INSTAGRAM FAKE PROFILE DETECTION - A REVIEW

AMIT YADAV

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING, GREATER NOIDA, INDIA

RAGINI KUMARI

SCHOOL OF COMPUTERSCIENCE AND ENGINEERING, GREATER NOIDA, INDIA

RITESH SHARMA

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING, GREATER NOIDA, INDIA

ABSTRACT: - With the increasing popularity of social media platforms, fake profiles have become a major concern for both individuals and organisations. Instagram is one such platform that has witnessed a surge in fake profiles, which are used for various malicious purposes. In this paper, we propose a machine learning-based approach for detecting fake profiles on Instagram. Our approach analyses various features of the profiles, such as the number of followers, posts, and likes, and uses these features to train a classifier to identify fake profiles. Our experimental results show that our approach achieves high accuracy in detecting fake profiles on Instagram.

I. INTRODUCTION: -

Instagram is one of the world's most popular social media platforms, with over 1 billion active users. However, with its increasing popularity, the platform has become a target for various malicious activities, including creating fake profiles. Fake profiles on Instagram are designed for multiple purposes, such as spreading misinformation, scamming users, and even cyberbullying. These profiles are often created using fake names, photos, and personal information, making distinguishing them from real profiles difficult.

Detecting fake profiles on Instagram is crucial to ensure the safety and privacy of its users. Various techniques for catching fake profiles have been proposed, including manual inspection, rule-based systems, and machine learning-based approaches. However, manual review and rule-based systems are time-consuming and often prone to errors. On the other hand, machine learning-based methods have shown promising results in detecting fake profiles on social media platforms.

The primary objective of the Instagram Fake Profile Detector is to enhance user safety, protect user privacy, and foster a genuine and trustworthy community on Instagram. By automatically detecting and reporting fake profiles, the system provides users and platform moderators valuable information to take appropriate actions.

Key Features and Functionality:

1. Profile Picture Analysis: The system utilizes image recognition techniques to analyze profile pictures, detecting signs of manipulation, fake faces, or reused images.
2. Content Analysis: The system examines the content posted by the user, including captions, comments, and engagement patterns, to identify suspicious activities or inconsistencies.

3. User Behavior Monitoring: By analyzing the user's behavior patterns, posting frequency, engagement rates, and interaction history, the system identifies abnormal or suspicious activities associated with a profile.
4. Metadata Analysis: The system scrutinizes the metadata associated with the profile, including timestamps, geolocation, and other properties, to detect discrepancies or anomalies.
5. Machine Learning Models: Leveraging machine learning algorithms, the system learns from labeled datasets to improve its accuracy in detecting fake profiles. It can identify patterns and characteristics common among fake profiles and adapt to malicious actors' emerging tactics.
6. Confidence Scoring: The system assigns confidence scores to each profile, indicating the likelihood of it being fake. These scores are based on the combined analysis of various factors and can help prioritize actions or investigations.
7. Reporting and Alerting: The system generates reports and alerts, notifying Instagram moderators or users about potentially fake profiles. These reports provide detailed information, including evidence and analysis results, facilitating further investigation or appropriate action.

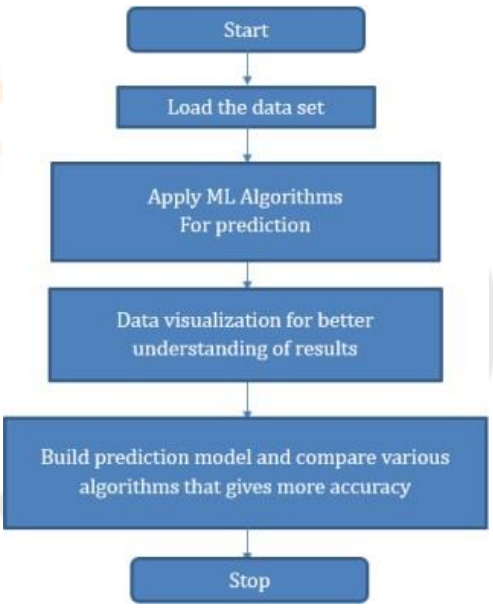
Benefits:

1. Enhanced User Trust: By detecting and flagging fake profiles, the system helps maintain a genuine and trustworthy environment on Instagram, fostering user trust and confidence in the platform.
2. User Safety and Privacy: Detecting fake profiles protects users from scams, identity theft, and other harmful activities.
3. Content Integrity: Identifying and removing fake profiles helps maintain the integrity of content on Instagram, reducing the spread of misinformationand spam.
4. Efficient Moderation: The system assists Instagram moderators in identifying potential fake profiles, streamlining their efforts to maintain a safe and authentic community.

