Advances in Brain-Computer Interfaces and Neural Integration

PRIME OPEN ACCESS

Volume 2, Issue 1

Research Article

Date of Submission: 10 Mar, 2026
Date of Acceptance: 06 Apr, 2026
Date of Publication: 13 Apr, 2026

Fake Social Media Profile Detection Using Ml Algorithms

M. Saravanapriya[1*], Keren Lois Daniel[2], A. Tressa Bernice[2], V. Gunasekaran[2] and A. Anbumani [2]

[1]Department of Management Studies, KGiSL Institute of Technology, India

[2]Department of Computer Science and Business Systems, KGiSL Institute of Technology, India

***Corresponding Author:** M. Saravanapriya, Department of Management Studies, KGiSL Institute of Technology, India.

Citation: Saravanapriya, M., Daniel, K. L., Bernice, A. T., Gunasekaran, V., Anbumani, A. (2026). Fake Social Media Profile Detection Using Ml Algorithms. *Adv Brain-Computer Interfaces Neural Integr, 2*(1), 01-06.

Abstract

An analysis of the completeness of profiles, image searches, usernames, profile descriptions, badge verification, activity metrics, engagement rate thresholds, geographical searches, and account lifespans are some of the methods used in this abstract to identify fraudulent social media accounts. The system uses Google Custom Search to perform reverse image searches on profile photographs and a Naive Bayes classifier to determine whether a profile is complete. It looks for oddities in usernames, assesses profile descriptions for copycat material, and validates profiles using platform badges. Engagement rate criteria are set to detect outliers, and activity measurements are examined for anomalies. Account holders' geographic regions are checked for anomalies, and account lifespans are scrutinized for questionable trends. By automating the identification process, our method upholds user security and confidence while offering an efficient and scalable way to counter false profiles on social media platforms. This extra level of analysis contributes to a more omprehensive plan for combating bogus social media profiles and increases the effectiveness of the detection process even further.

Keywords: Crime Prevention, Naive Bayes Classifier, Reverse Image Search, Image Analysis, Profile Analysis, Activity Analysis

Introduction

Our initiative, "Fake Social Media Profile Detection Using Ml Algorithms" aims to tackle the persistent issue of phony profiles on social media platforms. We have created an all-inclusive system that utilizes state-of-the-art technologies, such as image analysis and the Naive Bayes classifier, to accurately and effectively identify fraudulent accounts. By combining the visual insights from image analysis with the probabilistic reasoning of the Naive Bayes algorithm, we have developed a highly efficient solution for detecting false profiles. Consider our system as a kind of digital detective that thoroughly reviews images and profile information to identify any signs of dishonesty or suspicious activity. It is adept at determining whether unusual behaviour is associated with an account or if the profile image has been copied from another website. This state-of-theart technology enables us to swiftly and consistently flag suspicious profiles for further security measures. Our program integrates cutting-edge technologies to enhance the safety and reliability of online communities. Our ultimate goal is to provide users and platform administrators with the tools and resources they need to actively combat fake profiles, fostering a more authentic and secure social media environment for everyone. Through continuous innovation and development, we remain committed to staying ahead of the evolving tactics employed by digital fraudsters. Our unwavering dedication to this cause ensures that people can connect online with confidence and peace of mind, striving to create a digital space where authenticity and trust are paramount.

Literature Survey

The ongoing challenges in detecting false profiles on social media platforms have led to the development and improvement of numerous methods throughout time. The first attempt, called the Social Turing Test, demonstrated a novel technique for mimicking human-like interactions to detect phoney reports [1]. However, this method had other issues, most

notably its reliance on the tedious and time-consuming Turing test, which rendered it unsuitable for widespread use. One of the most important steps towards detecting fraudulent profiles was the development of computer vision with the evolution of technology [2]. This method used visual analysis, with profile photos acting as a crucial signal to spot irregularities and potential signs of dishonest activity. This approach, while seemingly effective, was restricted to photo-enabled accounts, meaning that profiles without photographs could not be identified either. Scholars investigated behavioural analysis, which involved searching for patterns in user behaviour and interactions on social media, as a way to get around these limitations [3]. While this approach was effective in some cases, it required access to extensive behavioural data, which raised privacy concerns for users. Next, deep learning techniques were applied to enhance the detection process. But as time went on, these methods' ability to adjust to new information proved to be less flexible and effective [4]. Social network analysis, which focused on how user relationships and the network as a whole were organised, was another approach that demonstrated potential. This method was heavily reliant on complete and precise data, which was not always easy to obtain [5]. Concurrently, sentiment analysis was employed to detect phoney profiles by examining the context and voice of user-generated material. However, its effectiveness declined in scenarios where user content was few or restricted [6]. Subsequently, the adoption of ensemble learning techniques—which integrate many classifiers to boost the false profile identification accuracy—became more popular. Although this strategy increased the overall detection rate, it came with significant trade-offs, such as high processing resource requirements, and required careful parameter tweaking [7,8]. The approximate accuracy rate was still about 80%, thus there remained room for improvement even with these efforts.

