Mani Manvith Devineni, Mohammed Ali-Khan, Rohita Sreya Namburi

Ramakrishna Koganti

CSE 5301.005

24 November 2024

Unmasking Digital Imposters:

Fake Social Media Detection Using Random Forest

Introduction

In today's day and age, social media platforms have become abundant, using the interactions of people to influence the way lives are lived, and ideas are built. As one of the use cases of social media is the ability to influence, users begin to come up with innovative and creative ways to bring their desired content to attract an audience and build on shared relationships. This influence can be beneficial as it can be used to build communities, educate one another on the ways the universe functions, and bring coverage to the day-to-day activities and events that occur around the globe. As a researcher from the American Public University states, "Social media wields cultural influence on fashion and food trends, family and adolescent health issues, world news and local events, political and community action events [Ashar 1]."

However, this interconnectedness and transmission of ideas towards the enlightenment and betterment of society can also take a darker turn, with certain influencers having the ability to manipulate audiences through the creation of fake social media profiles to create an illusion of an existing community that showcases certain harmful and malicious misinformation. For example,

according to the Air Force Security Forces Center, “Fake social media pages are a global problem. In the second quarter of 2022 alone, Facebook reported taking action on 1.4 billion bogus accounts. Fake social media accounts can promote phony products, spread scams and even share lies and misinformation [Joyner 10].” In another instant found by Dr. Olga Boichak, “In the US election, social bots were not only amplifying, but also spreading candidates’ messages, helping them reach new audiences [Boichak 2].” Such fake social media profiles can also coerce other users on the platform into harmful interactions, such as phishing attacks, which undermine the effectiveness and integrity of social media platforms. Frontiers in Physics states, “Social bots mimic human behavior to occupy an important role in public opinion research and amplify their influence through social networks, which in turn change human opinions and behaviors [Cai 3].” According to Frontiers in Psychology, “More aggressive and tailored messages launched on social media platforms may enhance negative consumption behaviors and there can be dire consequences for countries’ economies and for individuals’ indulging in these behaviors … [Pellegrino 12].”

The solution seems simple, allow real users to continue to interact with each other, while detecting fake user profiles and ensure platform integrity by removing disingenuous interactions. Luckily, social media platforms have been implementing such strategies to create a safer and more genuine interconnectedness with real users but falls short in certain areas. This introduces the limitations of the current methods of fake social media detection, and an opportunity for the usage of more advanced machine learning methodologies to combat bot activity. It is also important to note, that according to the Journal of Medical Internet Research, “Media use is just one aspect of this complex structure, but it has the potential to set context, add to a social support system, and touch individuals and their closely tied friends and family [Gruzd 8].”

Problem Definition

Given that social media platforms have implementations to detect and neutralize fake bot accounts to restore platform integrity, there are limitations to many of these detection methods. Traditional approaches rely on user reporting systems, as only when a user runs into bot made accounts do users flag the account, and the account is passed for further inspection. Another approach is rules based, which observes certain parameter to decide whether the account is legitimate or not. Researchers from Partnership on AI, state, "There is an urgent need for technology platforms to play a role in facilitating increased access to methods for detecting manipulated media, yet it became clear through the workshop that the realities of adversarial attacks lead platforms to avoid widely collaborating with others on detection methods [Stray 15]." These approaches are shown to be inadequate, as fake user profiles are starting to become more and more sophisticated. User flagging is slow and subject to user bias, as a more advanced bot may appear to be more real than the user realizes. This approach is also limited, as more bot accounts would be created, far outweighing the platforms detection management to detect, analyze, and further investigate the account. According to the Journal of Computer and Communications, "Depending on how many of your friends may have tags or connection histories, Facebook uses an algorithm to identify bots. The aforementioned guidelines can be used to spot bot accounts, but they fall short when it comes to human-made false accounts [Chakraborty 4]." Rule based approaches are limited to how the parameters for their detection are laid out, as more advanced bots can generate the necessary requirements to appear as a real account. According to Springer Nature researchers, "Content-based and social context-based approaches are the primary methods in fake news detection [Uppada 16]."

Cue machine learning models, as this is where a more advanced detection mechanism can be used to detect fake social media profiles. Implementing it in this fashion allows for a more

robust system to build that automatically detects fake profiles, as opposed to the manual user reporting methods, and takes in several parameters such as post frequency, follower count, friend relationships, and other important factors that play into deciding between authentic accounts and bot generated ones. By using a machine learning approach, it would result in the culmination of all necessary parameters as decided by the needed platform as well as being an automatic process to filter out such unwanted malicious bot accounts.

