Fake Profiles Identification in Online Social Networks using Machine Learning and NLP

[1]Geetha C Megharaj, [2]Premalatha D [3]Sachin S, [4]Sandhya S, [5]Vishnu Varshith R
Department of CSE, Dr. T. Thimmaiah Institute of Technology
KGF, India

Abstract: Online social networks (OSNs) now form an intrinsic part of daily life for millions, providing effortless communication over distances and timelines. With this, however, also exists the increasing threat of security issues caused by false profiles, which take advantage of user trust and invade personal information. It is imperative that accurate detection systems for such false profiles be used to ensure the security and integrity of these sites. Traditional classification methods have, however, been proven to be ineffective and less adaptable for the purpose. We research, in this study, the use of Machine Learning (ML) and Natural Language Processing (NLP) techniques for increasing the detection rate for false profiles. We study specifically the use of Support Vector Machines and Naive Bayes for classifying user accounts as real or false, thus enhancing the overall reliability of OSN environments.

Keywords: Fake Account Identification, learning datasets, training model Natural Language Processing, Prediction, Social Network

I. INTRODUCTION

Lakhs of people utilize online social platforms which have become a major part of digital life. They spend a lot of time interacting on different platforms. From interaction-centric services such as Facebook and Instagram to information-sharing platforms like Twitter and Google Buzz, OSNs cater to a huge range of user needs. However, as these platforms continue to grow, so do concerns surrounding user privacy and security. One of the most pressing challenges is the proliferation of fake profiles, which can be applied in identifying the theft, impersonation, and various forms of online abuse.

Users frequently divulge critical personal details while interacting with OSNs. Malicious actors may take use of this information whether it is provided openly or only within specific networks. One of the main effects of fraudulent accounts, identity theft, has long been a problem that affects millions of people worldwide.

Victims could experience monetary losses, harm to their reputation, or even legal repercussions. Sadly, a lot of OSNs have weak identity verification procedures and frequently have privacy settings with little protection, which makes users more vulnerable.

OSNs have inadvertently become fertile ground for identity theft, impersonation, and other cyber threats. While users are typically required to provide accurate information when registering, weak privacy settings and poor monitoring make it easier for attackers to create and exploit fake profiles. These activities range from social engineering attacks and online defamation to manipulative marketing and political propaganda.

User profiles on social media can generally be categorized into static and dynamic data. Static data includes demographic details and stated interests, typically provided during account creation. In contrast, dynamic data captures user behavior over time, such as interactions, content engagement, and location patterns. Many research efforts rely on both data types to detect fake profiles and suspicious activity. However, not all platforms make dynamic data accessible, limiting the effectiveness of certain detection techniques.

A huge range of method have been proposed in identifying the fake profiles and filter malicious content in OSNs, each offering its own advantages and limitations. Among the persistent challenges are issues like online harassment, privacy invasion, and content misuse—many of which are linked to fake accounts. These profiles often operate under false identities and engage in activities that undermine the user experience and platform security.

Despite the implementation of security measures like the Facebook Immune System (FIS), platforms like Facebook are still not accurate in identifying and blocking fraudulent profiles. This disparity emphasizes the necessity of more sophisticated, data-driven strategies. In this regard, combining Natural Language Processing (NLP) with Machine Learning (ML) presents exciting prospects for boosting overall OSN security and detecting fraudulent profiles.

The integration of machine learning and natural language processing has been demonstrated significant promise in recent years for identifying irregularities in user behavior and content. It is possible to flag accounts displaying suspicious characteristics by looking at posting behavior, metadata, and language trends. ML and NLP models are able to adjust and learn from changing threats, in contrast to old or traditional methods that mostly impact in manual evaluation or static feature sets.

II. RELATED WORK

In recent years the research efforts have been explored to various methodologies for detecting fake profiles and malicious behaviour in online platforms, often leveraging advancements in Natural Language Processing (NLP), user interaction analysis, and data mining techniques. Chai et al. presented a foundational study focusing on the integration of natural language dialog systems within user interfaces. Although their prototype utilized relatively basic NLP and human-computer interaction methods, user testing revealed a clear preference—particularly among novice users—for conversational interfaces over traditional menu-driven systems.

The study emphasized that in dynamic environments like ecommerce, effective dialog management holds more value than processing complex language structures. Furthermore, the authors proposed combining dialog-based navigation with menu-driven systems to provide more intuitive access to information. This hybrid approach may offer useful insights into designing adaptive interfaces for social platforms, especially when dealing with varying user behavior patterns.