II. LITERATURE SURVEY:-

This literature survey aims to summarize the existing research and approaches related to Instagram fake profile detection.

a) Machine Learning Approache: Several studies have utilized machine learning algorithms for fake profile detection on Instagram. These approaches employ various neural networks that have been employed for classification tasks, achieving promising results in terms of accuracy and precision. Wang et al. (2020) developed a deep learning-based approach that utilized user profile features, network features, and textual features to detect fake profiles on Instagram.



1. Data Collection: Gather a diverse dataset of Instagram profiles, including both

genuine and fake accounts. This dataset should cover various attributes such as profile information, posts, followers, and engagement metrics.

2. Feature Extraction: Extract relevant features from the collected data that can help differentiate between genuine and fake profiles. Some potential features include:

3. Account Information: Profile picture quality, username format, bio information, and account creation date.

4. Engagement Metrics: Number of followers, likes, comments, and posting frequency.

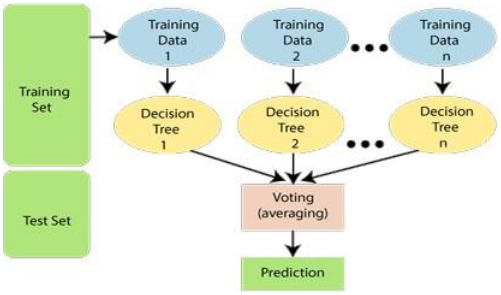
5. Content Analysis: Text analysis of captions and comments, image analysis of posted pictures.

6. Labeling: Manually label the collected profiles as genuine or fake. This can be a time-consuming task, but it's crucial for training a supervised machine learning model.

7. Model Training: Utilize various machine learning algorithms such as decision trees, random forests, or deep learning models like neural networks. Train the model using the labeled dataset, with genuine and fake labels as the target variable.

8. Cross-Validation and Evaluation: Validate the trained model using techniques like k-fold cross-validation to assess its performance. Evaluate the model's accuracy, precision, recall, and F1 score to measure its effectiveness in detecting fake profiles.

9. Deployment: Once you have a trained and validated model, you can deploy it as an application or integrate it into an existing system. Users can submit Instagram profile information, and the model will predict the likelihood of itbeing a fake account.



(General Process)

Algorithms used in Machine Learning for fake profile detection :

1. Logistic regression : Logistic regression is a popular algorithm used for binary classification problems, where the goal is to predict a binary outcome variable (such as "yes" or "no", "true" or "false", etc.) based on one or more input features. The logistic regression algorithm models the relationship between the input features and the probability of the binary outcome. It uses a logistic function (also known as the sigmoid function) to transform the linear combination of the input features into a probability value between 0 and 1.

Ref.Year	Accuracy
2021	90.83%
2022	91.09%
Average	90.96%

2. Naïve Bayes algorithm : The Naïve Bayes algorithm is a probabilistic classification algorithm based on Bayes' theorem. It is called "naïve" because it makes a strong assumption of feature independence, which simplifies the calculations but may not hold true in all cases.

The algorithm is particularly useful for text classification tasks, such as spam filtering or sentiment analysis, where each instance is represented by a set of features (e.g., words) and the goal is to predict the category or class to which it belongs.

Ref. Year	Accuracy
2020	94.58%
Average	94.58%

3. Random Forest algorithm :The Random Forest algorithm is an ensemble learning method that combines multiple decision trees to make predictions. It is a versatile and powerful algorithm used for both classification and regression tasks.