Predictive Model Selection

Given the nature of the problem, a machine learning model is needed to appropriately train itself on the data, the data being several genuine and disingenuous social media accounts along with the given parameters, and then yield the best, most accurate detection rate as compared to other machine learning data models. Another factor that must go into the selection of the data model is its ability to handle large complex datasets and have a proven ability to accurately differentiate between real and fake accounts. Based on these qualities, the chosen data models for continued analysis are: Random Forest, Support Vector Machines, and Neural Networks.

Random Forest is a data machine learning model based on decision trees and is suited for high dimensional datasets with complex relationships. According to the International Journal of Advanced Computer Science and Applications researcher, “The Random Forest (RF) technique is a collective learning approach that integrates numerous decision trees to build a more robust and precise forecasting model [Wang 18].” This model is proven to be robust as it is less prone to be overfit over the data because of its ensemble nature. Random Forest puts an importance on feature detection as it provides insight into certain features of the dataset, which in this case is the parameters that make an authentic social media profile different and distinguishable than one that is bot generated, such as post frequency, follower to following ratio and the interactions it has with

other users. Additionally, Random Forest is scalable, meaning it can handle large amounts of datasets, as this is crucial in building an autonomous detection system for billions of social media accounts, given that some of these will be inauthentic.

Support Vector Machines, abbreviated as SVM, is the second of the machine learning models used and tested. This method is particularly effective in binary classification problems, which in our case is the detection between real and fake accounts, by finding the greatest boundaries between the two types of profiles and their differences in their respective features. SVM also has strengths in working between both linear and nonlinear relations, as it maps onto higher dimensional spaces allowing for the flexibility to handle intricate datasets. However, it is also shown to be computationally expensive, but still serves as an excellent model to use.

Neural Networks is the last of the machine learning models used for the analyzation of this problem. It is a powerful machine learning model that emphasizes identifying patterns within a large dataset. These patterns are meant to be highly complex, and in our case, would be mapping a relationship between the parameters of accounts into categorizing them as either authentic or inauthentic. Neural Networks are excellent at handling nonlinear relationships, given the complexities of the parameters that are being searched and compared, this makes it ideal for behavioral pattern analysis. Additionally, this machine learning model adapts itself based on the data it receives, as given the fake social media accounts that the model is being trained on, this would mean that the model itself would begin to see a correlation between the attributes of fraudulent accounts and better program itself into detecting them.

Analysis

Given the problem understanding and the machine learning models and techniques, the most ideal methodology has yet to be decided. To accomplish this, some more steps are needed. Data needs to be collected, as the machine learning models need loads of social media accounts, including certain attributes that differentiate real accounts from fake accounts. Secondly, the determining features also need to be determined, based both on the understanding of what fake versus real accounts have as well as how they would be listed in the dataset itself. After all this setup is done, the models can then be trained and tested using the same dataset as well as more additional datasets. The different models discussed previously are then compared, given the most accurate prediction methods and can then be finalized into a working prototype or proof of concept. This includes a front-end webpage that can take in a social media account, along with its properties and decipher on whether it is a genuine account, or bot created.

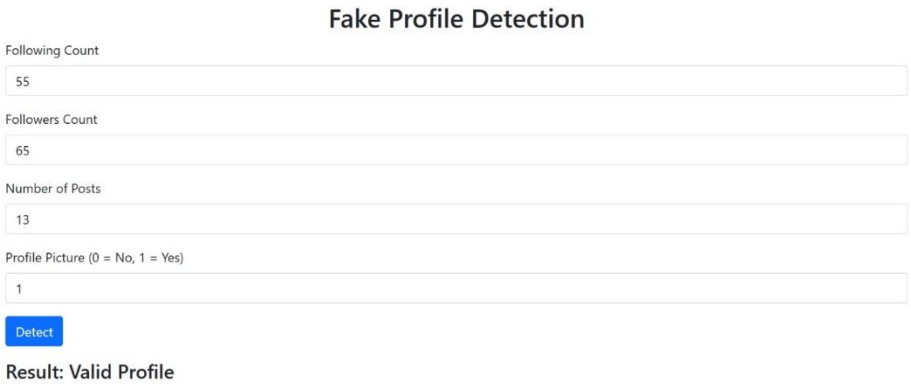

Figure 1. Front End Implementation of Final Detection Algorithm.

