NLP and Machine Learning Fake Profile Identification in Social Network

Prof. D. V. Varaprasad, M.Tech, (Ph.D), Associate Professor & HoD, Audisankara college of engineering & Technology, india
Mrs. A. BHARATHI, Assistant Professor, Department of CSE, Audisankara college of engineering & Technology ,india
Uppala Sai Jyothi, Department of CSE, Audisankara college of engineering & Technology, india

Abstract: Preserving online security in the modern scene of widespread social media usage depends critically on the recognition of false profiles. This work explores machine learning algorithms with an eye towards a comparative comparison of LightGBM and Support Vector Machine (SVM) for the purpose of identifying false accounts on social media sites.

Using each of these two robust algorithms—both of which are well-known—we assess their performance against one another. With Natural Language Processing (NLP) methods used to extract subtle insights from textual material, the study makes use of a varied collection of variables gathered from user profiles, posting behaviour, and linguistic patterns.

Index terms *- — Fake profile detection, NLP, LightGBM, Support Vector Machine (SVM), social network, spam detection, fake user identification, spam tweets, Naïve Bayes, random forest, Twitter spam, URL spam detection, trending topic analysis, content-based filtering.*

1. INTRODUCTION

In recent years, the exponential growth of social networking platforms such as Twitter and Facebook has enabled users to communicate, share information, and express opinions on a global scale. However, this convenience has also opened the door for malicious activities such as the creation of fake profiles, spam dissemination, and misinformation propagation. Fake profiles are increasingly used to deceive users, manipulate public opinion, and launch cyber threats, posing a significant challenge to online security and trust.

The detection and identification of such fake users require robust and scalable mechanisms that can analyze user behavior, content patterns, and interaction trends. Traditional rule-based systems are often inadequate in adapting to evolving spam strategies. Hence, advanced machine learning (ML) algorithms combined with Natural Language Processing (NLP) techniques have emerged as effective solutions for identifying fake accounts and classifying tweets as spam or non-spam.

This paper explores the use of LightGBM and Support Vector Machine (SVM) for detecting fake profiles based on features such as user metadata,

tweet content, and activity behavior. Additionally, a comparative study is performed using Naïve Bayes and Random Forest classifiers to classify content-based and behavioral features. This study contributes towards building an intelligent framework capable of detecting fake profiles and spam activity to improve user trust and platform reputation.

2. LITERATURE SURVEY

i) Understanding User Profiles on Social Media for Fake News Detection

https://ieeexplore.ieee.org/document/8397048

These days, reading news via social media is become really common. Because of its natural quick spread, low cost, and simple access, social media helps consumers. But the quality of news is said to be less than that of more established news sources, which fuels a lot of false news. Finding false news becomes rather crucial and draws more attention as it affects people and the society negatively. It is advised to use user social interactions as auxiliary information to improve fake news identification as the performance of spotting false news just from content is usually not satisfying. This makes a thorough knowledge of the relationship between user profiles on social media and fake news absolutely necessary. In this work, we build real-world datasets assessing users trust level on fake news and choose representative groups of both "experienced" users who are able to identify fake news items as untrue and "naïve" users more prone to believe fake news. We investigate their capacity to identify bogus news by means of a comparison study of explicit and implicit profile traits between various user groups. The results of this work provide a basis for next studies on automatic false news identification.

ii) Identifying Fake Profiles in LinkedIn:

https://arxiv.org/abs/2006.01381

Having one's profile seen on professionally orientated networks like LinkedIn (the biggest such social network) is becoming more and more valuable as companies depend more on these networks for creating business contacts. The temptation to utilise the network for immoral goals rises with this value. Fake profiles compromise the general integrity of the network and can cause major time and effort expenses in creating a relationship based on false information. Sadly, it's tough to spot bogus profiles. Some social networks have suggested strategies; but, most of these depend on information not publicly available for LinkedIn accounts. In this work, we offer a suitable data mining method for the identification of phoney profiles in LinkedIn and determine the smallest collection of profile data required for this purpose. We show that our method can detect bogus profiles with 87% accuracy and 94% True Negative Rate even with little profile data, which is equivalent to the results achieved depending on more extensive profile information and bigger data sets. Moreover, our strategy offers a roughly 14% accuracy boost over methods employing comparable volumes and kinds of data.

