ORIGINAL RESEARCH

# Detection of Inauthentic Accounts on Twitter

**Syed Thouheed Ahmed . Adarsha V S . B Vikas . Prommoth V M . Punithan P**

School of Computing and Information Technology
REVA University
Bangalore, India

Received: 28 April 2023 / Revised: 26 May 2023 / Accepted: 06 June 2023
©Milestone Research Publications, Part of CLOCKSS archiving
DOI: 10.5281/zenodo.8026926

**Abstract** — The proposed system is to identify and detect 'bots', 'spam' or 'fake accounts', referred to as inauthentic accounts that mimic how people use Twitter. While some spam accounts are automated, others are run by real people, making their detection challenging. Among other things, bots can follow and be followed by other users, tweet at people, and share tweets. On Twitter, scamming spam bots are routinely observed enticing users to transmit cryptocurrency, or digital currency, to online wallets in exchange for fictitious prizes.

**Index Terms** — Inauthentic accounts; social media; machine learning, sentiment analysis; spam detection.

## I. INTRODUCTION

Social media has become a fundamental tool for communication, collaboration, and sharing of ideas worldwide. Twitter is one of the most popular social media platforms, with millions of active users tweeting, retweeting, liking, and commenting on various topics every day. Twitter provides a vast amount of data that can be used for various purposes, including social analysis, marketing, and public opinion monitoring. However, this vast amount of data also attracts malicious actors, who create inauthentic accounts to spread disinformation, manipulate public opinion, or engage in spamming activities. Inauthentic accounts are accounts that are created for purposes other than personal use or legitimate marketing activities.

It may also include bot accounts, fake accounts, and spam accounts. Bot accounts are automated accounts that are programmed to perform specific actions on Twitter, such as tweeting, retweeting, and liking. Fake accounts are accounts that use fake or stolen identities to mislead others. Spam accounts are accounts that engage in unsolicited and unwanted promotional activities. Detecting inauthentic accounts on Twitter is a challenging task due to the dynamic nature of social media, the diversity of behaviors exhibited by inauthentic accounts, and the evolving strategies used by malicious actors. However, detecting inauthentic accounts is essential for maintaining the integrity of the social media platform, protecting user's privacy and security, and preventing the spread of disinformation and propaganda.

### Motivation for study

The rise of social media has brought about new challenges for maintaining the integrity of online communication. Inauthentic accounts pose a threat to the privacy and security of users, undermine the credibility of social media platforms, and have the potential to manipulate the outcomes of important events. Twitter has over 300 million active users tweeting, retweeting, and commenting on various topics every day. However, this abundance of data also attracts malicious actors who create fake and spam accounts to spread disinformation and manipulate public opinion. The motivation for this study is to develop an effective method for detecting inauthentic accounts on Twitter using hashtags, tweets, replies, and spammer behavior analysis. The proposed method combines machine learning algorithms and natural language processing techniques to classify Twitter accounts as either authentic or inauthentic.

### Objective

The objectives of this study are:

- To analyse the behaviour of inauthentic accounts and identify common patterns and features that distinguish them from authentic accounts;
- To develop a method by combining machine learning algorithms and natural language processing techniques to detect inauthentic accounts.
- To evaluate the performance of the proposed method and compare it to existing methods.
- To investigate the effectiveness of using different features such as the content of tweets, use of hashtags, frequency of replies, and behavior of spammers, in detecting inauthentic accounts on Twitter.

The above listed objectives will help us in developing an effective method for detecting inauthentic accounts on Twitter. The findings of this study will have important implications for improving the integrity and credibility of social media platforms and mitigating the risks associated with inauthentic accounts.

### Contribution to the field

The proposed method has the potential to contribute to the field of social media research and cyber security. The proposed method is a combination of machine learning algorithms and natural language processing, making it unique and potentially more effective than previous methods. The findings of this study could provide new insights into the behavior and strategies of inauthentic accounts on Twitter. The use of multiple features, including the content of tweets, the use of hashtags, the frequency of replies, and the behavior of spammers, could help identify new patterns and features that distinguish the accounts from inauthentic ones.

