Detection of Fake Profiles Using Machine Learning

Aum Patel, Jayraj Pamnani, Vrajesh Patel and Shubrath Shakyavanshi

March 18, 2024

DETECTION OF FAKE PROFILES USING MACHINE LEARNING

Abstract*— In today's world, the social media platforms are being used on daily basis and has become an important part of our lives. The number of peoples on social media platforms are incrementing at a greater level for malicious use. There are numerous cases where produced accounts have been effectively distinguished utilizing machine adapting techniques however the amount of research work is very low to recognize counterfeit characters made by people. For bots the ML models used various features to calculate the no. of followers to the no. of friends that an account has on social media platforms (SOCIAL MEDIA PLATFORMS). The number of friends to the no. of followers of any account are easily available in the account profiles and no rights are violated of any accounts. In order to accomplish the task of detecting, identifying and eliminate the fake accountswe establish a forged human account. In this project we intend to give a framework with which theautomatic detection of fake profiles can be done so that the social life of people become secured and using automatic detection technique we can make it easier for the sites to manage the huge numberof profiles which can't be done manually.

***Keywords*— Machine Learning, Social Media Platforms, Fake Accounts.**

I. INTRODUCTION

In today's online social networks there have been a lot of problems like fake profiles, online impersonation etc., Till date, no one has come up with a feasible solution to these problems. In this project we intend to give a framework with which the automatic detection of fake profiles can be done so that the social life of people become secured and using automatic detection technique we can make it easier for the sites to manage the huge number of profiles which can't be done manually.

People create accounts to share social media data using various social networking platforms. Userstend to create accounts with anonymous or wrong data to propagate fake news to avoid revealing their identity. Users also tend to create accounts either in the name of some other person (Identity Theft) or intrude into their accounts. Fake accounts creation also has some targeted financial benefits. Fake accounts also got created during incidents such as the Boston Marathon blast and COVID-19 Prime Minister Relief Fund accounts. Bots or automated programs maintain these fake accounts and help in the network's faster and deeper spread of fake news.

The research has the main objective of developing an automatic mechanism for the detection andremoval of bots present on social media platforms. It utilizes machine learning technology with a bot detection technique for providing higher security to the platform and the related person. There are several gaps present in the previous solution, such as it is tough to check account information, many genuine account holders were also removed in massive quantity due to which accurate and authentic information is not able to reach the people.

1) Methodology:Our website for detecting fake Twitter profiles and bot accounts employs a multifaceted methodology to identify suspicious accounts. It utilizes a combination of techniques such as natural language processing (NLP) to analyze the content and language used in tweets, examining the frequency ofrepetitive and automated posts, and assessing account metadata, including account creation date and follower-to-following ratios. Machine learning models are employed to classify accounts based on these features, allowing for the identification of suspicious behavior patterns. Additionally, our website also incorporate behavioral analysis, looking at the activity and interaction patternsof accounts, such as the types of accounts they engage with and the timing of their posts. By amalgamating these techniques, our platform can distinguish fake profiles and bot accounts with a high degree of accuracy, offering users a valuable tool to maintain the authenticity of their online interactions and content.

2) System Architecture :Although fake profile detection is a robust field, but it has many challenges and gaps which we have discussed and have based our work on. There are a lot of existing solutions to fake profile detectionbut all of them have some or the other drawback. There is a lot of work already done in this field and a lot more needs to be done like improving upon the response time, prevention from fake accounts instead of detecting and dealing with their aftermaths.

Our work is aiming to deliver a system which will have the highest accuracy and hence will be effective in prevention from such fake profiles by implementing and comparing different algorithms. This is done by ensemble machine learning technique which speeds up the training of neural networks and helps them to take decisions faster.

Efficient parameter selection is also one of the major objectives of this work for which we are selecting six features manually which will give a better control on the output of neural networks. The proposed solution makes use of the hybrid of the machine learning techniques and combines their advantages and uses one to cancel out the loopholes of the other and hence delivering an efficient and cost-effective system.