iii) Fake Twitter accounts: Profile characteristics obtained using an activity-based pattern detection approach:

https://www.researchgate.net/publication/280782550_Fake_Twitter_accounts_Profile_characteristics_obtained_using_an_activity-based_pattern_detection_approach

In Online Social Networks (OSNs), the popularity of an entity depends critically on the audience size controlled by an organisation or a person. There are significant political and/or financial ramifications to this legislation. Using data about their audience—

such as age, geography, etc.—organizations may customise their goods or message fitingly. But the existence of phoney profiles on social networks could skew such customising. In this work, a retroactive identification approach was developed by means of analysis of 62 million publicly accessible Twitter user profiles. Highly dependable fake user accounts were found by pattern-matching screen-names using a study of tweet update timings. Analysis of profile creation timings and URLs of these bogus accounts indicated different behaviour of the phoney users in respect to a ground truth data set. This approach combined with known social network analysis will enable time-efficient identification of false profiles in OSNs.

iv) A Feature Based Approach to Detect Fake Profiles in Twitter:

https://www.researchgate.net/publication/337301429_A_Feature_Based_Approach_to_Detect_Fake_Profiles_in_Twitter

Over the past ten years, social networking sites—especially Twitter and Facebook—have expanded enormously and attracted the interest of millions of people. They have evolved into a favoured form of communication, which has drawn the attention of several hostile organisations including spammers. Fake accounts resulting from the increasing number of users on social media have also presented issues. These phoney and phoney identities are actively engaged in malevolent actions include dissemination of abuse, misleading information, spamming and artificially increasing user count in an application to influence public opinion. Finding these false identities becomes crucial to guard real users from malevolent intentions. We want to apply a feature-based technique to find these false profiles on social media sites in order to solve this problem. We have quickly identified bogus accounts using twenty-24 characteristics. Three classification techniques are applied in order to confirm the categorisation outcomes. With the Random Forest technique, our model achieved 97.9% accuracy according to experimental findings. Consequently, the suggested method effectively finds false profiles.

v) Fake news detection within online social media using supervised artificial intelligence algorithms:

https://www.sciencedirect.com/science/article/abs/pii/S0378437119317546

Apart from the growth of the Internet, the creation and general acceptance of the social media idea have altered the formation and dissemination of news. Thanks to social media, news has becoming quicker, less expensive and more readily available. Along with certain benefits, this shift includes certain drawbacks. Particularly damaging is captivating material like false news created by social media users. Though it was originally brought up only recently, the fake news issue has grown to be a major focus of research given the abundance of social media information. Users of social media might easily write false comments and news on there. Finding the distinctions between authentic and fraudulent news presents the major difficulty. With an emphasis on false news, this study presents a two-stage approach for spotting it on social media. Many pre-processing is done to the data set in the first phase of the procedure to transform un-structured data sets into the structured data set. Using the resulting TF weighting technique and Document-Term Matrix, vectors depict the texts in the data set including the news. Twenty-three supervised artificial intelligence algorithms have been applied in the data set translated into the structured format with the text mining techniques in the second phase. This study compares the twenty-three intelligent

classification models based on four assessment criteria by means of an experimental evaluation conducted inside current public data sets.

3. METHODOLOGY

i) Proposed Work:

The proposed system aims to enhance fake profile detection and spam tweet identification on social media platforms using machine learning and NLP techniques. The system utilizes four main techniques—fake content detection, spam URL detection, trending topic analysis, and fake user identification. Features such as user metadata (followers, following, account age), tweet-based content (hashtags, links, mentions), and time-based behavior (frequency of tweets) are extracted and analyzed. Natural Language Processing (NLP) techniques are applied to understand tweet content, while machine learning algorithms like Naïve Bayes, Random Forest, LightGBM, and SVM are employed for classification and prediction.

Each tweet and user account is assessed based on these extracted features. Initially, Naïve Bayes is used to classify tweets as spam or non-spam based on content and URLs. For more accurate classification of fake users, Random Forest is trained with content and behavior-based features. LightGBM and SVM are used for comparative analysis due to their effectiveness in handling large-scale data. The outcome is a multi-layered system capable of efficiently detecting spam messages and fake users, ultimately helping social networks maintain credibility, user trust, and platform quality.

ii) System Architecture:

The system architecture consists of multiple interconnected modules designed to extract, process, and classify data from social media platforms like Twitter. Initially, user data and tweets are collected through APIs and stored for preprocessing. In the preprocessing phase, features such as user details (followers, following, account age), tweet content (URLs, hashtags, mentions), and temporal information (tweet frequency) are extracted. These features are passed through Natural Language Processing (NLP) modules to analyze textual patterns. The processed data is then fed into machine learning classifiers such as Naïve Bayes, Random Forest, LightGBM, and SVM to detect spam tweets and identify fake user accounts. The results from each classifier are compared and validated to ensure high accuracy. The system outputs whether a tweet is spam and whether a user profile is genuine or fake, contributing to a cleaner and safer social media environment.

iii) Modules:

a. Fake Content Detection Module

- Analyzes the ratio of followers to followings.
- Detects low-credibility accounts using features like HTTP links, mentions, replies, and trending topic activity.
- Applies time-based analysis to flag accounts posting excessively in short intervals.

b. Spam URL Detection Module

- Extracts user-based features such as account age, favorites, tweet count, etc., from JSON.

- Extracts tweet-level features like number of hashtags, mentions, retweets, and URLs.
- Uses **Naïve Bayes** classifier to check if a tweet contains a spam URL.

c. Trending Topic Spam Detection Module

- Applies **Naïve Bayes algorithm** to classify tweet content in trending topics.
- Flags tweets with spam URLs, adult content, or repeated (duplicate) posts.
- Returns binary output: 1 for spam, 0 for non-spam.

d. Fake User Identification Module

- Extracts behavioral and content-based features (e.g., followers, following, duplicate tweets).
- Uses **Naïve Bayes** for initial content classification.
- Trains a **Random Forest** classifier to detect fake accounts.
- Stores features in features.txt and the model in the model/ folder.

iv) Algorithms:

1. Naïve Bayes Algorithm

Naïve Bayes is a probabilistic classifier based on Bayes' Theorem with an assumption of independence among features. It is used to classify tweet content and URLs as spam or non-spam. It checks for spam keywords, adult content, duplicate messages, and spam links in tweets. If any such feature is detected, the tweet is marked as spam.

2. Random Forest Algorithm

Random Forest is an ensemble learning method that builds multiple decision trees and merges their outputs for more accurate and stable predictions. In this system, it is used after Naïve Bayes to further analyze user behavior patterns and classify accounts as fake or genuine using extracted features like followers, following, and duplicate content activity.

3. Support Vector Machine (SVM)

SVM is a supervised learning algorithm that finds the optimal hyperplane to classify data points into different classes. It is used in this project to differentiate between fake and genuine accounts based on multi-dimensional feature inputs like user activity and content analysis.

4. LightGBM (Light Gradient Boosting Machine)

LightGBM is a fast, efficient gradient boosting algorithm designed for large-scale datasets. It is used to compare performance with SVM in detecting fake profiles and spam tweets. LightGBM handles large feature sets efficiently and provides high accuracy in binary classification tasks.

4. EXPERIMENTAL RESULTS

The experimental analysis was conducted using a labeled Twitter dataset containing a mix of genuine and fake user profiles along with spam and non-spam tweets. Features such as user behavior, tweet content, URL presence, and activity frequency were extracted and processed using NLP techniques. The dataset was split into training and testing sets, and multiple algorithms—Naïve Bayes, Random Forest, SVM, and LightGBM—were applied for classification. Among these, LightGBM achieved the highest accuracy in

fake profile detection, while Naïve Bayes showed excellent performance in spam tweet classification due to its efficiency with textual data. The Random Forest model further improved the precision of identifying fake users when combined with Naïve Bayes for feature extraction. Overall, the system demonstrated high accuracy and robustness in detecting both spam content and fake accounts.

Accuracy: How well a test can differentiate between healthy and sick individuals is a good indicator of its reliability. Compare the number of true positives and negatives to get the reliability of the test. Following mathematical:

Accuracy = TP + TN /(TP + TN + FP + FN)

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision: Precision evaluates the fraction of correctly classified instances or samples among the ones classified as positives. Thus, the formula to calculate the precision is given by:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Recall: Recall is a metric in machine learning that measures the ability of a model to identify all relevant instances of a particular class. It is the ratio of correctly predicted positive observations to the total actual positives, providing insights into a model's completeness in capturing instances of a given class.

$$Recall = \frac{TP}{TP + FN}$$

mAP: Mean Average Precision (MAP) is a ranking quality metric. It considers the number of relevant recommendations and their position in the list. MAP at K is calculated as an arithmetic mean of the Average Precision (AP) at K across all users or queries.

