Grenze International Journal of Engineering and Technology, June Issue

Review of Fake Profile Classification and Identification on Social Networks

Samant Verma[1] and Dr. Shailja Shukla[2]
[1]PhD Research Scholar, Department of CSE, RNTU, Bhopal (m.p), India
Email: samantverma7@gmail.com
[2]Associate Professor, Department of CSE, RNTU, Bhopal (m.p), India
Email: shailja.sharma@aisectuniversity.ac.in

***Abstract*— The rapid proliferation of social media has expanded its use across diverse domains, including commerce, business promotion, political messaging, education, and entertainment. However, this upsurge in social media activity has also attracted malevolent actors who exploit these platforms for illicit activities. This research, as discussed in the paper, is dedicated to discerning the user and societal attributes essential for the identification of fraudulent reviews, misinformation, and rumors disseminated on social media channels. Moreover, the study offers insights into the potential applicability of this research for recognizing illicit user cohorts, analyzing the influence of political ideologies, and assessing the impact of military users aligned with specific ideologies. The paper additionally underscores the significance of intrusion detection mechanisms on social media and introduces a deep neural network-based model for the identification of counterfeit profiles, as well as distinguishing between active and inactive profiles in the realm of social media.**

***Index Terms*— Social Media, Intrusion Detection, Counterfeit Accounts, Disseminated Information, Social Media Mining, Social Media Spam.**

I. INTRODUCTION

This passage highlights how the amount of user- generated data on the internet is increasing as we move towards the Web 2.0 era. It notes that people use social media platforms and commercial websites like Facebook, Twitter, Yahoo, LinkedIn, Amazon, Yatra.com, and others, to express their knowledge and attitudes about election-related and other global-critical problems. Social media is noted to have significantly changed the way individuals communicate and engage in their everyday lives, by providing easy, quick, and trustworthy communication with others. Additionally, it states that the popularity of social media is growing, while user-generated business on the internet is also increasing at an exponential rate in the Web 3.0 era.
With the scalability of the social media platform for marketing, promotion, and campaigning expanding, spammer activity has also expanded significantly. The user's freedom to publish personal or social information on social media encourages spammers to abuse the platform for their own gain, and spam has spread to all kinds of digital communication.
Advertising has become an essential promotional tool for goods and service providers globally. Since the expression of the advertisement has a wide application area, it has many definitions that show similarities and

Grenze ID: 01.GIJET.10.2.187

differences between each other. In its simplest explanation, advertising promotes a good or service in general media in return for a specific price and announces it to large target audiences. Advertising is an economic activity as well as contains many elements. Advertising has gained an important place in the economy due to factors such as its activation with daily life, its effect on purchasing preferences, and shaping consumers' lives. The main actions that ensure the spread of marketing activities to large masses and areas in economic life are, of course, primarily the existence of efforts to support production, distribution and, in parallel, sales. In this respect, advertising has an economic character that accompanies distribution and contributes to show; therefore, it is vital for businesses.

In today's consumer society, businesses, services/products and brands play the leading role. Getting the service or development of a company also means acquiring a status. Advertising is the most effective method to use in gaining this status. Therefore, it is necessary to communicate constantly while carrying out these advertising activities. Having constant communication means staying in the customer's mind for a longer time. Therefore, the best way to keep the customer's sense is to provide advertising communication.

In summary, businesses and individuals are increasingly using social media evaluations to make decisions about purchasing, product design, and elections. Consumer reviews are considered to be more trustworthy than other types of reviews. However, opinion spammers post false or incorrect opinions, and some consumers post fake positive or negative evaluations for economic or personal reasons. Negative reviews can harm a company's reputation and reduce the perceived value of their products/services, while positive reviews can help build a company's reputation and boost sales.

II. Social Media

In the realm of online electronic communication, social media encompasses platforms that facilitate user connections and the formation of virtual communities dedicated to the exchange of information, novel ideas, messages, and various personal and public content. Websites designed to enable user interactions, information sharing, and the reception of news updates are classified as social media platforms, examples of which include Facebook, Twitter, YouTube, and many others.