In another study focused on professional networking platforms, researchers addressed the unique challenges of detecting fake profiles on LinkedIn. Due to strict privacy settings, LinkedIn exposes only limited public data, which makes traditional behaviour-based detection approaches less effective. To overcome this,Ms.Shalinda Adikari and Mr.Kaushik Dutta proposed a model that identifies a minimal but effective set of publicly available profile attributes crucial for distinguishing fake profiles. They also recommended tailored data mining and learning methods suitable for handling the constraints of LinkedIn's restricted data access. Their work highlights the necessity of platform-specific detection strategies, especially for networks where behavioural data is not fully accessible.

Z. Halim et al. introduced a novel approach that applies spatiotemporal mining to detect users involved in coordinated malicious activities. By leveraging Latent Semantic Analysis (LSA), the researchers were able to compare clusters formed through spatio-temporal co-occurrence with actual user networks. Their findings showed a strong correlation between the generated clusters and existing social ties, indicating the potential of LSA in uncovering hidden patterns of malicious behavior.

The quality and comprehensiveness of the selected feature set, however, the feature has a major impact in how well this strategy performs. While a broader feature set greatly improves detection accuracy, a smaller feature set may make it difficult for the system to identify dangerous content. This emphasizes how crucial careful feature selection is when creating models for detecting phony accounts.

Moreover, graph-based detection models have gained traction in recent years. These models treat users as nodes and their interactions (such as friendships, follows, or message exchanges) as edges, enabling the application of social network analysis techniques. Metrics like clustering coefficient, betweenness centrality, and community detection are used to distinguish genuine users from anomalies. Fake profiles often exhibit unusual patterns, such as high friend-request rates, low reciprocity, or connections to multiple suspicious accounts. While powerful, these graph-based methods require real-time access to network topology and may face challenges in largescale implementations due to computational complexity. Nonetheless, while been combined along with machine learning classifiers, they offer a compelling toolset for enhancing the accuracy of fake account identifying systems.

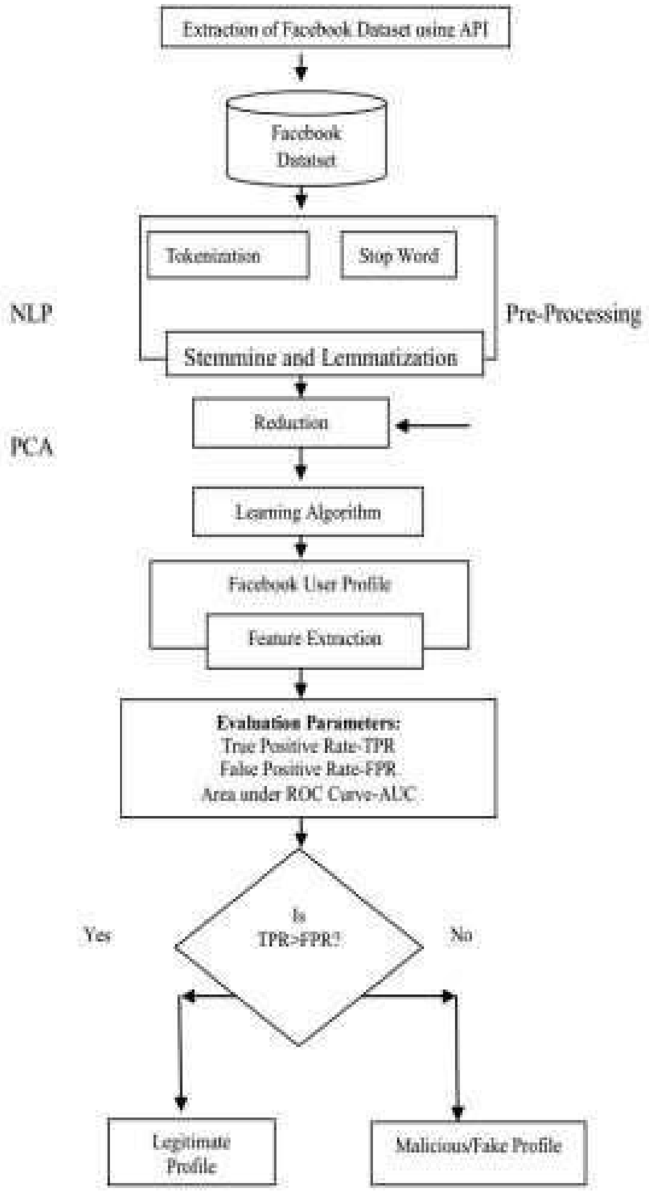
III. PROPOSED METHOD

A. Proposed System Architecture

We built a smart system to spot fake accounts on social media. It uses computer learning (ML) and language understanding (NLP) to look at profile information and how people act online..

