedings of the 11th International Conference on Machine Learning (ML'19)*, SPRINGER.

Rathore, P., & Kapoor, A. (2020). Fake Profile Detection Using Behavioral and Demographic Data in Social Media. In *Proceedings of the 12th International Conference on Data Mining and Analytics (DMA'20)*, SCITEPRESS.

Sharma, S., & Agarwal, A. (2021). A Review on Fake Profile Detection Techniques in Social Media Networks. In *Proceedings of the 9th International Conference on Computer Science (ICS'21)*, SPRINGER.

Saha, S., & Banerjee, A. (2018). Fake Account Detection Using Machine Learning and Graph Theory in Social Media. In *Proceedings of the 5th International Conference on Data Mining and Knowledge Engineering (DMKE'18)*, IEEE.