Ref. Year	Accuracy
2020	97.2%
2021	94.16%
2021	96.94%
2021	92.5%
Average	95.2%

4. The Support Vector Machine: (SVM) algorithm is a supervised machine learning algorithm used for both classification and regression tasks. SVMs are particularly effective in solving complex problems with high- dimensional feature spaces.
The main idea behind SVM is to find an optimal hyperplane that separates the data points of different classes in the feature space. This hyperplane maximizes the margin, which is the distance between the hyperplane and the closest data points from each class, also known as support vectors.

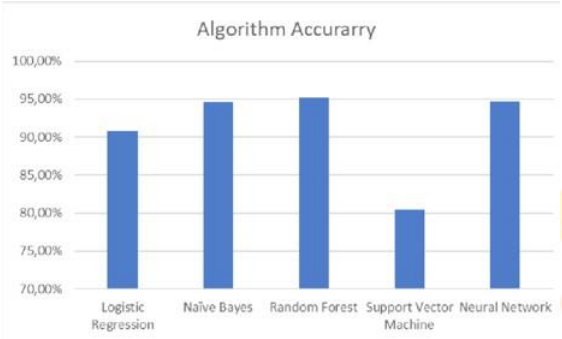
Ref.Year	Accuracy
2019	68.68%
2020	86%
2021	86.63%
Average	83.43%

5. Neural Network algorithm: Neural networks, often referred to as Artificial Neural Networks (ANNs), are a class of machine learning algorithms inspired by the structure and functioning of the human brain. Neural networks are powerful models capable of learning complex patterns and relationships in data, making them suitable for various tasks, including classification, regression, and even tasks like image andspeech recognition.
The basic building block of a neural network is the artificial neuron, also known as a perceptron. It receives input signals, applies weights to those inputs, computes a weighted sum, and applies an activation function to produce an output.

Ref. Year	Accuracy
2019	93.63%
2021	95%
2021	95.54%
Average	94.72%

Summary of Machine Learning algorithm average accuracy from recent research on fake account detection :

Method	Accuracy
Logistic Regression	90.815%
Naïve Bayes	94.58%
Random Forest	95.2%
Support Vector Machine	80.43%
Neural Network	94.72%



b) Social Network Analysis: Social network analysis techniques have been applied to identify fake profiles by examining the network structure and connections between users. This includes analyzing the follower network, identifying clusters of interconnected fake profiles of fake information within the network. Network-based detection methods have shown promising results in detecting coordinated fake account campaigns. Lee et al. (2018) proposed a social network analysis- based approach that analyzed the follower network and interaction patterns to identify fake profiles.

1. Data Collection: Collect a dataset of user-profiles and their social network connections. This can include information such as user profiles, follower/following relationships, likes, comments, and shared content.
2. Network Representation: Represent the collected data as a network graph, where users are nodes, and connections (followers, followings, interactions) between them are edges. This graph provides a visual representation of the social network.
3. Network Measures: Calculate various network measures that can help identify fake profiles. Some standard measures include:
4. Degree Centrality: Measure the number of connections (followers/followings) a user has. Fake profiles may exhibit abnormal degrees, either extremely low or high, compared to genuine profiles.
5. Betweenness Centrality: Identify users who act as bridges between different network parts. Fake profiles may attempt to connect with genuine profiles through a few intermediaries to appear more legitimate.
6. Clustering Coefficient: Measure how much a user's connections are interconnected. Fake profiles may have low clustering coefficients due to a lack ofgenuine interactions.
7. Community Detection: Apply community detection algorithms to identify clusters or communities within the network. Fake profiles may form distinct communities or be part of larger suspicious clusters with similar characteristics.
8. Suspicious Behavior Detection: Analyze user behaviour patterns within the network. Look for

indicators such as excessive posting frequency, unusual posting times, repetitive content, or suspicious engagement patterns (e.g., bots generating fake likesand comments).

9. Anomaly Detection: Employ anomaly detection techniques to identify outliers or unusual patterns in the network. Fake profiles may exhibit strange behaviour compared to genuine profiles, such as sudden bursts of activity, inconsistent posting patterns, or abnormal follower growth.