This Model focuses on identifying fake Facebook profiles by leveraging a structured pipeline composed of three main phases:

(1) NLP Pre-processing,
(2) Principal Component Analysis
(3) Machine Learning Algorithms.



. **Fig 1. Working Structure of the Application**

1. NLP Pre-Processing

Natural Language Processing pre-processing is an important feature in refining and preparing raw textual data for future analysis. This stage aims to reduce noise, optimize storage, and enhance classification performance. A key benefit is the reduction in index size, as stop words—common terms such as “and” “the,” or “is”—can constitute up to 30% of a document's word count but add little semantic value. Similarly, stemming—reducing words to simple forms—can shrink the indexing size by up to 50%.

The pre-processing pipeline includes several sub-steps:

1. Tokenization

The first crucial stage is tokenization, which divides an ongoing text stream into more manageable, significant chunks called tokens. These tokens, which stand in for words, phrases, or other important textual components, are the starting point for further Natural Language Processing (NLP) operations like feature extraction, classification, and parsing. By taking into consideration special characters like punctuation marks, hyphens, and brackets that could otherwise skew the text's structure, tokenization guarantees accurate and consistent text segmentation. As a result, the analysis that follows is easier to handle and more significant.

2. Stop Word Removal

Stop word removal follows tokenization and plays a significant role in refining the dataset. Stop words are common but generally non-informative words such as “and” “the,” or “is,” which do not provide valuable insight when classifying text data. Since these words appear frequently across many texts, they add little to the differentiation between genuine and fake profiles. By eliminating stop words, the model reduces noise, improving both processing efficiency and the effectiveness of classification algorithms. This step helps the system focus on more meaningful words that contribute to distinguishing between legitimate and fraudulent content.

3. Stemming and Lemmatization

Stemming and lemmatization are employed to reduce words to their base or root forms. Both processes aim to standardize the text, ensuring that different word forms of the same root are treated equivalently. Stemming is a heuristic approach that removes suffixes from words, often resulting in approximations of the base form. For example, “running” might be reduced to “run.” Lemmatization, in the other phase, it is been an more important method that uses linguistic rules and a dictionary to return the correct base form of a word, such as converting “better” to its lemma “good.” While stemming can be quicker, lemmatization is more precise, and together, these techniques ensure that the context is normalized, reducing variability and improving the performance and feature evalutation for the machine learning models.

2. Principal Component Analysis (PCA)

To minimize the dataset's complexity maintaining its most informative features, PCA is used. High-dimensional data is been formatted into new coordinate system using this statistical method, in such cases the directions of largest variance are represented by the axes (principal components). In addition to simplifying the feature space and increasing the efficiency of learning algorithms, PCA aids in visualizing feature associations and locating underlying patterns that differentiate authentic profiles from fraudulent ones.

3. Learning Algorithms

To classify user profiles accurately, the system integrates two supervised learning algorithms: **Support Vector Machine** and **Naive Bayes**.

Support Vector Machine

SVM is able to find the simple datasets that separates datas from different classes with the maximum margin. In the data of the fake account detection, the SVM identifies boundary conditions that best distinguish genuine users from deceptive ones. The closest data points to the separating hyperplane, known as support vectors, play a critical role in defining this boundary.

Naïve Bayes

This probabilistic classifier operates under the assumption of feature independence. Despite its simplicity, it is highly effective for high-dimensional data such as user metadata and textual content. Naïve Bayes calculates the probability that a profile belongs to a similar class based on the likelihood of observed features, such as timing, content language, or geographic information. Even when features are interdependent, the classifier performs remarkably well in detecting patterns typical of fake accounts.