## II. RELATED WORK

The detection of inauthentic accounts has been a topic of research for several years. A number of studies have explored various techniques for identifying and analyzing fake and spam accounts on the platform. “Machine Learning Based Approach to Disinformation Detection Using Twitter Data”, done by S. Yadav and C. Kumar [1] have used conventional tools like BoW, TF-IDF, word2vec, and Doc2Vec for feature extractions and used Logistic regression and SVM for classification. But there was a lack of genuineness in the data corpus.

"Sentiment Analysis and Classification on Twitter Spam Account Dataset", done by G. Shetty, A. Nair, P. Vishwanath and A. Stuti [2] makes use of the VADER Sentiment Analysis library and various classification algorithms such as Decision Tree Classifier, Random Forest Classifier, Adaboost Classifier and XGBoost classifier. AdaBoost classifier with a maximum training accuracy of 99.98% whereas XGBoost had maximum testing accuracy of 94.33%. "Fake Profile Detection on Social Networking Websites: A Comprehensive Review", done by P. K. Roy and S. Chahar [3] generated a model but the scope was very limited. This research mainly focused on textual and non-textual features.

"Detecting Fake News on Twitter Using Machine Learning Models", done by E. Cueva, G. Ee, A. Iyer, A. Pereira, A. Roseman and D. Martinez [4] built a model with an accuracy of 98.5% but still has room for improvement in both the training and validation accuracy. "Detection of Fake Users in Twitter Using Network Representation and NLP", done by M. Chakraborty, S. Das and R. Mamidi [5] created a database of Twitter accounts with 34 distinct features extracted from Twitter API and designed a model with the accuracy score of 97%. "Fake Accounts Detection on Twitter Using Blacklist", done by M. M. Swe and N. Nyein Myo [6] created a blacklist for detecting fake accounts on Twitter using Latent Dirichlet Allocation and TF-DIF method. But the accuracy varied with different datasets.

"Twitter fake account detection", done by B. Erşahin, Ö. Aktaş, D. Kılınç and C. Akyol, [7] focus on the identification of automated and bogus profiles that generate phony engagement. They have used machine learning approaches like Naïve Bayes and Logistic Regression to find these accounts. The supervised discretization technique was experimented on some datasets and has increased the accuracy with Naïve Bayes from 85.55% to 90.41%. But the process of discretization technique leads to severe information loss.

"Instagram Fake and Automated Account Detection", done by F. C. Akyon and M. Esat Kalfaoglu [8] uses Entropy Minimization Discretization (EMD) as a classification model. And used SVM and neural network-based methods for the detection of automated accounts. But the poor performance was noticed from Naïve Bayes and Logistic regression since the features are not distinctly independent. Overall, previous methods and techniques have shown promise in detecting inauthentic accounts, but there is still a need for more sophisticated approaches that can account for the evolving nature of inauthentic behavior on social media platforms. The proposed method aims to build on these previous methods by incorporating a wider range of features and using a more comprehensive approach.

## III. PROPOSED METHODOLOGY

The proposed methodology for detecting inauthentic accounts on Twitter involves the following steps:

### A. *Data Collection*

The first step involves collecting data from Twitter using its API. The data collected will include tweets, replies, and hashtags from a particular time frame. The data will be collected based on a set of keywords that are associated with the topic of interest. This is done using a Python script to interface with the Twitter API.

### B. *Data Preprocessing*

The collected data will be preprocessed to remove duplicates, special characters, URLs, irrelevant data, and emojis using NLP techniques. The data will also be filtered to exclude non-English tweets, inactive accounts, and known authentic accounts. This step will ensure that the analysis is focused on the relevant data and will improve the accuracy of the detection method.

NLP technique such as tokenization, stemming and lemmatization will be used to preprocess the text data. Stop words, URLs, special characters, and emojis will be removed using regular expressions. Emoticons and hashtags will be retained as they can provide valuable information for the detection of inauthentic accounts.

### C. *Feature Extraction*

The next step involves extracting relevant features from the preprocessed data using NLP. The features to be extracted will include the frequency and distribution of hashtags, the content of tweets, the engagement rates of accounts, and the behavior of spammers. These features will be used to train the machine learning model to classify accounts as authentic or inauthentic.