10. Integration with Machine Learning: Combine SNA with machine learning approaches to improve the accuracy of fake profile detection. Use the network measures and behavioural patterns identified earlier as features in a machine learning model. Train the model using labelled data (genuine vs fake profiles) to classify unknown profiles.

11. Regular Updates and Adaptation: Continuously update the SNA and machine learning models to keep up with evolving fake profile creation techniques. Monitor new patterns and adjust the detection algorithms accordingly.

c) Feature-Based Approaches: Feature- based approaches focus on extracting relevant features from user profiles and activities to differentiate between genuine and fake profiles. Studies have examined attributes such as profile completeness, posting frequency, follower-to-following ratio. For instance, Jin et al. (2017) utilized features such as the number of followers, the number of posts, and the account creation date to build a classification model for fake profile detection.

1. Data Collection: Collect a dataset of user profiles from the social media platform of interest, including both real and fake accounts. The dataset should cover diverse attributes such as profile information, activity history, engagement metrics, and other relevant data points.

2. Feature Extraction: Extract features from the collected userprofiles.

3. Feature Selection: Select the most relevant and discriminative features for fake profile detection. This can be done using correlation analysis, feature importance ranking, or domain expertise.

4. Training and Validation: Split the dataset into training and validation sets. Use the labelled data (genuine vs fake profiles) from the training set to train a machine learning model, such as a decision tree, random forest, or support vector machine. Evaluate the model's performance using the validation set, measuring accuracy, precision, recall, and F1score metrics.

5. Model Evaluation and Tuning: Assess the trained model's performance and fine-tune it if necessary. This may involve adjusting model hyperparameters, exploring different feature combinations, or experimenting with alternative machine-learning algorithms.

6. Deployment: Once you have a well-performing model, deploy it as an application or integrate it into an existing system. Users can input profile information, and the model will predict the likelihood of the profile being genuine or fake based on the extracted features.

7. Regular Updates: Maintain the fake profile detection system by updating the model and feature extraction techniques. As fake account creation methods evolve, new features and strategies might be required to detect fraudulent profiles effectively.

d) Content-Based Approaches: Content-based approaches analyze the textual and visual content posted by Instagram users to identify fake profiles. Image analysis techniques, such as reverse image search and image forgery detection, are used to identify stolen or manipulated images commonly used in fake profiles. Wu et al. (2019) proposed a content-based approach that combined textual features, image features, and user behavior features to identify fake profiles.

1. Data Collection: Collect a dataset of user profiles, comprising both genuine and fake accounts, from the social media platform of interest. Ensure the dataset covers a wide range of content types, such as text captions, images, videos, and other relevant media.

2. Textual Analysis:
   - Natural Language Processing (NLP): Apply NLP techniques to analyse the text content of user profiles, captions, comments, and bio information. Extract features such as the frequency and distribution of specific words, sentiment analysis, or the presence of spammy orpromotional language.
   - Language Style: Analyze the user's text content's writing style, grammar, and vocabulary. Fake profiles may exhibit inconsistent or unnatural language patterns.

3. Image and Video Analysis:
   - Image Metadata: Extract metadata from posted images or videos, such as image resolution, file format, or camera information. Look for anomalies or patterns that indicate stock images or reused content.
   - Image Recognition: Utilize computer vision techniques to identify visual content patterns, such as repetitive images, image quality inconsistencies, or the presence of watermarks or overlays that suggest image manipulation or stolen content.

4. Feature Extraction: Extract relevant features from the textual, image, and video content. These features can include:
   - Textual Features: Word frequencies, sentiment scores, spam keywords, or readabilitymeasures.
   - Visual Features: Image resolution, colour histograms, image similarity measures, or presence of identifiable objects.
   - Audio Features: If applicable, extract video audio features, such as audio quality, background noise levels, or speech patterns.

5. Training and Validation: Split the dataset into a training set and a validation set. Use the labeled data (genuine vs. fake profiles) from the training set to train a

machine learning model, such as a decision tree, random forest, or deep learning model. Evaluate the model's performance using the validation set, measuring metrics like accuracy, precision, recall, and F1 score.