IV. RESULTS

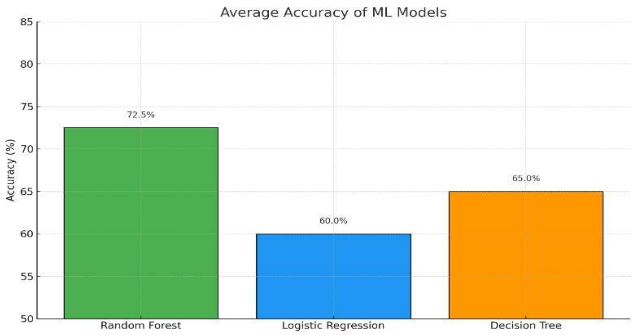


Fig 2: Three machine learning models—Random Forest, Logistic Regression, and Decision Tree—were evaluated for their average classification accuracy in identifying fraudulent profiles on social networks. The models make use of features that have been extracted using NLP and metadata analysis methods. The best accuracy was obtained by Random Forest, demonstrating its resilience in managing diverse data and minimizing overfitting.

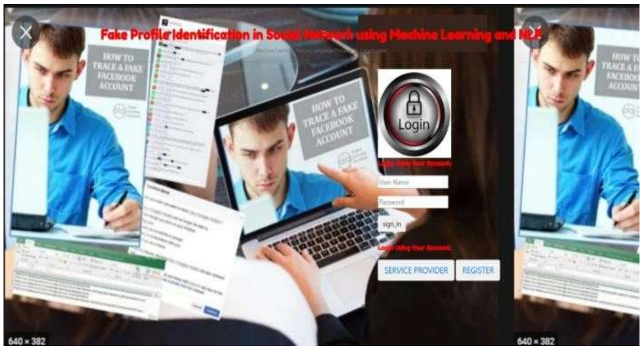
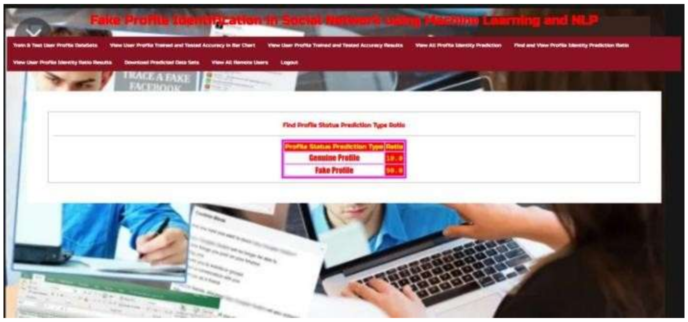

Fake Profile Identification in Social Network using Machine Learning and NLP
HOW TO TRACE A FAKE FACEBOOK ACCOUNT
Login
User Name
Password
sign_in
SERVICE PROVIDER
REGISTER
640 × 382
640 × 382

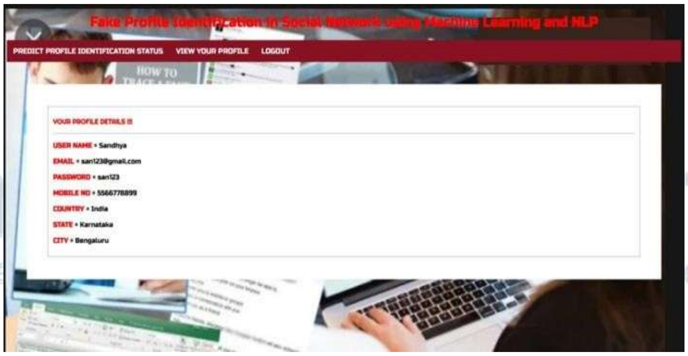
Fig 3: The suggested fake profile identification system's user interface. To identify phony profiles on social networks, the system uses machine learning and natural language processing algorithms. It supports real-time interaction and analysis and has login modules for both users and service providers.


Fake Profile Identification in Social Network using Machine Learning and NLP
Train & Test User Profile DataSets
View User Profile Trained and Tested Accuracy in Bar Chart
View User Profile Trained and Tested Accuracy Results
View All Profile Identity Prediction
Find and View Profile Identity Prediction Ratio
View User Profile Identity Ratio Results
Download Predicted Data Sets
View All Remote Users
Logout