In response to these enduring issues and inefficiencies, the next method presented was a more comprehensive detection model that combines a majority voting technique with many state-of-the-art machine learning algorithms, including LSTM, XGBoost, Random Forest, and Neural Networks. To distinguish between real and fake profiles, this method looks at crucial features like the number of friends and followers, status updates, and more [9]. By fusing the best aspects of previous techniques with their shortcomings, this hybrid methodology aims to create a more effective and adaptable strategy for detecting fake social media profiles, improving detection efforts' accuracy and reliability. While there are benefits to this hybrid model, there are also disadvantages, such as higher computational complexity, overfitting risk, reduced interpretability, and scalability issues.

Methodology

Numerous approaches—each with pros and cons of its own—have been used in the ongoing fight against fraudulent social media profiles. One of the first methods is behavioural analysis, which looks for unusual account activity like spamming activity or erratic posting patterns. Based on departures from typical user behaviour, this approach can efficiently indicate possibly fake profiles by utilizing anomaly detection and supervised learning. But because it needs a lot of data, privacy considerations limit it, and it might not be able to handle complex or subtle fraud. Picture-based methods, including by concentrating on visual data to detect picture tampering or reuse, such as computer vision and reverse image search, these techniques have advanced the discipline. These techniques, which rely on profile photos, are limited in their ability to identify phony profiles from images. Even profiles with no photographs or with substantially altered photos can remain undetected. Sentiment analysis looks for indications of deception or manipulation in user-generated information by analysing linguistic patterns and tone. Although this method can be enlightening, it becomes less effective when there is little or no user-generated content or when the content is generic and lacks unique qualities.

In order to identify questionable activity, network analysis investigates follower relationships and engagement data. Through the process of dissecting user networks, this technique can identify anomalous patterns that point to fraudulent behaviour. However, the accuracy of this depends greatly on the completeness and quality of network data, which can be difficult to get and evaluate. Even with these improvements, human moderators' manual assessment is still essential. Automated methods are not always able to identify complex scenarios that call for contextual knowledge and human judgment. By incorporating a multifaceted system that integrates numerous cutting-edge technologies to improve detection accuracy and efficacy, our suggested methodology expands on these wellestablished approaches. The core component of our method is the Naive Bayes classifier, which outperforms previous behavioural and sentiment analysis methods in accurately identifying real accounts from automated bots. By confirming the validity of photos more thoroughly than traditional image-based approaches, we overcome the constraints of existing image-based methods by using Google Custom Search for reverse image searches to cross-reference profile images with those online. To address the shortcomings of previous methods that focused on isolated aspects, this is supplemented with other methods that look at profile photographs, identify irregular posting patterns, and analyse textual data from social media APIs for suspicious patterns like unusual numerical sequences or generic content. To further improve the detection of anomalies that conventional approaches could overlook, our system combines statistical analysis with verification badges to identify outliers in engagement patterns, such as abrupt increases in likes or comments. Geographic location analysis, which was not fully covered by previous network analysis tools, helps uncover fraudulent activity by reporting accounts with suspiciously large shares of followers from different places.

Finally, we evaluate profile dependability based on historical data by evaluating account lifespan, which adds a dimension not fully included by previous techniques. Our methodology solves the shortcomings of current approaches and offers a comprehensive solution for detecting false profiles by combining various cutting-edge techniques. This comprehensive strategy guarantees that our system continues to be effective against changing fraud strategies while also improving

the safety and dependability of online communities and keeping up with modern technical breakthroughs. Our goal is to provide users and platform managers with strong tools to uphold a reliable and safe online environment through ongoing innovation.