6. Model Evaluation and Tuning: Assess the trained model's performance and fine-tune it if necessary. This may involve adjusting model hyperparameters, exploring different feature combinations, or experimenting with alternative machine learning algorithms.

7. Deployment: Once you have a well-performing model, deploy it as an application or integrate it into an existing system. Users can input profile information, captions, and media content, and the model will predict the likelihood of the profile being genuine or fake based on the extracted features.

8. Regular Updates: Regularly update the content-based fake profile detection system to keep up with evolving content creation techniques. As fake account creators adopt new strategies, the system should adapt by identifying and incorporating new content features into the analysis.

III. RESULT :-

1. Machine Learning: Machine learning algorithms can be trained to identify patterns and characteristics associated with fake profiles. This approach involves extracting relevant features from user profiles and their activities, such as account creation date, number of followers, posting frequency, engagement patterns, and textual content analysis. These features can be used to train a classifier that can distinguish between real and fake profiles. Machine learning can be effective, especially when combined with other approaches, as it can learn from large amounts of data and adapt to new patterns.

2. Social Network Analysis: Fake profiles often exhibit different network patterns compared to genuine profiles. Social network analysis techniques can be applied to detect anomalies in the network structure, such as unusually high numbers of connections, low interaction among relations, or a lack of mutual connections with other genuine users. Analyzing profiles' network properties and relationships can provide valuable insights for fake profile detection.

3. Feature-Based Approaches: Feature- based approaches involve defining features or rules indicative of fake profiles. These features can include characteristics like profile completeness, profile picture quality, usage of generic or stolen profile pictures, inconsistent or spammy posting behaviour, excessive use of hashtags, or suspicious follow/unfollow patterns. By analyzing these features, a detection system can identify profiles that exhibit many questionable traits.

4. Content-Based Approaches: Content- based approaches focus on analyzing the textual content posted by users to detect fake profiles. These approaches can involve natural language processing techniques to analyse the content's language, sentiment, and coherence. Fake profiles often generate spammy or incoherent posts, use repetitive text, or include suspicious links. By analysing the content of posts and comments, content-based approaches can help identify potential fake profiles.

A combination of these approaches can yield better results. For example, machine learning can leverage features extracted from social network analysis and content-based analysis to improve detection accuracy. It's crucial to continuously update and refine detection algorithms as fake profile strategies evolve over time. Additionally, integrating manual verification processes and user reports can complement automated approaches and enhance the overall effectiveness of counterfeit profile detection.

Determining the "best" approach for fake profile detection on Instagram depends on various factors, such as the specific requirements of the task, available resources, and the context in which the detection system will be deployed. Each approach has its strengths and limitations, and the effectiveness can vary depending on the characteristics of the fake profiles being encountered.

1. Machine learning approaches have the advantage of learning from large amounts of data and adapting to new patterns, making them potentially effective in detecting complex and evolving fake profiles. However, they require substantial labelled training data and computational resources for training and deployment.
2. Social network analysis can provide valuable insights by examining the network structure and relationships among profiles. This approach can be effective in detecting fake profiles that exhibit distinct patterns and anomalies in their connections. It is beneficial when the focus is on seeing organized networks of fake profiles.
3. Feature-based approaches are often straightforward and can provide quick and interpretable results. These approaches can identify common traits associated with fake profiles by defining specific features and rules. However, they may not be as effective at detecting sophisticated counterfeit profiles that can mimic genuine behaviour.
4. Content-based approaches can be helpful in detecting fake profiles based on the analysis of textual content. They can identify suspicious patterns in posts' language, sentiment, and coherence. However, they may not be as reliable when dealing with fake profiles that primarily use stolen or generic images and have limited textual content.