Data Collection.

Data is collected primarily from Kaggle. The dataset is then processed and cleaned, allowing only the necessary information to be passed over to the machine learning model.

Feature Engineering.

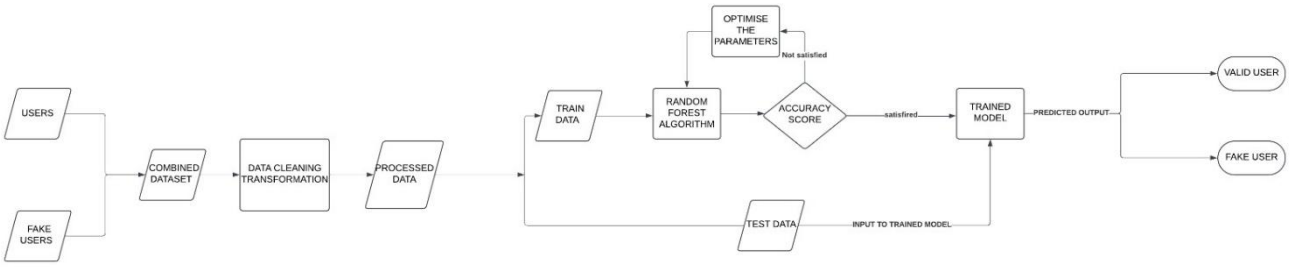

Figure 2. Pipeline of Final Detection Implementation.

For this prototype, the social media platform of Instagram is chosen. Different platforms would have different users and analytics that would need to be modeled for the platform appropriately. The features chosen from the dataset and for the machine learning model is the post frequency, follower to following ratio, and engagement patterns.

A genuine account would post casually, while a bot account would either not post content, in the case of being more used for fraudulent phishing scams, or post consistently or in high frequencies, as is the case for those spreading misinformation and trying to sway public opinion. Subsequently, it is not the only determinant, as real accounts may also not post frequently or at all, and others may post regularly, as they may belong to a legitimate organization or media network, as these organizations have dedicated media teams whose purpose is to regularly bring informed content to its constituents.

Additionally, genuine users tend to have an equal number of followers and accounts followed. Media organizations and institutions may have more followers than account followed. However, bots tend to be following more accounts than followed by, as these false accounts are interested in reaching out to people to spread misinformation or send malicious messages, which becomes an important feature to be analyzed within the datasets for the detection implementation.

Lastly, genuine users would have an appropriate number of messages sent to other accounts, while false accounts would have an abnormal number of messages sent, as they attempt to influence or coerce other users. Genuine users would like and comment posts as they gradually see through their respective content feed, while bots would not generally like, and the comments, if any, would be repetitive and this engagement activity are passed over to the machine learning model to decipher between real and fake accounts.

Model Training and Evaluation.

Given the preprocessed and cleaned dataset, as well as the identifications for the model to decide between real and fake accounts, the next step is to feed this information into a machine learning model. To reiterate, Random Forest, Support Vector Machines, and Neural Networks are the modeling tools used to judge these account datasets. First the dataset is split into an eighty percent training dataset and twenty percent testing dataset. The training dataset is used for the model to learn what feature makes a real account a real account and a fake account a fake account. The testing dataset is then used to verify if the machine learning model has adequately learned how to decipher between the two.

With the testing dataset tested, the results are further broken up into true positives, where true accounts are detected as true accounts, true negatives, where false accounts are detected as false accounts, false positives, where false accounts are mistakenly detected as true accounts, and false negatives, where true accounts are mistakenly flagged as false accounts. These results are then normalized, to compensate for the difference in the size of the original dataset for each category. This procedure is then repeated for all machine learning models.