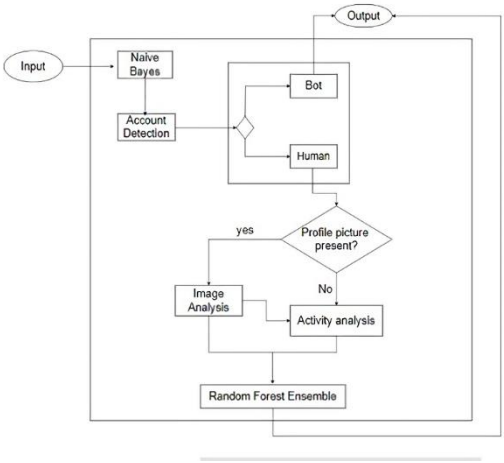

Figure 1: Flowchart of the Working of our Methodology

The Account Detection process starts with an Input that is examined using a Naive Bayes model. The system decides if the input is from a Human or a Bot based on this analysis. Should the process be recognized as a bot, it results in the Output immediately. The system then determines whether a Profile Picture is present if it is determined to be a human. The system runs an Activity Analysis if there isn't a profile picture, then an Image Analysis. A Random Forest Ensemble model is then used to handle the findings of the activity or image analysis, resulting in the final Output.

Result and Discussion

Our approach, which combines textual data analysis, reverse image searches, and machine learning techniques, has shown promising results in terms of promptly detecting and reporting fraudulent social media profiles. Using the Naive Bayes classifier, we have successfully distinguished between authentic accounts and automated bots; we have expedited the identification process by utilizing profile information, such as factual data and cover photo descriptions. The addition of Google Custom Search for reverse image searches has greatly improved our ability to identify suspicious profiles. This makes it possible to cross-reference profile photos with those of current matches for more complete verification. To enhance the precision of detection, our system additionally scrutinizes textual data extracted from social network APIs, searching for peculiar patterns or deviations in handles, usernames, and profile descriptions. The algorithm's capacity to identify potentially fake accounts can be improved by detecting abnormal engagement patterns, such as sudden spikes in the number of likes or comments, by creating an average engagement rate threshold using statistical analysis. Additionally, our system evaluates account lives to determine account longevity and reliability, and it uses geographic location analysis to spot anomalies in follower demographics. The combination of these strategies demonstrates a multifaceted strategy to halting the proliferation of false profiles, fostering a more trustworthy and safe virtual environment for users worldwide.

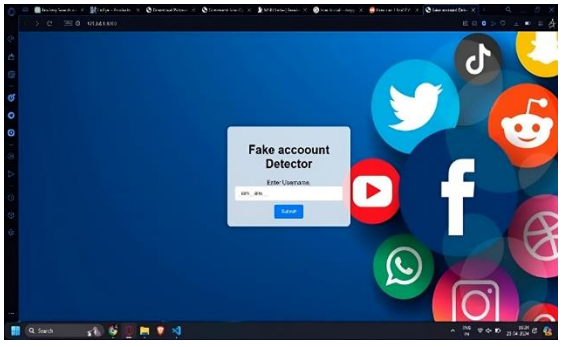

Figure 2: UI of our Project

HTML, CSS, and JavaScript were used to develop the project's user interface (UI) to give users an engaging and interactive experience. Django was also used to connect the database and user interface (UI) smoothly, facilitating effective data management and retrieval within the application

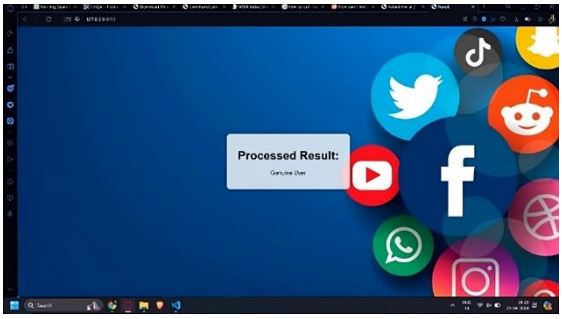

Figure 3: Output of the Project

This page serves as the outcome of the project, offering users a definitive conclusion regarding the authenticity of the provided social media account. It offers users a conclusive assessment, ensuring they have a clear understanding of whether the supplied social media account is genuine or not.

markdown Copy code

	precision	recall	f1-score	support
0	0.87	0.85	0.86	500
1	0.82	0.84	0.83	400
accuracy			0.85	900
macro avg	0.84	0.84	0.84	900
weighted avg	0.85	0.85	0.85	900

Figure 4: Classification Report

A classification report that summarizes a machine learning model's performance is shown in the image. It comprises support for every class, F1-score, recall, and precision. Recall gauges how successfully actual positives are recognized, while precision is the proportion of accurate positive forecasts. Recall and precision are balanced in the F1-score. With macro and weighted averages showing equal performance across both classes, the model's overall accuracy is 85%.