Find Profile Status Prediction Type Ratio

Profile Status Prediction Type	Ratio
Genuine Profile	10.0
Fake Profile	90.0

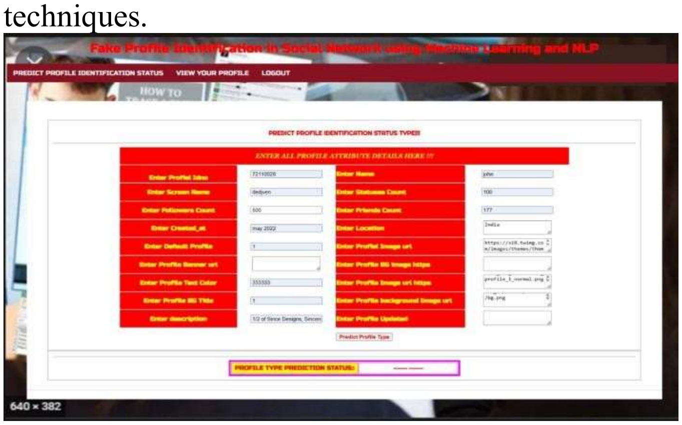
Fig 4: Using machine learning and natural language processing, the fake profile identification system for social networks uses an interface for user profile details. In order to enable individualized profile verification and identification, the dashboard shows important user data such as username, email, password, cellphone number, and location information.


Fake Profile Identification in Social Network using Machine Learning and NLP
PREDICT PROFILE IDENTIFICATION STATUS
VIEW YOUR PROFILE
LOGOUT

YOUR PROFILE DETAILS !!!

USER NAME • Sandhya
EMAIL • san123@gmail.com
PASSWORD • san123
MOBILE NO • 5566778899
COUNTRY • India
STATE • Karnataka
CITY • Bengaluru

Fig 5: Interface for entering profile information in the system for identifying fraudulent profiles. User-inputted parameters include screen name, follower count, and profile creation date. The system makes predictions about the authenticity of the social network profile by using machine learning and natural language processing techniques.


Fake Profile Identification in Social Network using Machine Learning and NLP
PREDICT PROFILE IDENTIFICATION STATUS
VIEW YOUR PROFILE
LOGOUT
Predict Profile Type
640 × 382

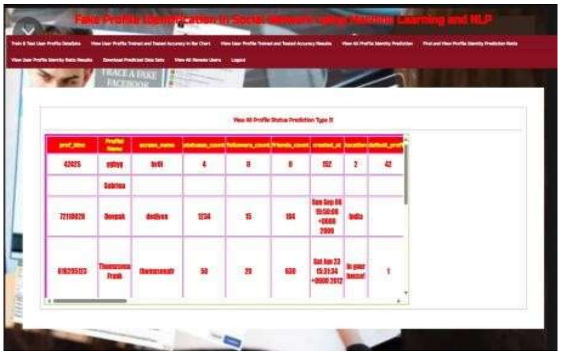
PREDICT PROFILE IDENTIFICATION STATUS TYPE!!!

ENTER ALL PROFILE ATTRIBUTE DETAILS HERE !!!

Enter Profiel Idno	72110026	Enter Name	john
Enter Screen Name	dedyen	Enter Statuses Count	100
Enter Followers Count	600	Enter Friends Count	177
Enter Created_at	may 2022	Enter Location	India
Enter Default Profile	1	Enter Profiel Image url	https://s18.twimg.com/images/themes/them
Enter Profile Banner url		Enter Profile BG Image https	
Enter Profile Text Color	333333	Enter Profile Image url https	profile_1_normal.png
Enter Profile BG Tile	1	Enter Profile background Image url	/bg.png
Enter description	1/2 of Since Designs, Sincere	Enter Profile Updated	

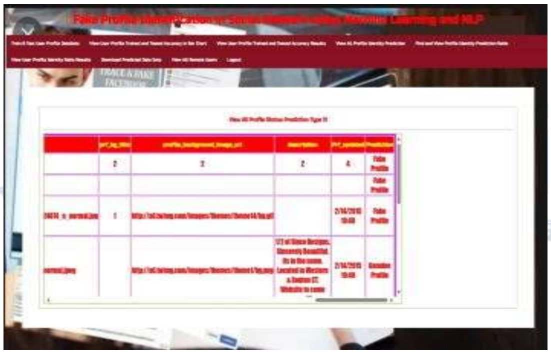
PROFILE TYPE PREDICTION STATUS: ---- ----

Fig 6: A graphic representation of the profile status prediction ratio that shows the results of classifying real and fraudulent social network profiles. The machine learning and natural language processing (NLP)-based algorithm can detect unusual patterns of activity that are usually linked to fraudulent accounts, predicting 90% of profiles as phony and 10% as real.