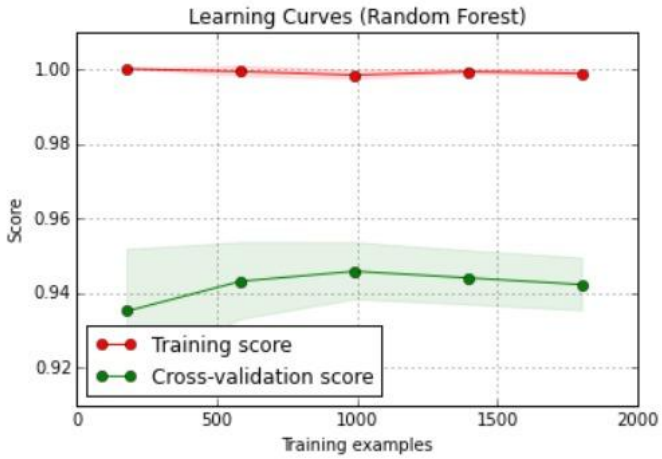

Figure 3. Learning Curve for Random Forest (Note Random Forest is better that SVM)

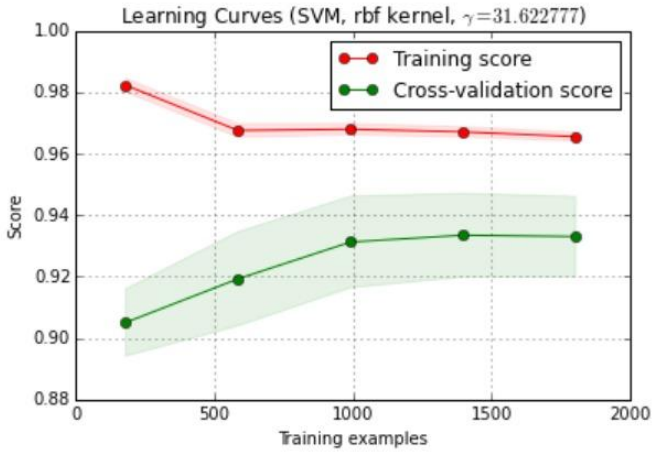

Figure 4. Learning Curve for SVM (Note SVM is less than Random Forest)

Model Comparison

With all three machine learning models used with the training dataset, and the result of implementing on the testing dataset, the results of the success rate of the models, which is number of true positives and true negatives against the false positives and false negatives, are then compared between the three models.

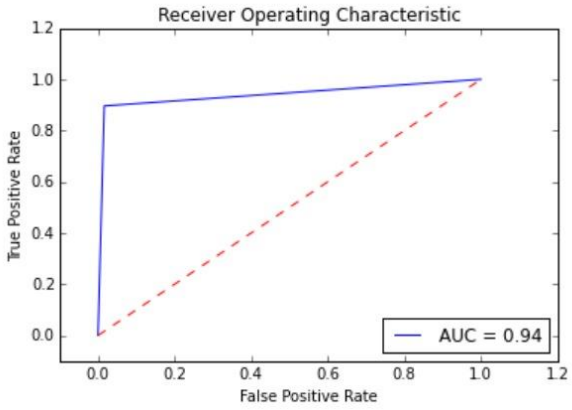

Figure 5. ROC Curve for Random Forest

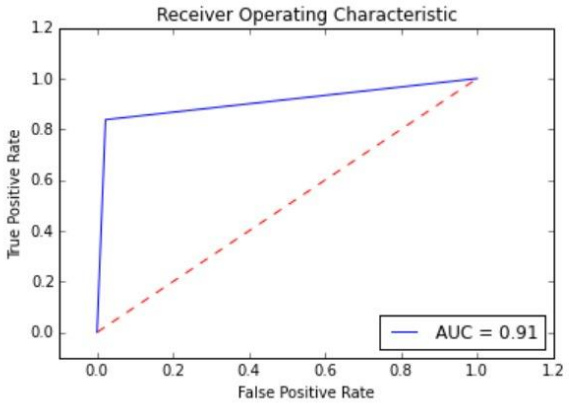

Figure 5. ROC Curve for SVM

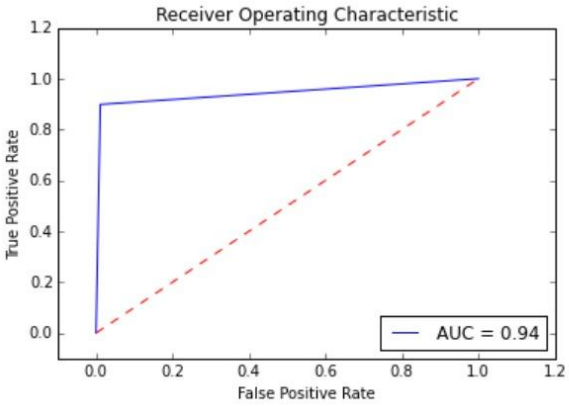

Figure 6. ROC Curve for Neural Networks

Conclusion

Based on the analysis and results from the trials of the three different data machine learning models, Random Forest yielded the highest accuracy, with a success rate of nearly 94%. This is due to the model's ability to handle high dimensional data, with complex relationships as it classifies the different social media accounts as real or fake. Random Forest also proved the importance of such chosen features for model fitting as post frequency, follower to following ratio, and engagement activity were necessary in making the meaningful classification decision.