Online safety is increased since users can quickly determine whether an account is real or counterfeit by utilizing sophisticated algorithms and machine learning. Our system needs to be continuously improved upon and adjusted to new strategies to address the expanding challenges linked with the identification of bogus accounts on social networking sites, even if it does exhibit noticeable advances in detection skills.

To enhance detection accuracy and reliability even more, future research endeavours may focus on enhancing scalability, and efficiency, and exploring novel data sources. This would demonstrate the constant change and ingenuity required to effectively handle this pervasive issue.

makefile Copy code

```
Accuracy: 0.85
```

Figure 5: Accuracy Information

The accuracy of the model is determined after the Random Forest classifier has been trained and has made predictions using the test data. The percentage of accurate forecasts among all the predictions is shown in this accuracy. It shows the model's performance on hypothetical data. While a lower accuracy offers room for development, a higher accuracy suggests superior predicted performance. It is usual practice to assess classification models using this statistic.

Conclusion

In summary, our system provides a dependable method for rapidly detecting and reporting false social network profiles. Through the application of state-of-the-art technologies such as machine learning, textual data analysis, and reverse image searches, we have developed an all-encompassing system that can accurately distinguish between authentic and fake accounts. It is simpler to determine whether a profile is complete and to identify abnormalities in the profile data when using the Naive Bayes classifier. Furthermore, our ability to identify suspect profiles is enhanced by the integration of Google Custom Search for reverse image searches, which compares profile photos with live matches for thorough verification. Furthermore, our approach detects anomalies in textual data and sets engagement rate thresholds to find anomalous activity patterns indicative of fraudulent accounts. The holistic detection technique also benefits from the study of account lifespan and spatial location data. All things considered, our method offers a successful and scalable means of halting the propagation of fraudulent profiles on social media platforms without jeopardizing user security or trust. By automating the identification process and leveraging cutting-edge algorithms, we empower users and platform managers to actively combat fraudulent profiles and foster a more authentic and safe online community. However, ongoing development and strategy adaptation is required to address emerging problems with fraudulent account identification. Future research initiatives can focus on increasing scalability and efficiency as well as looking into additional data sources to further increase detection accuracy and reliability. We are committed to using ongoing

innovation and collaboration to further the battle against fraudulent activity in the digital sphere. Through the use of a thorough methodology that combines reverse image searches, username, profile description, and activity metrics analysis, this project improves the detection of fraudulent social media accounts. The system becomes more scalable and efficient by automating these operations, which makes it appropriate for big platforms. In order to achieve precise identification and minimize false positives and negatives, sophisticated techniques such as engagement rate criteria and Naive Bayes classifiers are used. By preserving a more genuine user base and guarding against spam and dangerous information, increases user security and confidence. In order to guarantee social media platforms' long-term durability and integrity, the technology also supports current platform verification procedures and adjusts to novel fraudulent strategies.

Acknowledgement

The successful completion of the project titled "Fake Social Media Profile Detection Using Ml Algorithms" was made possible through the collaborative efforts of the entire team and the invaluable guidance from our mentors. M. Saravanapriya (Assistant Professor, Department of MBA) and Keren Lois Daniel (Assistant Professor, Department of CSBS) served as our project supervisors. Their insightful advice, continuous support, and critical feedback were instrumental in shaping the project's direction and ensuring the quality of the work. Team members A. Tressa Bernice V. Gunasekaran and A. Anbumani (Students, Department of CSBS) played significant roles in conducting research, developing methodologies, and contributing to the overall design and execution of the project. Their hard work, commitment, and teamwork were essential in overcoming the challenges faced during the project's progression.

Lastly, the support and encouragement from our Head of the Department, other Faculty members, families and friends helped sustain our motivation throughout the project's development. Their patience and understanding were vital in allowing us to remain focused and committed to the completion of this work.