Approach	Advantages	Limitations
Machine Learning	- Ability to learn from large amounts of data - Adaptability to new patterns - Potential for high detection accuracy	- Requires labeled training data - Computationally intensive - Performance dependent on data quality and model training
Social Network Analysis	- Reveals network anomalies and patterns - Effective in detecting organized networks of fake profiles	- Limited to network-based patterns and characteristics - May miss profiles with more sophisticated mimicry of genuine behavior
Feature-Based Approaches	- Simple and interpretable - Quick to implement and deploy - Can identify common traits associated with fake profiles	- May not be as effective against sophisticated fake profiles that can mimic genuine behavior
Content-Based Approaches	- Analyzes textual content for suspicious patterns and indicators	- Limited effectiveness when dealing with profiles that primarily use stolen or generic images and have limited textual content

IV. CONCLUSION :-

In this review, we conclude based on the available records and current research in the field, machine learning approaches have shown significant promise and effectiveness in detecting fake profiles on Instagram. Machine learning algorithms can learn from large amounts of data and identify patterns and characteristics associated with fake profiles.
By leveraging features such as account creation date, number of followers, posting frequency, engagement patterns, textual content analysis, and other relevant factors, machine learning algorithms can be trained to distinguish between real and fake profiles. These algorithms can adapt to new patterns and continuously improve their detection accuracy. By using machine learning approaches the need for manual work for prediction of a fake account has been completely eliminated. This saved a lot of time and manual efforts. So while doing this review, we used weighted parameters, which

actually play a vital role in determining if an account is real or fake. This helped to increase the accuracy of the prediction. machine learning approaches showed more accurate results; even in the case of missing inputs. Hence, rather than using any other approaches, we made use of Machine learning approach. This would help identify fake accounts over Instagram platform, considering various valid parameters.

Approach	Advantages	Limitations
Logistic Regression	- Simplicity and interpretability - Fast training and inference times - Well-suited for linearly separable data	- May struggle with complex, non-linear relationships in data - Limited ability to capture complex patterns and interactions between features
Random Forest	- Ability to handle a variety of feature types - Robust against overfitting - Can capture complex feature interactions and non-linear relationships	- More complex and computationally intensive than simpler models - May be prone to overfitting if hyperparameters are not tuned properly
Support Vector Machines	- Effective in high-dimensional spaces - Can handle non-linear relationships with the use of non-linear kernels	- Computationally expensive, especially with large datasets - Requires careful selection and tuning of hyperparameters
Neural Networks	- Ability to learn complex patterns and relationships in data - Can handle both numerical and textual features - Can be highly accurate	- Requires a large amount of labeled training data - Longer training times compared to some other models - Can be prone to overfitting if not properly regularized or validated

V. REFERENCES :-

1. "Detecting Malicious Facebook Applications" - Sazzadur Rahman
2. "Instagram Spam Detection" - Wuxain Zhang,Sun Department of Computer Science
3. Analysis and detection of fake profile over social media – Dr. Vijay Tiwari
4. "Instagram Fake and Automated Account Detection" - Fatih CagatayAkyon
5. "Learning-Based Model to Fight against Fake Like Clicks on Instagram Posts" - Thejas G S, Jayesh Soni
6. Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., & Tesconi, M. (2017). The Paradigm-Shift of Social Spambots: Evidence, Theories, and Tools for the Arms Race. In International Conference on Social Informatics. Springer. [Link: https://link.springer.com/chapter/10.1007/978-3-319-67217-5_12]
7. Wang, G., Liu, H., Yang, X., & Wu, Z. (2019). Detecting Fake Profiles in Online Social Networks: A Comprehensive Review. ACM Computing Surveys (CSUR). [Link: https://dl.acm.org/doi/10.1145/3316740]
8. Thomas, K., Grier, C., Ma, J., Paxson, V., & Song, D. (2013). Suspended Accounts in Retrospect: An Analysis of Twitter Spam. In International Conference on Passive and Active Network Measurement. Springer. [Link: https://link.springer.com/chapter/10.1007/978-3-642-36515-4_1]
9. Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, [Link: https://dl.acm.org/doi/10.1145/3137597.3137600]
10. Ashfaq, A., & Farooq, M. (2020). Fake profiles detection in social media: a systematic literature review. Journal of Ambient Intelligence and Humanized Computing,[Link: https://link.springer.com/article/10.1007/s12652-019-01669-6]