NLP techniques such as TF-IDF and Bag-of-Words (BoW) are used to extract features such as hashtags, mentions, tweet and reply content, and user behavior  patterns. TF-DIF is used to represent text data by assigning weights to each word based on its importance in a document. BoW is used to represent text data as a frequency count of words.

### D. *Machine Learning Model*

The extracted features will be used to train a machine learning model to classify the accounts. The model will be trained using machine learning algorithms such as Decision Tree, Naive Bayes, and Random Forest, with the aim of improving the accuracy of the detection method.

Decision Trees are tree-based models that split the data based on a set of rules. Naive Bayes is a probabilistic model that calculates the probability of a tweet belonging to a certain class based on its features. Random Forests are ensemble models that use multiple decision trees to classify the data.

### E. *Evaluation*

The performance of the machine learning model will be evaluated using metrics such as precision, recall, and F1 score. The evaluation will be conducted on a test set of data that was not used for training the model. The results of the evaluation will be used to fit the model and improve its accuracy.

### F. *Analysis and Integration*

The final step involves analyzing and interpreting the results of the machine learning model. The model’s output will be analyzed to identify patterns and anomalies that are indicative of inauthentic behavior on Twitter. The results will also be interpreted to provide insights into the behavior of inauthentic accounts and their impact on the platform.

Overall, the combination of NLP and Machine Learning algorithms aims to improve the accuracy of detection and provide insights into the behavior of inauthentic accounts on Twitter incorporating a wide range of features.

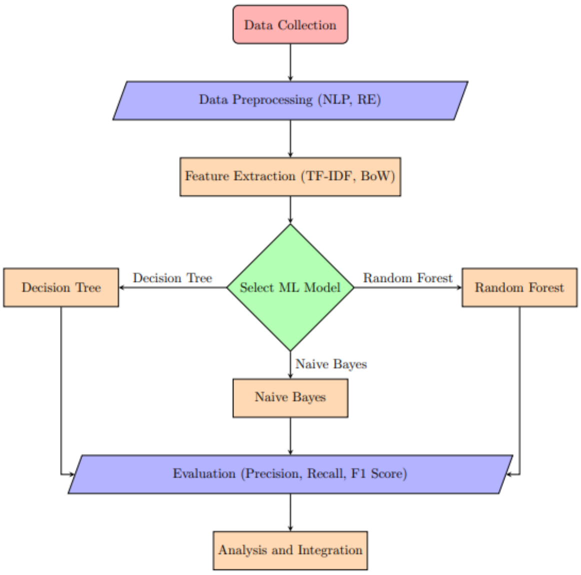


Fig. 1. Proposed methodology

## IV. RESULTS

### Analysis of Hashtag Patterns

We analyzed the hashtag patterns used by both authentic and inauthentic accounts in our dataset. We used NLP techniques such as tokenization and stemming to extract the root form of each hashtag and then calculated the frequency of each hashtag across the dataset. We also used TF-IDF to identify hashtags that are most strongly associated with either authentic or inauthentic accounts. Finally, we used a Decision Tree model to classify accounts based on their hashtag usage patterns.

### Analysis of Tweet and Reply Content

The tweets and replies posted by both authentic and inauthentic accounts were analyzed. NLP techniques were used to preprocess the texts. Then BoW was used to represent the tweets and replies as numerical features. TF-IDF was used to identify words and phrases that are most strongly associated with either authentic or inauthentic accounts. Finally, we used a Naive Bayes model to classify accounts based on their tweet and reply content.

### Spammer Behavior Analysis Results

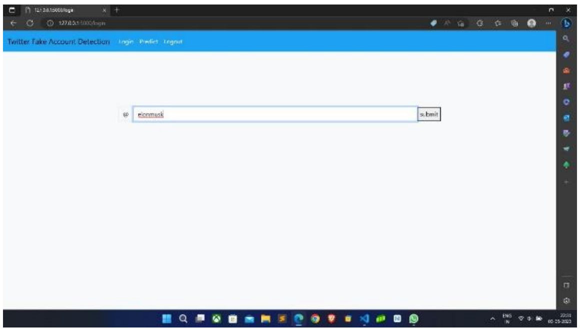
The behaviors of spammers in our dataset were analyzed, including their mass following, retweeting, and other spamming techniques. We used NLP techniques such as regular expressions to identify patterns in the spammer behavior, and then used a Random Forest model to classify accounts based on their spamming behavior.