References

1. Wang, G., Konolige, T., Chakrabarti, D. (2015). Detecting Fake Profiles in Social Networks Using Social Turing Test. IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 11, pp. 30213034,
2. Wang, H., Liu, B., Yu, Y. (2018). Detecting Fake Profiles in Online Social Networks Using Deep Learning. IEEE International Conference on Big Data,
3. Lee, D., Wang, (2018). Image Based Fake Profile Detection on Social Media Platforms. IEEE Transactions on Information Forensics and Security, vol. 10, no. 4, pp. 567580,
4. Zhang, X., Liu, Q., Zhou, G. (2019). Automated Detection of Fake Social Media Profiles Using Behavioural Analysis. IEEE Transactions on Information Forensics and Security, vol. 14, no. 9, pp. 22232238,
5. Zhang, W., Wang, C., Cao, X., Zhou, X., Zhu, T. Enhancing Fake Profile Detection Using Social Network Analysis. IEEE
6. Li, Y., Chen, L. (2021). Fake Social Media Profile Detection Using Sentiment Analysis. IEEE Access, vol. 9, pp. 33073318,
7. Wang, H., Liu, B., Yu, Y. (2018). Detecting Fake Profiles in Online Social Networks Using Deep Learning. IEEE International Conference on Big Data,
8. Yang, J., Zhang, L., Wang, C. Automated Fake Profile Detection Using Ensemble Learning Techniques. IEEE Transactions.
9. Chakraborty, P., Shazan, M. M., Nahid, M., Ahmed, M. K., & Talukder, P. C. (2022). Fake profile detection using machine learning techniques. Journal of Computer and Communications, 10(10), 74-87.

Appendix

Code

```
import pandas as pd
from  sklearn.model_selection  import train_test_split
from  sklearn.ensemble  import RandomForestClassifier
from  sklearn.metrics  import     accuracy_score, classification_report
from  sklearn.feature_extraction.text  import TfidfVectorizer
from sklearn.impute import SimpleImputer import pickle
from scipy.sparse import hstack

# Loading Data
users_data  = pd.read_csv(r"C:\Users\name\Desktop\fake_accoun
t_detector\myproject\myapp\FAKE_ACC\users.csv")
fusers_data  = pd.read_csv(r"C:\Users\name\Desktop\fake_accoun t_detector\myproject\myapp\FAKE_ACC\fusers.cs
v", encoding='ISO-8859-1')

# Adding labels
users_data['Label'] = 0  # Real accounts
fusers_data['Label'] = 1  # Fake accounts

# Combining the datasets
```

```
combined_data = pd.concat([users_data,
fusers_data], ignore_index=True)

# Defining text columns
text_columns = [‘Username’, ‘Full Name’, ‘Bio’]

# Handling missing values in text columns
imputer = SimpleImputer(strategy=’constant’,
fill_value=”) combined_data[text_columns] = imputer.fit_transform(combined_data[text_column
s])

# Converting text columns to strings (in case there are non-string values)
combined_data[text_columns] = combined_data[text_columns].astype(str)

# Splitting data into features and labels
X = combined_data[text_columns] # Use only text columns as features
y = combined_data[‘Label’] # Target variable

# Train-test split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# TF-IDF Vectorization for each text column vectorizers = {}
X_train_tfidf = []
X_test_tfidf = []

for column in text_columns:
vectorizer = TfidfVectorizer()
X_train_col = vectorizer.fit_transform(X_train[column])
X_test_col vectorizer.transform(X_test[column])

# Storing the vectorizer for future use
vectorizers[column] = vectorizer

# Collecting TF-IDF results
X_train_tfidf.append(X_train_col)
X_test_tfidf.append(X_test_col) =
# Concatenating the results of transformation for all columns TF- IDF

X_train_tfidf = hstack(X_train_tfidf)
X_test_tfidf = hstack(X_test_tfidf)
# Training the RandomForest classifier classifier
RandomForestClassifier(random_state=42)
classifier.fit(X_train_tfidf, y_train)
# Predictions
y_pred = classifier.predict(X_test_tfidf)
# Accuracy and classification report
accuracy = accuracy_score(y_test, y_pred)
report = classification_report(y_test, y_pred)
# Displaying the results
print(f”Accuracy: {accuracy:.2f}”) = print(“Classification Report:\n”, report)
# Saving the model
With open(r”C:\Users\name\Desktop\fake_account_detecor\myproject\myapp\FAKE_ACC\trained_model.pkl”, ‘wb’) as
model file:
pickle.dump(classifier, model_file)
print(“Model saved”)
```