Fake Profile Identification in Social Network using Machine Learning and NLP
Train & Test User Profile DataSets
View User Profile Trained and Tested Accuracy in Bar Chart
View User Profile Trained and Tested Accuracy Results
View All Profile Identity Prediction
Find and View Profile Identity Prediction Ratio
View User Profile Identity Ratio Results
Download Predicted Data Sets
View All Remote Users
Logout
TRACE A FAKE FACEBOOK

View All Profile Status Prediction Type II

prof_id	Profile Name	screen_name	statuses_count	followers_count	friends_count	created_at	location	default_prof
42025	[illegible]	[illegible]	4	0	0	152	2	42
	Sabrina							
72110028	Deepak	[illegible]	1234	15	184	Sun Sep 06 15:50:08 +0000 2009	India	
618295123	Thomasena Frank	[illegible]	50	29	638	Sat Jun 23 15:31:34 +0000 2012	In your house!	1


Fake Profile Identification in Social Network using Machine Learning and NLP
Train & Test User Profile DataSets
View User Profile Trained and Tested Accuracy in Bar Chart
View User Profile Trained and Tested Accuracy Results
View All Profile Identity Prediction
Find and View Profile Identity Prediction Ratio
View User Profile Identity Ratio Results
Download Predicted Data Sets
View All Remote Users
Logout
TRACE A FAKE FACEBOOK

View All Profile Status Prediction Type III

	[illegible]	profile_background_image_url	description	[illegible]	Prediction
	2	2	2	4	Fake Profile
					Fake Profile
[illegible]_normal.jpg	1	[illegible]		2/14/2015 10:40	Fake Profile
normal.jpeg		[illegible]	[illegible]	2/14/2015 10:40	Genuine Profile

Fig 7: The fake profile identification system uses machine learning and natural language processing to predict user profile details. Analyzed attributes such screen name, status count, followers, and background image metadata are shown in the interface. Based on learnt patterns and profile traits, the system assesses each profile's validity and assigns a "Fake Profile" or "Genuine Profile" classification.

V. CONCLUSION

This study explored a way to automatically find fake accounts on social network like Facebook. We combined machine learning (algorithms that learn from data) with natural language processing (which helps computers understand text). We cleaned and prepared text from Facebook profiles using NLP techniques. Then, we have used two machine learning models, SVM and Naïve Bayes, to classify profiles as real or fake. The results shows that using the combination of methods improved how accurately we could identify fake profiles, which can help make online platforms safer.

REFERENCES

[1] M. Fire, F. Günther, and S. Fritsch, ―Strangers intrusion detection—Detecting spammers and fake profiles in social networks based on topology anomalies,‖ Human Journal, vol. 1, no. 1, pp. 26–39, 2012.

[2] S. Kannan and V. Gurusamy, ―Preprocessing techniques for text mining,‖ 2015. [Online]. Available: Published on March 5, 2015.

[3] S. Adikari and K. Dutta, ―Identifying fake profiles in LinkedIn,‖ in PACIS 2014 Proceedings, AISeL, 2014.

[4] [4] Z. Halim, M. Gul, N. ul Hassan, R. Baig, S. Rehman, and F. Naz, ―Malicious users' circle detection in social network based on spatiotemporal co-occurrence,‖ in Proc. Int. Conf. on Computer Networks and Information Technology (ICCNIT), July 2011, pp. 35–390.

[5] Y. Liu, K. Gummadi, B. Krishnamurthy, and A. Mislove, ―Analyzing Facebook privacy settings: User expectations vs. reality,‖ in Proc. 2011 ACM SIGCOMM Conf. Internet Measurement Conf., ACM, 2011, pp. 61–70.

[6] S. Mahmood and Y. Desmedt, ―Poster: Preliminary analysis of Google's privacy,‖ in Proc. 18th ACM Conf. on Computer and Communications Security, ACM, 2011, pp. 809–812.

[7] T. Stein, E. Chen, and K. Mangla, ―Facebook immune system,‖ in Proc. 4th Workshop on Social Network Systems, ACM, 2011.

[8] S. Abu-Nimeh, T. M. Chen, and O. Alzubi, ―Malicious and spam posts in online social networks,‖ Computer, vol. 44, no. 9, pp. 23–28, 2011.

[9] J. Jiang, C. Wilson, X. Wang, P. Huang, W. Sha, Y. Dai, and B. Zhao, ―Understanding latent interactions in online social networks,‖ in Proc. 10th ACM SIGCOMM Conf. on Internet Measurement, ACM, 2010, pp. 369–382.

[10] P. Kazienko and K. Musial, ―Social capital in online social networks,‖ in Knowledge-Based Intelligent Information and Engineering Systems, Springer, 2006.