InSight Bulletin: A Multidisciplinary Interlink International Research Journal
Peer Reviewed International, Open Access Journal.
ISSN : 3065-7857 / Website: https://ibrj.us / Volume-3, Issue-2 / February - 2026

Original Article

Identification of Social Media Accounts Using Big Data Analysis on Social Media Platforms

Mohit Shevkar[1], Dr. Disha Ganatra[2]
[1]Assistant Professor, PCU, Sate, Maval
[2]Assistant Professor, Atmiya University

Manuscript ID:
IBMIIRJ -2026-030230

Submitted: 08 Jan. 2026

Revised: 12 Jan. 2026

Accepted: 06 Feb. 2026

Published: 28 Feb. 2026

ISSN: 3065-7857

Volume-3

Issue-2

Pp. 151-155

February 2026

Correspondence Address:
Mohit Shevkar,
Assistant Professor, PCU, Sate, Maval
Email: mohit.shevkar@gmail.com

OPEN ACCESS

Quick Response Code:

Web. https://ibrj.us

DOI: 10.5281/zenodo.21353923

DOI Link:
https://doi.org/10.5281/zenodo.21353923

Abstract

The rapid growth of social media platforms has led to an unprecedented increase in user-generated data, creating significant challenges in accurately identifying and correlating social media accounts across platforms. Traditional detection approaches, which primarily rely on content-based analysis or small-scale heuristic methods, often suffer from limited scalability and reduced effectiveness in handling large and heterogeneous datasets. To address these limitations, this research proposes a big data–driven framework for identifying and linking social media accounts using comprehensive behavioral analysis. The proposed approach leverages multiple data dimensions, including user behavioral patterns, metadata attributes, activity logs, and network relationships, to establish robust identity correlations across platforms. A distributed processing architecture based on Hadoop and Apache Spark is implemented to enable scalable, real-time analysis of massive social media datasets. Machine learning models are developed to learn complex patterns and relationships within the data, facilitating accurate cross-platform identity correlation and detection of suspicious or related accounts. Experimental evaluations demonstrate that the proposed framework significantly improves identification accuracy and scalability compared to traditional content-centric methods. The results highlight the effectiveness of integrating big data analytics with machine learning techniques for large-scale social media account identification. This research contributes a scalable and efficient solution for enhancing trust, security, and integrity in online social networks, making it suitable for real-time deployment in modern social media environments.

Keywords: *Social Media Security, Fake Account Detection, Behavioral Analytics, Big Data Frameworks, Cross-Platform Analysis and Accuracy.*

Introduction

Social media sites have now become a common everyday phenomenon on the digital landscape. Sites such as Facebook, Instagram, LinkedIn, and X (Twitter) have billions of active subscribers. The sites also create a vast amount of heterogeneous data every day. Though the volume raises the level of connectivity and accessibility to these sites on a global scale, it also increases the amount of fake and duplicate profiles that are anonymous and robot-controlled. The effect is on the quality of data and site security. Identity theft, spam distribution, malicious acts coordinated through these robots, and generation of false information can happen. The task of finding authentic users and correlating identities among various social networking platforms has been a challenge. This occurs because many users spread their identities across various accounts, either in terms of their profile details, names, or writing styles. The current state of identification systems mainly relies on content analysis, similarity search, or heuristic-based identification. These models pose limitations in handling massive scales of rapidly growing social networking platforms, especially when the task involves noisy inputs. Also, their accuracies mainly lack scalability to accommodate various platforms for integration.The size, variety, and velocity of the social media data require efficient, big data-focused analytical solutions that are able to process the never-ending data streams in a efficient manner. Traditional data processing paradigms are ineffective in dealing with massive data for real-time processing or for the purpose of correlating identities on a massive scale. On the other hand, distributed computing models, such as Hadoop and Apache Spark, allow data storage and efficient processing for feature derivation or decision-making on massive data.

How to cite this article:
Shevkar, M. & Ganatra, D. (2026). Identification of Social Media Accounts Using Big Data Analysis on Social Media Platforms. InSight Bulletin: A Multidisciplinary Interlink International Research Journal, 3(2), 151–155. https://doi.org/10.5281/zenodo.21353923

Proper identification of the social media accounts requires a broad scope analysis and vision that encompasses areas beyond text. Some of the activities performed on the account, including the pace at which the account posts, the nature of the account that interacts with the account in question, the pace at which the account is active at different times of the day, and the manner in which the account users interact with the account can assist in acquiring information on the authenticity of the account identity. Contemporary advancements in machine learning and deep learning have been shown to have immense potency in addressing the complexities involved in identifying social media accounts.

Methodology:

In this study, a big data-driven approach is used for social media account identification and correlation. There are various stages involved in this approach, namely data acquisition, preprocessing, feature identification, distributed processing, and model-based account identification. These stages are intended for large-scale social media data.