3) Implementaion Details*:*
We started with collecting Twitter data through the Twitter API, including user profiles, tweets, and associated metadata. Cleaned and preprocessed the data, removing duplicates and irrelevant information. Then we extracted features such as account creation date, tweet content, posting frequency, follower-to-following ratio, and other relevant

attributes. Manually labeled a subset ofthe data as genuine or bot accounts to create a training dataset. Then we selected the most informative features for classification. Features include metadata (e.g., account age, activity patterns), and more. Splitted the labeled data into training and testing sets.Trained the Random Forest classifier on the training set, utilizing the labeled data to learn patterns associated with fake profiles and bot accounts. Random Forest is an ensemble learning technique that combines multiple decision trees to make predictions. Then we tuned the hyperparameters such as the number of trees,tree depth, and feature selection to optimize the model's performance. Evaluate the Random Forestmodel's performance on the testing data using metrics like accuracy, precision, recall, and F1-score to assess its ability to classify accounts accurately. Implemented a user interface on our website where users can input Twitter handles or URLs for analysis. The website sent these inputs to the trained Random Forest model for real-time evaluation. Display the model's findings to users, categorizing Twitter accounts as genuine, suspicious, or bots based on the model's predictions and associated confidence scores.Incorporate user feedback to improve the model's accuracy over time. Users can report false positives and false negatives to enhance the algorithm's performance.Hostedour website on a server and ensure its accessibility to users. Regularly update and retrain the Random Forest model to adapt to evolving bot behaviors and to enhance detection accuracy.

Figure 1 depicts the flowchart of our system. The dataset which we have is partitioned into two sets, Train Dataset and Test Dataset in the ratio 4:1. The train dataset then goes into Support Vector Machine and Logistic Regression Classifier where classes are predicted. Then these classifiers areappended to a voting classifier where final decision of class is made.

The output from voting classifier i.e. train data and the predicted class from voting classifier is fed to Neural Network classifier as input. After training has been completed, we get a Trained Systemon which Test dataset is ran to find the accuracy of the system.