Conversely, Support Vector Machines and Neural Networks also demonstrated strong potential, with success rates of 90% and 93% respectively. It is highly possible that both other methods can also be utilized, albeit by tweaking the current feature engineering needed to sample from the social media account. This in turn proves that providing a more advanced way to classify such social media accounts, machine learning would in fact make social media platforms safer for users to interact with and bring a better sense of platform integrity.

Recommendations

All in all, there are a few more ways this method can be further enhanced and improved upon. Note that this research and implementation was done for one specific platform, and that other platforms would have different metrics and deciding factors to authenticate legitimate users. Thus, implementations would need to be built dependent on the specific platform, but the general concept can still be extrapolated from this assessment. Secondly, social media evolves, just as people find new and innovative ways to express themselves and their ideas, interactions within social media are prone to change as well. This means machine learning models will have to adapt and be regularly updated and remodeled. For instance, as detection methods are used to filter out

inauthentic accounts that deem to spread misinformation, a rival to the elimination process will be artificially created content, as they are beginning to become an artistic way of expressing or exploring different ideas on one end, but also becoming a tool for misinformation to critical news stories on another. According to researchers from Harvard Kennedy School, "According to various voices, including some leading AI researchers, generative AI will make it easier to create realistic but false or misleading content at scale, with potentially catastrophic outcomes for people's beliefs and behaviors, the public arena of information, and democracy [Simon 14]." Hence, the filtering out of misinformation of artificially generated content will be a hurdling endeavor to tackle soon.

Despite all the limitations and obstacles in detecting fake social media profiles, one point is clear. Machine learning is an incredibly necessary tool to analyzing the vast amount of information being generated very minute on social media platforms, to promote a place where genuine people can express their thoughts freely and come together to form communities based on integrity.

Works Cited.

1. Ashar, Linda C. “Social Media Impact: How Social Media Sites Affect Society: American Public University.” *APU*, 2 May 2024, www.apu.apus.edu/area-of-study/business-and-management/resources/how-social-media-sites-affect-society/. Accessed 22 Nov. 2024.
2. Boichak, Olga. “Can Bots Influence Elections with the ‘Megaphone Effect’?” *The University of Sydney*, 17AD, www.sydney.edu.au/news-opinion/news/2021/02/17/can-bots-influence-elections-with-the-megaphone-effect.html. Accessed 22 Nov. 2024.
3. Cai, Meng, et al. “Differences in Behavioral Characteristics and Diffusion Mechanisms: A Comparative Analysis Based on Social Bots and Human Users.” *Frontiers*, 4 Apr. 2022, www.frontiersin.org/journals/physics/articles/10.3389/fphy.2022.875574/full. Accessed 22 Nov. 2024.
4. Chakraborty, Partha, et al. “Fake Profile Detection Using Machine Learning Techniques.” *SCIRP*, Scientific Research Publishing, 13 Oct. 2022, www.scirp.org/journal/paperinformation?paperid=120727. Accessed 22 Nov. 2024.
5. Chakraborty, Partha, et al. “Fake Profile Detection Using Machine Learning Techniques.” *SCIRP*, Scientific Research Publishing, 13 Oct. 2022, www.scirp.org/journal/paperinformation?paperid=120727. Accessed 22 Nov. 2024.
6. Clavin, Don. “Famous Phishing Incidents from History.” *Famous Phishing Incidents from History | Hempstead Town, NY*, www.hempsteadny.gov/635/Famous-Phishing-Incidents-from-History. Accessed 22 Nov. 2024.
7. Goyal, Aditya. “Twitter-Bot Detection Dataset.” *Kaggle*, 31 May 2023, www.kaggle.com/datasets/goyaladi/twitter-bot-detection-dataset. Accessed 22 Nov. 2024.