Social media platforms empower individuals to express their viewpoints, share photos and videos, and engage with friends and fellow users. The convenience offered by these platforms has led to a widespread adoption among internet users, serving as an effortless avenue for disseminating information to a broader audience.

The era of social media has witnessed a surge in user participation across diverse domains, ranging from commercial activities, marketing, and political propaganda to educational initiatives and entertainment. The exponential growth of social media usage has been instrumental not only in the educational sector but also in the domains of marketing and promotion. However, this surge in activity has also attracted individuals who engage in negative behaviors, such as the creation of counterfeit profiles aimed at tarnishing the reputation of businesses.

In the field of social network analysis, individuals are represented as nodes in a graph, and their relationships are depicted as edges. Community detection within these networks typically involves graph analysis, where various methodologies are applied to partition the network into distinct groups. These methodologies often emphasize defining the structural attributes that distinguish one group from another, ensuring that communities exhibit higher connection density internally compared to externally, or clustering nodes with proximity based on specific similarity criteria.

Classical data clustering algorithms are often employed to tackle these challenges. Nevertheless, many data clustering issues are computationally intensive and fall within the category of NP-hard problems. As a response, simplified algorithms have been developed to provide approximate solutions. For the ensemble detection problem, techniques like graph segmentation, spectral clustering, divisive algorithms such as Girvan Newman, modularity-based optimization algorithms, and statistics-based methodologies are frequently employed.

The identification of malicious activities draws upon both user-specific and social features to identify nefarious communal content. This methodology is valuable for discerning fraudulent reviews, counterfeit news, and the propagation of baseless rumors within social media platforms. Furthermore, it serves as a tool for influencing passively engaged malicious users based on their ideological alignment, predicting criminal activities, detecting criminal user groups, and assessing the influence of political ideologies within the context of election campaigns.

III. Social Media Mining

Social media mining is the process of extracting valuable insights from vast quantities of social data, with a particular focus on identifying unlawful or suspicious activities. This field has established the foundational principles and algorithms for efficiently

analyzing extensive social media datasets, drawing from a multidisciplinary approach that encompasses computer science, data mining, machine learning, social network analysis, network science, ethnography, statistics, optimization, and mathematics [4].
The primary objective of social media mining is the discovery of fresh and beneficial knowledge derived from social network data. It is important to note that social media data often includes sensitive user information, which necessitates anonymization before publication or sharing with data mining researchers. This step is crucial to safeguard the privacy of social media users.

IV. SOCIAL MEDIA SPAM

In recent years, the ubiquitous problem of spam has extended its reach into all facets of digital communication. The remarkable growth of user bases on popular social media platforms like Facebook, Twitter, YouTube, and others has inevitably attracted the attention of spammers. The openness of these platforms, where users can freely share personal and social content, provides an enticing environment for spammers to exploit for their own gain. In the early days of the digital era, spammers primarily targeted email and web search engines, prompting information scientists to actively address these issues. In more recent times, researchers have directed their efforts toward mitigating email and search engine spam, as well as developing techniques for detecting social media spam, as depicted in Figure 1 [5].
Given the prevalence and potential hazards associated with social network spammers, the research community has been proactive in devising methods for their detection. To identify social network spammers, researchers employ a range of approaches, including supervised, unsupervised, and graph algorithms. One example involves the use of the support vector machine and parameter J to model the prior distribution of spammers' proportions. The adjustment of parameter J seeks to strike a balance between accuracy and recall improvement while incorporating gain. Additionally, the chi-square test is utilized to identify distinctive characteristics [6].