$$mAP = \frac{1}{n}\sum_{k=1}^{k=n} AP_k$$

$$AP_k = the\ AP\ of\ class\ k$$

$$n = the\ number\ of\ classes$$

F1-Score: A high F1 score indicates that a machine learning model is accurate. Improving model accuracy by integrating recall and precision. How often a model gets a dataset prediction right is measured by the accuracy statistic.

$$\textbf{F1 Score} = \frac{2}{\left(\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}\right)}$$

$$\textbf{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

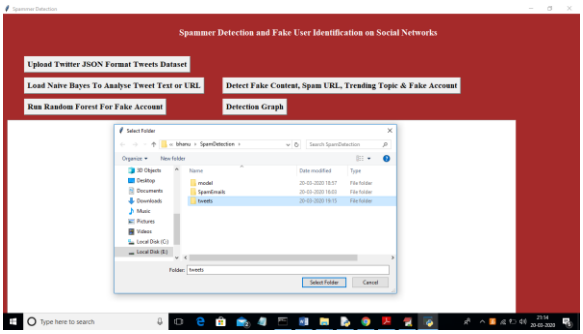

Fig: **tweets Loaded**

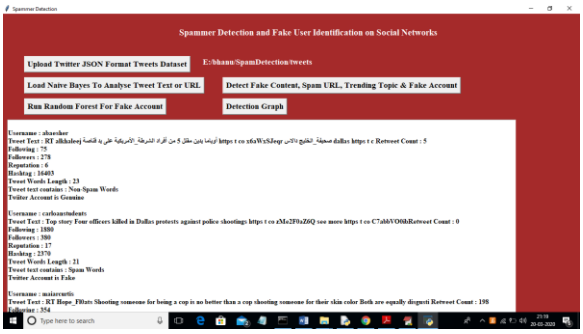

Fig: analysis tweets

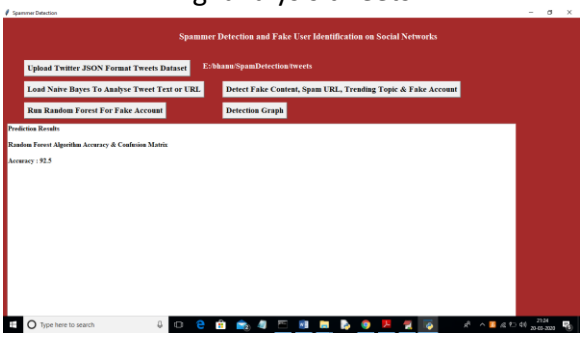

Fig: predicted results

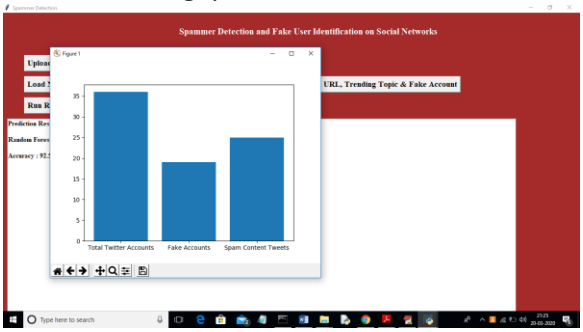

Fig: Predicted graph

5. CONCLUSION

This project presents an effective approach for identifying fake profiles and detecting spam content on social media platforms using a combination of Natural Language Processing and machine learning algorithms. By analyzing user behavior, tweet content, and activity patterns, the system successfully classifies tweets and accounts using classifiers like Naïve Bayes, Random Forest, SVM, and LightGBM. The experimental results confirm that LightGBM and Random Forest provide high accuracy in fake profile detection, while Naïve Bayes performs well for spam classification. This multi-algorithmic approach enhances the reliability and security of online social networks by reducing the spread of misinformation and spam activity.

6. FUTURE SCOPE

In the future, this system can be extended to support real-time detection of fake profiles and spam tweets across multiple social media platforms, not just Twitter. Deep learning models such as LSTM and BERT can be integrated to better understand complex language patterns and improve classification accuracy. Additionally, integrating user interaction patterns like retweet chains and hashtag propagation can help detect coordinated spam campaigns. The system can also evolve into a browser plugin or API service to assist end-users and developers in filtering fake content dynamically.

REFERENCES

1. Romanov, Aleksei, Alexander Semenov, Oleksiy Mazhelis, and Jari Veijalainen. "Detection of fake profiles in social media-Literature review." In International Conference on Web Information Systems and Technologies, vol. 2, pp. 363-369. SCITEPRESS, 2018.