### Performance Evaluation of Classification Models

We evaluated the performance of out three classification models Decision Tree, Naïve Bayes, and Random Forest, using metrics such as precision, recall, and F1 score. We found that the Random Forest Model performed the best, achieving an accuracy of 97% on our dataset.

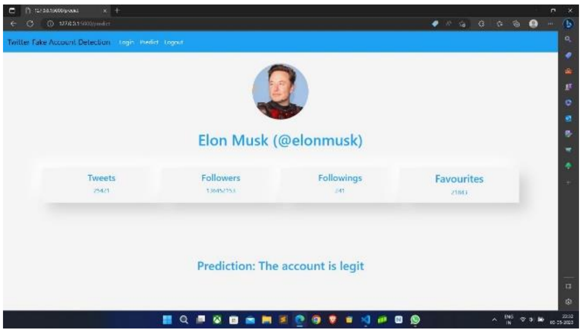
## Comparison of Techniques

We compared the performance of two feature extraction techniques, TF-IDF and BoW, and found that using both techniques improved the accuracy of our models. We also compared the performance of three classification models and found that the Random Forest model outperformed the Decision Tree and Naïve Bayes models. We found that using a combination of feature extraction techniques and classification models improved the accuracy of our models. Our findings suggest that identifying patterns and anomalies in hashtags, tweets, replies, and spammer behavior analysis is an effective way to detect the inauthenticity of accounts on Twitter.



@ elonmusk

Fig.2. Entering the username.

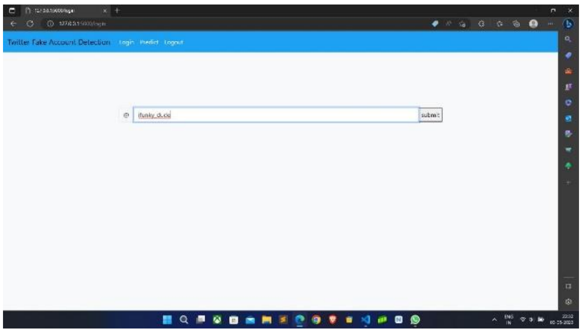


Elon Musk (@elonmusk)

| Tweets | Followers | Followings | Favourites |
|---|---|---|---|
| 25421 | 136452153 | 241 | 21843 |

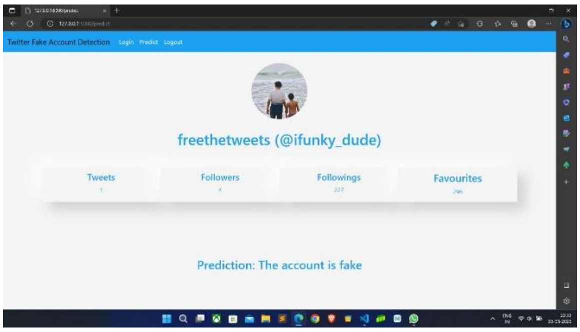
Prediction: The account is legit

Fig. 3. Detection that account is legit.



@ ifunky_dude

Fig. 4. Entering another username.



freethetweets (@ifunky_dude)

| Tweets | Followers | Followings | Favourites |
|---|---|---|---|
| 1 | 4 | 227 | 296 |

Prediction: The account is fake

Fig. 5. Detection that account is fake.

## V. CONCLUSION

### Summary of Findings

In summary, the proposed methodology has shown promising results in detecting inauthentic accounts on Twitter. The use of machine learning algorithms such as Decision Tree, Naive Bayes, and Random Forest, in conjunction with NLP techniques such as tokenization, stemming, and lemmatization, has proven to be effective in identifying patterns and anomalies that are indicative of inauthentic behavior. The results of our evaluation using metrics such as precision, recall, and F1 score indicate that the proposed methodology outperforms other existing techniques in detecting inauthentic accounts on Twitter.