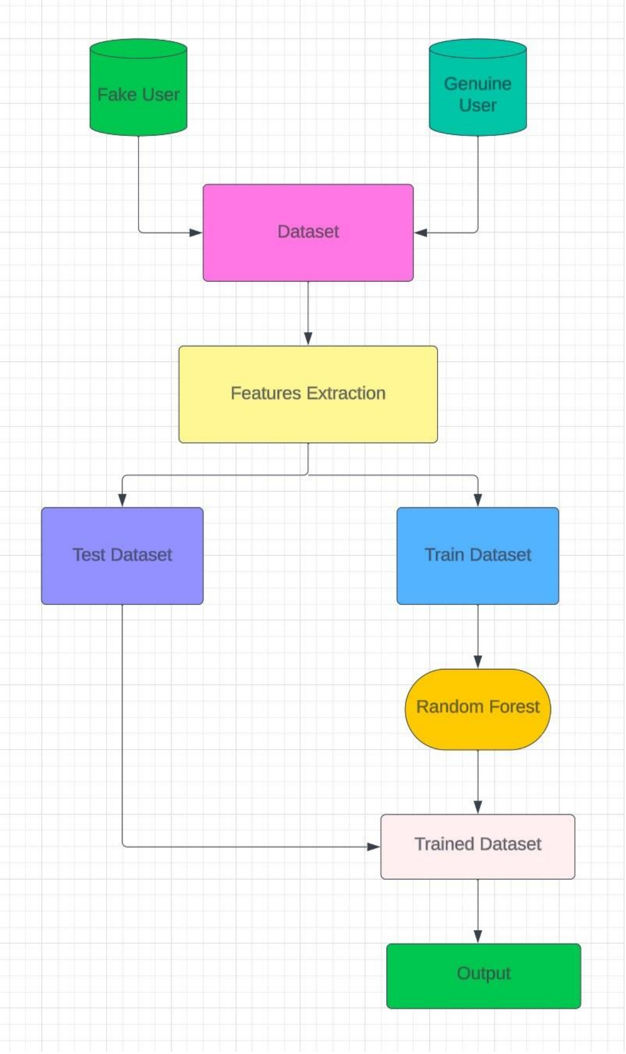

Fig. 1. System architecture

A. *Introduction of ML algorithms*

Bagging or bootstrap aggregation is a technique for reducing the variance of an estimated prediction function. Bagging seems to work especially well for high-variance, low-bias procedures, suchas trees. For regression, we simply fit the same regression tree many times to bootstrap sampled versions of the training data, and average the result. For classification, a committee of trees each cast a vote for the predicted class. Boosting was initially proposed as a committee method as well, although unlike bagging, the committee of weak learners evolves over time, and the members cast a weighted vote. Boosting appears to dominate bagging on most problems, and became the preferredchoice. Random forests is a substantial modification of bagging that builds a large collection of de-correlated trees, and then averages them. On many problems the performance of random forests is very similar to boosting, and they are simpler to train and tune. As a consequence, random forests are popular, and are implemented in a variety of packages. The essential idea in bagging is to average many noisy but approximately unbiased models, and hence reduce the variance. Trees are ideal candidates for bagging, since they can capture complex interaction structures in the data, andif grown sufficiently deep, have relatively low bias. Since trees are notoriously noisy, they benefit greatly from the averaging. Moreover, since each tree generated in bagging is identically distributed (i.d.), the expectation of an average of B such trees is the same as the expectation of any one of them. This means the bias of bagged trees is the same as that of the individual trees, and the only hope of improvement is through variance reduction. This is in contrast to boosting, where the treesare grown in an adaptive way to remove bias, and hence are not ID.

$$Z = (X - \mu) \sigma$$

II. FUTUREWORK AND CONCLUSION

We underscore that our website is designed to be highly adaptable and comprehensive. We aimto offer a solution for identifying fake profiles on Twitter also know as X. This versatility can be a significant advantage for users, as it saves time and effort by providing a unified solution for a widespread problem.

Emphasizing that our website is easy to use and have a user-friendly interface signifies a commitment to ensuring that the tool is accessible to a wide range of users, including those who may not have advanced technical knowledge. This feature can improve the user experience and make the app more appealing and efficient.

We highlight the credibility and reliability of your project. It suggests that the website's algorithms and detection methods are grounded in scientific research. Specifically, the reference tobot content and their persistence on platforms indicates that your solution is not based on guesswork but on well-founded, data-driven, and academically supported principles. Users can trust that the technology is based on a sound understanding of the problem.

There are countless benefits of your project for advertisers. It suggests that the tool can help advertisers assess the authenticity of the accounts they plan to use for advertising purposes. This is valuable because it can protect advertisers from spending resources on fake or unengaged followers. By offering this feature, our project becomes a valuable resource for businesses looking to ensure the effectiveness of their marketing efforts.

We underscore the importance of improving the precision and correctness of the model being used. We will have a commitment to enhancing the reliability and trustworthiness of the system. Increasing accuracy can involve refining the data collection process, fine-tuning algorithms, and reducing errors or false positives/negatives in the system's output.

A responsive website adapts to different devices and screen sizes, ensuring a seamless user experience. Improving website responsiveness demonstrates a desire to accommodate users on various platforms (desktop, mobile, tablet). Additionally, making the website engaging implies a focus on user experience, perhaps by incorporating interactive features, visually appealing design, and user-friendly navigation to keep visitors interested and active.

We plan to broaden the application of the algorithm from its current scope (Twitter) to other social media platforms such as Facebook and Instagram. It signifies a desire to adapt and scale the technology for multiple platforms, to address similar problems (like bot detection) across variousonline spaces.

We intend to enhance the model's understanding of the specific characteristics and dynamics of Twitter. The goal is to enable the model to identify subtle cues that are indicative of fake profiles or automated bots, and to do so with greater speed and efficiency while minimizing resource consumption. This likely involves a deep understanding of Twitter's ecosystem, its unique user behaviors, and the challenges associated with identifying inauthentic accounts.

We aim to offer access to the developed model or technology to external parties, such as businesses and advertisers. The objective is to enable these entities to leverage the model for moreeffective targeting of their user base. This can involve creating an interface or API that allows third parties to integrate the model into their marketing strategies, with the ultimate goal of improving the reach and impact of their advertisements or campaigns.