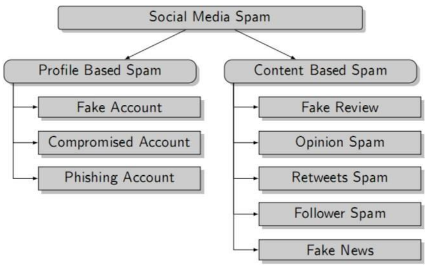

Figure 1: Social Media User Features

Another technique involves constructing comparable groups based on text MD5 similarity and URL clustering, followed by the assessment of each group's spammer status using distribution coverage and temporal burst analysis [7]. Furthermore, a solution for determining the ideal weight is achieved through spectral decomposition, aiding in the identification of spammers on platforms like Twitter [8]. Unsupervised algorithms depend on reasonable similarity indicators, while supervised algorithms rely on distinguishing feature indicators. Spammers often employ external features to evade detection, though this approach is quite effective. However, the vast scale and sparsity of graph data present challenges in terms of computational complexity and efficiency.

V. REVIEW SPAM (OPINION SPAM)

In the contemporary digital landscape, social media reviews have assumed an increasingly central role for both individuals and organizations in shaping their decisions concerning purchases, product development, and even electoral choices. Of all review types, consumer reviews are widely regarded for their credibility. Recent research revealed that a substantial 68% of consumers routinely consult reviews before finalizing their purchase decisions [9]. However, within this context, the issue of opinion spam, involving the propagation of deceptive or inaccurate reviews, has gained prominence, often driven by various motivations, including financial incentives [9,10].
(a) Negative reviews, for example, can tarnish the reputation of a rival entity and erode the perceived value of its offerings [11].
(b) Conversely, positive reviews have the potential to enhance the standing of a business and function as effective promotional tools for their products and services [11].

In recent times, the practice of opinion spamming has evolved into a lucrative industry. Notably, approximately 71% of users engaging in retweeting activities have been identified as automated bots [12-13]. In this landscape, advertising agencies have taken to compensating individuals for producing fraudulent reviews or encouraging their clients to craft biased or paid assessments. These deceptive practices are frequently facilitated through a range of mechanisms aimed at promoting counterfeit reviews.

VI. RELATED WORK

Previous research addressing the community recognition problem predominantly leans towards link analysis methodologies [3-6]. Communities are generally defined as groups of entities with shared specific characteristics, and the choice of similarity metrics is contingent on the network's unique context. Traditional methods frequently leverage the network's topological attributes when delineating communities. The key objective underpinning these methods is the optimal partitioning of the network into distinct communities, and to evaluate their efficacy, the modularity parameter is commonly employed [7]. To enhance modularity, various optimization-based algorithms have been developed.

An alternative approach in community detection involves classifying individuals based on their shared interests as inferred from content analysis [8-11]. Users active on social networks frequently exchange diverse and meaningful data, encompassing media shares, blogs, and emails. The presumption is that individuals sharing similar content likely share common interests. Community detection can be realized through clustering methods utilizing attributes extracted from these shared contents. It's worth noting that when assessing both approaches, methods centered on the network's topological features tend to overlook user interests, whereas those emphasizing content analysis may disregard the network's structural aspects [12]. To address this issue, two-stage community detection approaches are emerging [13]. These approaches initially break down the network based on content analysis and subsequently employ link analysis within each delineated community.

The Kernighan-Lin algorithm [5] stands as one of the pioneering graph segmentation techniques, which identifies communities by randomly dividing the graph into two equivalent parts and iteratively swapping node pairs between these divided sections. The spectral bisection approach, relying on Fiedler vector coordinates, is another frequently utilized method, facilitating boundary identification. Discriminative algorithms, like the Girvan-Newman algorithm, center on splitting communities based on middle link values (edge betweenness). Given that ensemble detection in complex networks presents an NP-hard problem, a multitude of heuristic optimization algorithms has emerged. These approaches adopt the modularity function as an objective to optimize the network's community partitioning. For instance, a simulated backgammon approach supports global optimization by redistributing nodes across communities to increase modularity. Additionally, heuristic algorithms such as the tabu search algorithm are applied to address the community detection problem. Another frequently used approach is the modularity optimization algorithm, which leverages modularity as a stopping criterion in a greedy hierarchical clustering framework [7].