Data Collection

The publicly available data is obtained from various social media platforms using platform-specific APIs as well as web scraping in compliance with the respective policies on usage of the data. The sources of the obtained data may include user profile information, postings, time stamps, interaction records (likes, shares, comments), and social network relations.

Data Processing:

The raw social media data undergoes cleaning and standardization to eliminate inconsistencies, noise, and missing data. The preprocessing tasks performed on the data are removal of duplicates, normalization of text and meta data, synchronization of times, and anonymization of sensitive data. The data is made uniform in the format it takes, irrespective of the social media platform.

Feature Extraction:

The important attributes are gleaned on varied dimensions to efficiently extract the traits of the user identity. The content attributes encompass the text format of the user posts, while the behavior attributes encompass the posting behavior of the user. The temporal attributes portray the user behavior in the timelines, and the network attributes illustrate the user connections on the network.

Distributed Data Processing:

For making the system scalable and manageable for processing a huge amount of data, the proposed system utilizes distributed computing techniques such as Hadoop for storing and processing data in a distributed manner, and it also utilizes Spark for processing in the RAM of computers.

Account Identification and Correlation:

The machine learning techniques are designed for classification and detection of the social media accounts based on the extracted features. Moreover, the graph-based models are also utilized for analyzing the structural information and extract relational patterns among the respective social media accounts. The cross-platform identity correlation mechanism relies on the combination of behavior-based proximity, matching based on features, and network proximity.

Performance Evaluation

The proposed system will be analyzed using the appropriate performance metrics, which include accuracy, precision, recall, F1-score, scalability, and processing time. An experimental evaluation will also be carried out in order to measure the performance of the framework in properly identifying and correlating the social media profiles even in a large-scale environment.

Designed System:

The designed system is developed as a scalable end-to-end data processing solution. For data extraction, social media information is processed through social media APIs along with data connectors. Data preprocessing, extraction, as well as machine learning-based classification/clustering algorithms, are designed for efficient processing of a massive amount of social media data using Apache Spark. For social media account identification, machine learning algorithms along with graph analytics concepts are implemented. Performance evaluation is done through various metrics such as Accuracy, Precision, Recall, as well as Scalability.

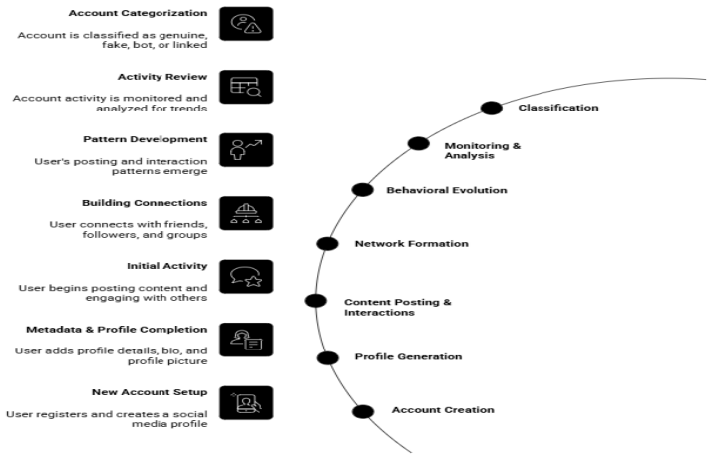

Fig: Life Cycle of Social Media Account

Account Creation: When you make a media account that is the beginning of its life cycle.
This step has parts to it:

- **Registration Information:** When you sign up for an account on the site you have to give them some information aboutyourself. This information is very important. It includes things like your email address, phone number, name and date of birth. The site also looks for things that do not seem right.

For example: if you use an email address that is a problem. It is like waving a flag that says something is not right. If you give them information that is a sure way to raise some suspicions about your Registration Information. The site is looking out for this kind of thing, with your Registration Information so it is best to be honest.

- **IP Address and Geolocation:** When someone makes an account the internet address they use at that time can give us an idea of where they're. So if the internet address changes all of a sudden. If they start using special servers or programs to hide their location it might mean the user is trying to keep their location a secret. The IP Address is, like a map that shows where the user is. If they use a proxy server or a VPN it is like they are trying to cover up their real location, which is the location that the IP Address shows.
- **Account Creation Timestamp:** The time when an account is created can be a clue. If many accounts are created in a time it can look like a bot is making them or that someone is trying to fake things on purpose, with Account Creation Timestamp. This can make us think that the Account Creation Timestamp is not legitimate.
- **CAPTCHA Solving:** The kind of practice followed for solving CAPTCHAs at the time of account creation may also give some clue. Bot operators generally use automated services to solve CAPTCHAs.