2. Adikari, Shalinda, and Kaushik Dutta. "Identifying fake profilesin linkedin." arXiv preprint arXiv:2006.01381 (2020).

3. Kaubiyal, Jyoti, and Ankit Kumar Jain. "A feature based approach to detect fake profiles in Twitter." In Proceedings of the 3rd International Conference on Big Data and Internet of Things, pp. 135-139. 2019.

4. Elovici, Yuval, F. I. R. E. Michael, and Gilad Katz. "Method for detecting spammers and fake profiles in social networks." U.S. Patent 9,659,185, issued May 23, 2019

5. Elyusufi, Y. and Elyusufi, Z., 2019, October. Social networks fake profiles detection using machine learning algorithms. In The Proceedings of the Third International Conference on Smart City Applications (pp. 30 40). Springer, Cham.

6. Ozbay, F.A. and Alatas, B., 2020. Fake news detection within online social media using supervised artificial intelligence algorithms. Physica A: Statistical Mechanics and its Applications, 540, p.123174.

7. Gurajala, S., White, J.S., Hudson, B. and Matthews, J.N., 2015, July. Fake Twitter accounts: profile characteristics obtained using an activity-based pattern detection approach. In Proceedings of the 2015 International Conference on Social Media & Society (pp. 1-7).

8. Ramalingam, D. and Chinnaiah, V., 2018. Fake profile detection techniques in large-scale online social networks: A comprehensive review. Computers & Electrical Engineering, 65, pp.165 177.

9. [9]. Ojo, Adebola K. "Improved model for detecting fake profiles in online social network: A case study of twitter." Journal of Advances in Mathematics and Computer Science (2019): 1 17.

10. Basha, A. M., Rajaiah, M., Penchalaiah, P., Kamal, C. R., & Rao, B. N. (2020). Machine learning-structural equation modeling algorithm: The moderating role of loyalty on customer retention towards online shopping. *Int. J*, *8*, 1578-1585

11. Basha, A. M., et al. "Machine learning-structural equation modeling algorithm: The moderating role of loyalty on customer retention towards online shopping." *Int. J* 8 (2020): 1578-1585.

12. Basha, A. M., Rajaiah, M., Vijayakumar, O., Haranath, Y., & Srinivasulu, T. (2020). Green and lean industrial engineering practices in selected manufacturing units in Andhra Pradesh: statistical analysis. *International Journal*, *8*(5).

13. Basha, AM Mahaboob, et al. "Green and lean industrial engineering practices in selected manufacturing units in Andhra Pradesh: statistical analysis." *International Journal* 8.5 (2020).

14. Basha, A. M., Reddy, P. C., & Murthy, G. R. K. (2019). Moderating role of security and reliability on high customer satisfaction: Relationship among ease of use-content-service quality with respect to customer satisfaction in digital banking transactions

15. Basha, A. M. M. (2020). HR analytics using R-machine learning algorithm: Multiple linear regression analysis. *International Journal of Innovative Technology and Exploring Engineering, 9*(5), 1179-1183.

16. Basha, A. M., Ankaiah, B., Srivani, J., & Dadakalander, U. (2020). Real estate analytics with respect to Andhra Pradesh: Machine learning algorithm using R-Programming. *International Journal of Scientific & Technology Research*, *9*(4), 2140-2144.

17. Rajaiah, M., Rao, R. N., Nagendra, K. V., Basha, A. M., Rao, B. N., & Vijayakumar, O. (2020). Mathematical modeling on marketing analytics. *Journal of Critical Reviews*, *7*(19), 4189-4198.

18. Venkatrayulu, C., Gurumoorthi, S. K., Basu, S., Jayalakshmi, M., & Basha, A. M. (2023). The Mediating Role of Digital Marketing Practices in Relationship between the Standard Marketing Strategies and the Market Growth of Fisheries and Aqua Products in India. *Journal of Survey in Fisheries Sciences*, *10*(2S), 2292-2300.

19. Rekha, Y. C., Basha, A. M., & Prasad, P. S. (2020). Mediation effect of organizational performance in the relationship between human resource development practices and organizational learning.

20. Basha, M. (2014). Statistical Analysis on the Sustainable development of India with Reference to Agricultural Sector. *International Multidisciplinary Research Journal*, 1-4.