8. Gruzd, Anatoliy, and Caroline Haythornthwaite. “Enabling Community through Social Media.” *Journal of Medical Internet Research*, U.S. National Library of Medicine, 31 Oct. 2013, pmc.ncbi.nlm.nih.gov/articles/PMC3842435/. Accessed 22 Nov. 2024.
9. “Guide to Data Cleaning: Definition, Benefits, Components, and How to Clean Your Data.” *Tableau*, www.tableau.com/learn/articles/what-is-data-cleaning. Accessed 22 Nov. 2024.
10. Joyner, Bo. “Beware of Fake Social Media Accounts.” *Air Force Security Forces Center*, 19 Jan. 2023, www.afsfc.af.mil/News/Article-Display/Article/3271499/beware-of-fake-social-media-accounts/#:~:text=Fake%20social%20media%20pages%20are,even%20share%20lies%20and%20misinformation. Accessed 22 Nov. 2024.
11. Kumar, K. Yashwanth, and B. Vani. “An Optimal Approach for Fraud Detection by Comparing Random Forest Algorithm and Support Vector Machine Algorithm for Credit Card Transaction with Improved Accuracy.” *AIP Publishing*, 21 Nov. 2023, pubs.aip.org/aip/acp/article-abstract/2821/1/070028/2922723/An-optimal-approach-for-fraud-detection-by?redirectedFrom=fulltext. Accessed 22 Nov. 2024.
12. Pellegrino, Alfonso, et al. “The Dark Side of Social Media: Content Effects on the Relationship between Materialism and Consumption Behaviors.” *Frontiers in Psychology*, U.S. National Library of Medicine, 28 Apr. 2022, pmc.ncbi.nlm.nih.gov/articles/PMC9096894/#:~:text=More%20aggressive%20and%20tailored%20messages,credit%20card%20over%20usage%20which. Accessed 22 Nov. 2024.
13. Santilli, Paul. *The Evolution of Social Media Algorithms*, 1AD, www.scip.org/news/661348/The-Evolution-of-Social-Media-Algorithms-From-Chronological-to-Intelligent-Feeds-.htm. Accessed 22 Nov. 2024.

14. Simon, Felix M., et al. “Misinformation Reloaded? Fears about the Impact of Generative AI on Misinformation Are Overblown: HKS Misinformation Review.” *Misinformation Review*, 22 Nov. 2023, misinforeview.hks.harvard.edu/article/misinformation-reloaded-fears-about-the-impact-of-generative-ai-on-misinformation-are-overblown/. Accessed 22 Nov. 2024.
15. Stray, Jonathan. “Manipulated Media Detection Requires More than Tools: Community Insights on What’s Needed.” *Partnership on AI*, 25 Jan. 2022, partnershiponai.org/manipulated-media-detection-requires-more-than-tools-community-insights-on-whats-needed/. Accessed 22 Nov. 2024.
16. Uppada, Santosh Kumar, et al. “Novel Approaches to Fake News and Fake Account Detection in Osns: User Social Engagement and Visual Content Centric Model.” *Social Network Analysis and Mining*, U.S. National Library of Medicine, 10 May 2022, pmc.ncbi.nlm.nih.gov/articles/PMC9089299/. Accessed 22 Nov. 2024.
17. Wang, Binghui, et al. *Detecting Fraudulent Users in Online Social Networks Via ...*, 2017, home.engineering.iastate.edu/~neilgong/papers/GANG.pdf. Accessed 22 Nov. 2024.
18. Wang, Huan. “Research on the Application of Random Forest-Based Feature Selection Algorithm in Data Mining Experiments.” *International Journal of Advanced Computer Science and Applications (IJACSA)*, The Science and Information (SAI) Organization Limited, Jan. 2023, thesai.org/Publications/ViewPaper?Volume=14&Issue=10&Code=IJACSA&SerialNo=54. Accessed 22 Nov. 2024.
19. Whitfield, Brennan. “Random Forest: A Complete Guide for Machine Learning.” *Built In*, 8 Mar. 2024, builtin.com/data-science/random-forest-algorithm. Accessed 22 Nov. 2024.

20. Yadav, Amit, et al. *Instagram Fake Profile Detection - A Review*, 7 July 2023, www.ijnrd.org/papers/IJNRD2307303.pdf. Accessed 22 Nov. 2024.

21. Zhao, Yunpeng, et al. "Data and Model Biases in Social Media Analyses: A Case Study of Covid-19 Tweets." *AMIA ... Annual Symposium Proceedings. AMIA Symposium*, U.S. National Library of Medicine, 21 Feb. 2022, pmc.ncbi.nlm.nih.gov/articles/PMC8861742/. Accessed 22 Nov. 2024.