Profile & Metainformation Creation: Once the account exists, the typical course of events would begin with the building of a profile and then the connection of the metadata, which would consist of:

- **Profile Picture Analysis:** The profile picture can also be analyzed for assessing whether it is a plagiarized picture, whether it is a stock photo, or whether it is a computer-generated photograph. Reverse image searching can also be used in tracing the origin of photographs.
- **Bio & Description**: The language and formatting employed within the bio & description can be tested for any logical inconsistencies. Typically, a fake profile will include a vague or nonsensical description.
- **Username:** The username can also be checked for some patterns such as alphabetical letters or sequential numbers which are commonly used by bots.
- **Account Age:** This is another factor that should be taken into account. An account that is newly opened is of greater suspicion than an older one.
- **LANGUAGE:** The language of the profile as well as the profile information needs to correspond to the language of the claimed location or interests.

Content Posting and Interactions:

- **Content Frequency:** When you look at the kind of things being posted and how often they are posted, like texts or images or videos you can see a pattern. Bot accounts usually post the Content Type like the same texts or the same images or the same videos or they post links that take you to bad websites that are trying to trick you with spam. The Content Type and Frequency can really help you figure out if it is a Bot account or not.
- **Content Source:** We need to look at where the content comes from to see if it is really original or if it was taken from else. The content might be copied from another place. Even made by a computer program if someone is trying to fake an account. We have to check the content source to make sure the content is genuine and not stolen from someone. The content source is important to figure out if the content's real or if it was made by a robot, for a fake account.
- **Follower Engagement:** The number of likes and comments and shares and retweets that an account gets can tell us how engaged people are with that account and how influence it has. Some accounts have a lot of likes and comments but they might be fake you know, like bots or something.
- **Interaction Patterns:** The people the user talks to on these accounts can tell us what the user is interested in and who they know. If the user mostly talks to accounts that seem fake or bad that is a problem. The Interaction Patterns of the user are important to look at because they can show us the users interests and associations. For example the accounts the user interacts with can give us clues about what the user likes and who the user trusts. So, if the user interacts with suspicious accounts, we should be careful and look into it more because the Interaction Patterns can raise suspicions, about the user.
- **Sentiment Analysis:** Analyzing the sentiment of the posts made on the content as well as the comments made on the account can show the overall tone of the account. Those involving themselves in spreading misinformation or hate speech are usually identified as problematic.

Network Formation: The degree of connectivity the account has with other users is also a very integral part of its social presence.

- **Ratio of Follower to Following**: Ratio of follower to following might be the signal of authenticity of the account. If there exist many followers and very few following, it might be considered an important or so-called influencer account, or if there exist many following and very few followers, it might be entering the follow-unfollow act.
- **Mutual Connections:** The number of mutual connections between the two accounts can help reveal the extent to which these two accounts are related. The greater the number of mutual connections, the less spam the accounts are.
- **Network Density:** Social network density measures in the vicinity of the account might help to determine the extent to

which the account is integrated within the social context. If the account has a high network density, then the chances of the account being genuine would increase.

- **Community Affiliation:** Information about the communities the account belongs to may exist in relation to the account, including interests and affiliations. Participations in the communities of spam and mis-information's could invoke warnings.

Behavioral Patterns Evolution

- **Consistency of Activity:** This is also something that can make a difference. How active an account is, over time. If an account has periods where it's really active and then nothing happens for a while it might be a sign of a bot account. The accounts activity level can change a lot. That is not what you usually see with a real account. The Consistency of Activity is important to look at when you are trying to figure out if an account's real or not.
- **Network Growth:** The speed at which an account grows its network dictates whether or not it is more fake or real. Sudden growth sprees could indicate automated follower buying. • Dynamic Interaction: This may point to an interaction between the account and other users changing. When the pattern does, it perhaps points toward some change or shift in its social behavior.

Monitoring and Analysis

- **Real-Time Monitoring:** One method for monitoring accounts is accomplished through real-time monitoring.
- **Historical Data Analysis**: There can be patterns which are not easily identifiable when viewing historical information.
- **Anomaly Detection:** Anomaly detection on the activities of an account may be employed to determine possibly suspect accounts.
- **Machine Learning Models**: These can be applied to determine the type of accounts depending on the data and behavior that the accounts hold.
- **Category of Account:** Real, Fake, Bot, or Linked Account
- **Genuine Account:** It defines an account which represents a genuine individual and participates in genuine interactions.
- **Fake account:** This is an account that is created through fraudulent or deceptive information.
- **Bot Account:** An account which is controlled through a computer software with the aim of automating actions such as posting content and/or following users.
- **Linked Account:** This account is referred to as being tied to the accounts, either through ownership or joint operations.