An additional approach to community detection involves clustering individuals based on the analysis of content within the network. Topic models have been developed to summarize individuals' interests and group those with shared topics into the same community. The well-known Latent Dirichlet Allocation (LDA) model is an example. These topic models have been adapted for social networks. Models such as CUT, CART, and CT are used to cluster network members with common interests. However, it's important to note that these topic models often overlook the topological network structure, resulting in lower modularity values.

In this comprehensive review, we've summarized 11 publications from the period between 2015 and 2020 in Table 1. The table includes six columns, detailing the study's references, core tasks, approaches, algorithms, data collection names, and the research gaps identified during the study.

TABLE I: SUMMARY OF RELATED WORK

s.no	**Task**	**Approach**	**Algo*b***	**FET*d***	**ResearchGap**
[1]	Deceptivespam review detection	SentenceWNN	CNN	POS	ReviewWeight
[2]	Network-based Spam Detection	Rank-Basedlabeling	GraphSMM	Net Spam	Information diffusion
[3]	LP based Spam Detection	Semi Supervisedlearning	LM Algo	POS, LIWC	Associated Multimedia Review

[4]	Temporal Spam Detection	Outlier Detection	ForestAlgo	Stream Buffer	Honest Recommendation
[5]	Spam Detectionvia review deviation	UnSupervised Learning	Deep Learning	Content analysis	Semantic Review Deviations
[6]	Incentivized Reviews Detection	Word ofmouth	ExChange Theor	Maximum likelihood	ReviewQuality
[7]	Fake reviews Fltering	Content classification	SVM	N-Gram	Incorporation of reviewerprofile
[8]	Fake Review Classifcat ion	MachineLearning	Ada Boost	TF-IDF	Control Proliferation
[9]	Probabilistic Review Graph Classification	Multi- Modal Framework	Neural Networks	N-gram	Node Spamicity computation
[10]	Intelligentreview spam detection	Outlierfactor model	HTTP TopicSearch	POS	DynamicOutlier factor
[11]	Intelligentreview spam detection	Multi- iterative graph-basedmodel	Feature Fusion , Machine learning	BOW	Influencespam activities

VI. PROPOSED SOLUTION

Identifying and classifying fake profiles on social media involves considering a variety of features and employing algorithms to analyze patterns. Here are some common features and characteristics that can be used for this purpose:

1. Profile Completeness:

- Presence of a profile picture
- Filled-out profile information (bio, location, education, job, etc.)
- Verified contact details

2. Activity Patterns:

- Posting frequency and consistency
- Interaction patterns (likes, comments, shares)
- Time of activity (e.g., consistent posting times or unusual patterns)

3. Unusual Behavior:

- Abnormal liking or commenting patterns
- Excessive posting or activity in a short period
- Rapid increase in friends/followers

4. Network Analysis:

- Analysis of friends or followers to identify suspicious patterns
- Presence of known fake accounts in the network

5. Content Quality:

- Coherent and meaningful posts
- Original content vs. repetitive or copied content
- Multimedia content quality (e.g., image resolution)

6. Language and Writing Style:

- Consistency in language and writing style across posts
- Use of proper grammar and language conventions

7. Image Analysis:

- Reverse image search to detect stock or stolen images
- Consistency in profile pictures and other images

8. Account Age and Creation Time:

- Recent account creation with sudden activity spikes
- Number of posts over time since account creation

9. Privacy Settings:

- Level of privacy settings (public, private, restricted)
- Consistency in privacy settings with user behavior

10. Geolocation:

- Consistency in location information with other profile details
- Unusual or inconsistent locations for the user's claimed residence

11. Machine Learning Models:

- Training models on labeled datasets to classify fake profiles
- Features extracted from user behavior and profile details for model training

12. Bot Detection Techniques:

- Analyzing patterns indicative of automated behavior (e.g., bots)
- Detection of scripted interactions or posting patterns