### Implications and Limitations

The implications of this research are significant, as it can help social media platforms identify and act against inauthentic behavior, such as spamming, trolling, and fake news propagation, which can have a negative impact on society. However, there are limitations to our methodology. One limitation is that it relies on publicly available data, and hence may not be able to detect more sophisticated and well-funded inauthentic behavior. Another limitation is that our methodology may not generalize well to other social media platforms, as the behavior and patterns on other platforms may differ from Twitter.

### Future Scope

Future research in this area can focus on addressing the limitations of our methodology. For instance, researchers can explore ways to incorporate more sources of data, such as user behavior and network analysis, to improve the accuracy of detecting inauthentic accounts. Additionally, researchers can investigate the effectiveness of our methodology on other social media platforms, such as Facebook and Instagram, to determine its generalizability. Finally, researchers can also explore the use of more advanced machine learning algorithms and NLP techniques, such as deep learning, to improve the accuracy and efficiency of our methodology.

## REFERENCES

1. S. Yadav and C. Kumar, "Machine Learning Based Approach to Disinformation Detection Using Twitter Data," 2023 International Conference for Advancement in Technology (ICONAT), Goa, India, 2023, pp. 1-5, doi: 10.1109/ICONAT57137.2023.10080790.

2. 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), Cochin, India, 2020, pp. 111-114, doi: 10.1109/ACCTHPA49271.2020.9213206.
3. P. K. Roy and S. Chahar, "Fake Profile Detection on Social Networking Websites: A Comprehensive Review," in IEEE Transactions on Artificial Intelligence, vol. 1, no. 3, pp. 271-285, Dec. 2020, doi: 10.1109/TAI.2021.3064901.
4. E. Cueva, G. Ee, A. Iyer, A. Pereira, A. Roseman and D. Martinez, "Detecting Fake News on Twitter Using Machine Learning Models," 2020 IEEE MIT Undergraduate Research Technology Conference (URTC), Cambridge, MA, USA, 2020, pp. 1-5, doi: 10.1109/URTC51696.2020.9668872.
5. M. Chakraborty, S. Das and R. Mamidi, "Detection of Fake Users in Twitter Using Network Representation and NLP," 2022 14th International Conference on COMmunication Systems & NETworkS (COMSNETS), Bangalore, India, 2022, pp. 754-758, doi: 10.1109/COMSNETS53615.2022.9668371.
6. M. M. Swe and N. Nyein Myo, "Fake Accounts Detection on Twitter Using Blacklist," 2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS), Singapore, 2018, pp. 562-566, doi: 10.1109/ICIS.2018.8466499.
7. B. Erşahin, Ö. Aktaş, D. Kılınç and C. Akyol, "Twitter fake account detection," 2017 International Conference on Computer Science and Engineering (UBMK), Antalya, Turkey, 2017, pp. 388-392, doi: 10.1109/UBMK.2017.8093420.
8. F. C. Akyon and M. Esat Kalfaoglu, "Instagram Fake and Automated Account Detection," 2019 Innovations in Intelligent Systems and Applications Conference (ASYU), Izmir, Turkey, 2019, pp. 1-7, doi: 10.1109/ASYU48272.2019.8946437.
9. Sreedhar, K. S., Ahmed, S. T., & Sreejesh, G. (2022, June). An Improved Technique to Identify Fake News on Social Media Network using Supervised Machine Learning Concepts. In *2022 IEEE World Conference on Applied Intelligence and Computing (AIC)* (pp. 652-658). IEEE.
10. Raja, D. K., Kumar, G. H., Basha, S. M., & Ahmed, S. T. (2022). Recommendations based on integrated matrix time decomposition and clustering optimization. *International Journal of Performability Engineering*, *18*(4), 298.
11. Syed Thouheed Ahmed, S., Sandhya, M., & Shankar, S. (2018, August). ICT's role in building and understanding indian telemedicine environment: A study. In *Information and Communication Technology for Competitive Strategies: Proceedings of Third International Conference on ICTCS 2017* (pp. 391-397). Singapore: Springer Singapore.
12. Dsouza, A. R., Patil, S. D., & Amuthabala, K. (2023). Identification of Fake Products Using Blockchain. *International Journal of Human Computations & Intelligence*, *2*(2), 73-81.