Implementation:

The system described is designed as a scalable and end-to-end data processing solution with a focus on social media account discovery and connection at a mass level. Data is obtained from publicly available sources with platform-specific APIs and distributed data connectors. Data is stored in a distributed file system with guaranteed accessibility, fault tolerance, and scalability with increasing data volume. Apache Spark is utilized for data preprocessing as well as feature extraction owing to its capability to process data in memory as well as perform parallel computation work efficiently. The process of feature engineering is done by Spark SQL and Spark MLlib to develop various features such as behavioral features, temporal features, metadata features, as well as network features. Both supervised as well as unsupervised learning models using Spark MLlib identify fake as well as bot-generated accounts.Graph-based analysis using Spark GraphX is used to link accounts across the different platforms. User interactions and relationships are modeled as graphs, and similarity measures combined with network proximity metrics are applied to detect potential matches between accounts of different platforms. The proposed approach has been developed in Python using PySpark, running on a distributed multi-node cluster built on commodity hardware such as multi-core processors, high-memory nodes, and distributed storage. Standard performance metrics such as accuracy, precision, recall, and the F1-score are used for classification, whereas scalability and processing time feature in the measurements. The experiment results prove that this framework has effectively and accurately found the suspicious accounts with strong scalability and efficient processing for a large volume of data. These results are further analyzed in the section Results and Discussion, particularly in how feature selection, model performance, and system scalability influence overall identification accuracy.

Conclusion:

The proposed work provided a scalable big data-driven solution for the identification and linkage of social media profiles on various platforms. The utilization of behavioral patterns, meta-data features, time-dependent activity, and relationships with respect to networks helps in overcoming the drawbacks of content-focused and non-scalable profile identification techniques. The combination of distributed processing engines, Hadoop and Apache Spark, helps in effectively handling big data-driven social web data and making the system suitable for real-time and near real-time processing. Machine learning techniques and graph, based models were implemented to not only detect fake and bot, generated accounts with high accuracy but also enable cross, platform identity correlation. The experimental evaluation showed that the framework put forward by the authors is capable of achieving better accuracy, precision, and recall metrics, in addition to retaining scalability and computational efficiency of a high order as the data volume increases. Such findings corroborate the potential of big data analytics and advanced learning techniques as a winning combination for robust identification of social media accounts. In essence, the system designed by the authors offers a useful and scalable approach to the problem of trust, security, and data integrity in online social networks. Subsequent research will aim at integrating more data modalities, adaptive learning strategies, and privacy, enhancing techniques to not only enhance cross, platform identity correlation but also facilitate the system's deployment in dynamic social media scenarios.

Acknowledgement
We would like to express our sincere gratitude to all those who contributed to the successful completion of this research work titled "Identification of Social Media Accounts Using Big Data Analysis on Social Media Platforms."We are deeply thankful to our institution for providing the necessary academic environment, infrastructure, and technical support required for conducting this research. We extend our heartfelt appreciation to our colleagues and peers for their valuable suggestions, constructive feedback, and encouragement throughout the study.
Financial support and sponsorship
Nil.

Conflicts of interest
The authors declare that there are no conflicts of interest regarding the publication of this paper.

References
1. A. Dharmireddy and M. D. Gottipalli, "Social Networking Sites Fake Profiles Detection Using Machine Learning Techniques," Asian Journal for Convergence in Technology (AJCT), 2023. https://www.asianssr.org/index.php/ajct/article/view/1299
2. G. Ravi Kumar et al., "Use of Machine Learning to Identify Fake Profiles," IJRASET, 2024. https://www.ijraset.com/research-paper/machine-learning-to-identify-fake-profiles
3. A. Agrawal et al., "Detecting Fake Accounts on Social Media Using Machine Learning," Journal of Mobile Computing, Communications & Mobile Networks, 2024. https://journals.stmjournals.com
4. G. Harsha Vardhan et al., "Detecting Fake Social Media Profiles," IJIRSET, vol. 13, no. 11, 2024. https://ijirset.com
5. D. Sharma and N. Singh, "A Review of Deep Learning Approaches for Fake Profile Detection on Social Networking Sites," IJSRSET, 2025. https://ijsrset.com
6. "Detecting Fake Accounts on Instagram Using Machine Learning," Algorithms (MDPI), 2024. https://www.mdpi.com
7. H. Kuruvilla et al., "Spotting Fake Profiles in Social Networks via Keystroke Dynamics," arXiv, 2023. https://arxiv.org/abs/2311.06903
8. J. Wu et al., "BotShape: A Novel Social Bots Detection Approach," arXiv, 2023. https://arxiv.org/abs/2303.10214
9. B. Deepika et al., "Blockchain Enabled Social Network for Detecting Fake Accounts," IJCSIT, 2024. https://ijsrcseit.com
10. "Unmasking Deception: Detection of Fake Profiles in Online Social Ecosystems," Journal of Big Data, Springer, 2025. https://link.springer.com