13. API Usage Patterns:

- Monitoring API usage patterns to identify automated or scripted activity
- Analyzing the rate and frequency of API calls

14. External Verification:

- Cross-referencing profile details with external sources to verify authenticity
- Verifying employment, education, or other claimed information

15. Social Graph Analysis:

- Analyzing the structure and connectivity of the social graph to identify anomalies
- Identifying clusters of fake profiles or unusual network patterns

It's important to note that the effectiveness of these features may vary, and combining multiple features often leads to more accurate detection. Social media platforms continually evolve their strategies for identifying fake profiles, and detection algorithms need to be updated regularly to stay ahead of emerging trends in online deception.

This work endeavors to provide a user-centric explanation of deep neural network technology, with the primary objective of classifying information as either harmful, fake, or genuine. It also seeks to establish the link between user feature similarity and the number of connections in order to identify "sock puppet" accounts on social media platforms. The proposed solution involves the development of a framework built upon deep neural networks to distinguish between malicious, deceptive, and authentic users on social media platforms. Standard parameters will be employed to evaluate and compare the framework's performance. Additionally, this work aims to detect and eliminate inactive profiles on social media platforms to free up valuable space. The deep neural network-based framework will be employed to differentiate between active and inactive users, and any idle profiles identified will be removed. Furthermore, efforts will be made to create a framework that restricts new profiles from posting until they are verified.

To incorporate a friend verification process alongside email verification, the following steps could be implemented:

1. Design a registration page requiring new users to provide their email address, password, and the email addresses of at least one friend who will vouch for their account.
2. Dispatch a verification email to the email addresses furnished by the new user. This email should contain a unique link that friends must click to validate the new user's account.
3. Once a friend activates the link, the new user's account is marked as verified, granting them access.
4. Employ middleware to scrutinize the user's verification status with every post request. If the user is unverified, they will be prevented from posting and directed to a page explaining the necessity of account verification.
5. Establish a user-friendly admin interface for verifying or unverifying a user's account.

6. Consider implementing a limit on the number of friends who can verify a user's account to forestall misuse of the system.
7. Maintain comprehensive logs of registration and verification activities for analysis and troubleshooting.
8. Rigorously test the framework to ensure it functions as intended.

This approach elevates security and assists in deterring the creation of fake profiles. It is worth noting that this process may introduce some inconvenience for legitimate users who may not have friends available or willing to verify their accounts.

VII. Conclusion

Detecting anomalies on social media is a demanding and intellectually stimulating endeavor. The prevailing method in literature involves labeling end users as malicious based on their profile and user-generated data. Nonetheless, the limited incorporation of graphical and linguistic components results in reduced diversity and accuracy. Conventional techniques often disregard language and opinion-related features, which hinder the identification of user perspectives and ideologies on various subjects, such as products, politics, global and local issues, and more. To foster an efficient model for classifying anomalies on social media platforms, a framework rooted in social theory is introduced, integrating user graphical and linguistic characteristics that contribute to the shaping of user interactions on social media platforms. This paper proposes the development of a deep neural network-based model to identify sock puppet accounts across social media. The culmination of this work results in a solution that enhances scalability within large networks and curtails malicious behavior across social media platforms.

References

[1] Y. Chen and S. F. Wu, "FakeBuster: A Robust Fake Account Detection by Activity Analysis," *2018 9th International Symposium on Parallel Architectures, Algorithms and Programming (PAAP)*, 2018, pp. 108-110, doi: 10.1109/PAAP.2018.00026.

[2] A. D. Frunze and A. A. Frolov, "Methods for Detecting Fake Accounts on the Social Network VK," *2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus)*, 2021, pp. 342-346, doi: 10.1109/ElConRus51938.2021.9396670.

[3] H. Alvari and P. Shakarian, "Hawkes Process for Understanding the Influence of Pathogenic Social Media Accounts," *2019 2nd International Conference on Data Intelligence and Security (ICDIS)*, 2019, pp. 36-42, doi: 10.1109/ICDIS.2019.00013.