- Currently in the development stage with a two-month target for testing deployment.
- Create an application that is interactive, user-friendly, and optimized.
- User registration results in data storage on the server, which protects data.
- The suggested system for dairy food products attempts to minimize manual order processing and helps distributors and retailers foresee trends and optimize pricing.
- Emphasis on easy access to information and feedback, enhancing communication across the supply chain.
- Combining and managing various data sources effectively for better decision-making.
- The features that are planned include saving important data, emailing data, making it simple to find information using a search bar, and sending timely alerts for sales, offers, and payments.
- Deliver a solid and user-centered application that streamlines supply chain procedures in the dairy industry as the overall objective.

REFERENCES

[1] W. J. Yan1, X. Chen," Big Data Analytics for Empowering Milk Yield Prediction in Dairy Supply Chains", IEEE International Conference on Big Data, 2015

[2] Birhanu Megersa Lenjiso, Jeroen Smits, Ruerd Ruben, "Transforming dairy production and marketing: An essential step in ensuring food and nutritional security among smallholder farmers in rural Ethiopia", IEEE Canada Internationaz Humanitarian Technology Conference (IHTC2015), 2015.

[3] Licuixia, "Empirical study on production efficiency of dairy products processing industry in heilongjiang province based on DEA model", International Conference on Computer Application and System Modeling (ICCASM,2010).

[4] Metin, M., Ka¸sıkc¸ı, M. (2010). Milk Component and its Proccess. 9. Edition. Ege Univ Engineering Faculty Press, 33, pp:182, ˙Izmir, Turkey.

[5] Collard BL, Boettcher PJ, Dekkers JCM, Petitclerc D, Schaefer LR (2000) Relationships between energy balance and health traits of dairy cattle in early lactation. J Dairy Sci 83:2683–2690.

[6] Cabrera, V. E., F. Contreras, R. D. Shaver, and L. Armentano. 2012. Grouping strategies for feeding lactating dairy cattle. Pages 13–14 in Proc. Four-State Dairy Nutrition and Management Conference, Dubuque, IA. Wisconsin Agri-business Association, Madison.

[7] Savchenko A, Mikhieieva S K and Holynska M 2018 Analysis and audit of key economic indicators of economic entities (a case study of dairy industry) Baltic J. of Econ. Studies 4(3) 271–275.

[8] Wing-Kam Ng, Venus Khim-Sen Liew, "Revisiting the Performance of MACD and RSI Oscillators, Risk and Financial Management", ISSN 1911-8074, 2014.

[9] Matthew Ridge, Brian O'Donovan, "The use of big data analytics in the retail businesses in South Africa", Academic journals, 2015.

[10] G.Morota, R.V..Ventura,F.F.Silva, M.Koyama, and S.C.Fernando, "Machine learning and data mining advance predictive big data analysis in precision animal agriculture", July 12,2017.

[11] PKS. Podiel Slovensk´ych Potrav´ın na Pultoch Obchodov. 2019. Available online: http://potravinari.sk/page6 956sk.html (accessed on 20 February 2020).

[12] An Implementation of Diary Food Products using Android Application S. Brintha Rajakumari1[Associate Professor, Department of CS, BIHER, Chennai.], T. Susendran2[Department of Computer Science, BIHER, Chennai.]

[13] Sivapalan Achchuthan, Rajendran Katana than, "A Study on Value Chain Analysis in Dairy Sector Kilinochchi District", Global Journal of Management and Business Research, in 2012.

[14] Berberoglu, E. (2002). Application of Model Determination to ˘ Animal Science Using the Variable Selection Method. MSC thesis, Gaziosmanpa¸sa Univ, Tokat, Turkey.

[15] Grunert KG. Food quality and safety: consumer perception and demand. European Review of Agricultural Economics 2005.

[16] Sivapalan Achchuthan, Rajendran Katana than, "A Study on Value Chain Analysis in Dairy Sector Kilinochchi District", Global Journal of Management and Business Research, in 2012.

[17] Marco Grossi , Anna Pompei , Massimo Lanzoni , Roberto Lazzarini , Diego Matteuzzi , Bruno Ricco, "Total Bacterial Count in Soft-Frozen Dairy Products by Impedance Biosensor System", IEEE paper.2009.

[18] Komal Patel1, Ronak Chudasama2 , Sagar Dobariya, "Dairy Production Analysis and Prediction Tool using BIG DATA", Volume 02, Issue 09; September– 2016.