[4] P. Sowmya and M. Chatterjee, "Detection of Fake and Clone accounts in Twitter using Classification and Distance Measure Algorithms," *2020 International Conference on Communication and Signal Processing (ICCSP)*, 2020, pp. 0067-0070, doi: 10.1109/ICCSP48568.2020.9182353.

[5] L. Caruccio, D. Desiato, and G. Polese, "Fake Account Identification in Social Networks," *2018 IEEE International Conference on Big Data (Big Data)*, 2018, pp. 5078-5085, doi: 10.1109/BigData.2018.8622011.

[6] J. Hodson and B. Traynor, "Design Exploration of Fake News: A Transdisciplinary Methodological Approach to Understanding Content Sharing and Trust on Social Media," *2018 IEEE International Professional Communication Conference (ProComm)*, 2018, pp. 1-5, doi: 10.1109/ProComm.2018.00008.

[7] K. Zarei, R. Farahbakhsh, N. Crespi, and G. Tyson, "Impersonation on Social Media: A Deep Neural Approach to Identify Ingenuine Content," *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2020, pp. 11-15, doi: 10.1109/ASONAM49781.2020.9381437.

[8] G. Shetty, A. Nair, P. Vishwanath, and A. Stuti, "Sentiment Analysis and Classification on Twitter Spam Account Dataset," *2020 Advanced Computing and Communication Technologies for High Performance Applications (ACCTHPA)*, 2020, pp. 111-114, doi: 10.1109/ACCTHPA49271.2020.9213206.

[9] F. C. Akyon and M. Esat Kalfaoglu, "Instagram Fake and Automated Account Detection," *2019 Innovations in Intelligent Systems and Applications Conference (ASYU)*, 2019, pp. 1-7, doi: 10.1109/ASYU48272.2019.8946437.

[10] Z. Shahbazi and Y. -C. Byun, "Fake Media Detection Based on Natural Language Processing and Blockchain Approaches," *in IEEE Access*, vol. 9, pp. 128442-128453, 2021, doi: 10.1109/ACCESS.2021.3112607.

[11] N. Singh, T. Sharma, A. Thakral, and T. Choudhury, "Detection of Fake Profile in Online Social Networks Using Machine Learning," *2018 International Conference on Advances in Computing and Communication Engineering (ICACCE)*, 2018, pp. 231-234, doi: 10.1109/ICACCE.2018.8441713.

[12] S. Khaled, N. El-Tazi, and H. M. O. Mokhtar, "Detecting Fake Accounts on Social Big Data (Big Data)," *2018, pp. 3672-3681, doi: 10.1109/BigData.2018.8621913.

[13] S. Mazepa, S. Banakh, A. Melnyk, S. Pugach, O. Yavorska, and N. Golota, "An Ontological Approach to Detecting Fake News in Online Media," *2021 11th International Conference on Advanced Computer Information Technologies (ACIT)*, 2021, pp. 531-535, doi: 10.1109/ACIT52158.2021.9548394.

[14] E. Van Der Walt and J. Eloff, "Using Machine Learning to Detect Fake Identities: Bots vs Humans," *in IEEE Access*, vol. 6, pp. 6540-6549, 2018, doi: 10.1109/ACCESS.2018.2796018.
[15] N. C. Lê, M. -T. Dao, H. -L. Nguyen, T. -N. Nguyen, and H. Vu, "An Application of Random Walk on Fake Account Detection Problem: A Hybrid Approach," *2020 RIVF International Conference on Computing and Communication Technologies (RIVF)*, 2020, pp. 1-6, doi: 10.1109/RIVF48685.2020.9140749.
[16] D. S. Alorini, D. B. Rawat, and G. S. Alorini, "On the Influence Blocking Maximization for Minimizing the Spreading of Fake information in Social Media," *2020 Spring