https://ieeexplore.ieee.org/document/10181604

https://dl.acm.org/doi/10.1145/2659651.2659736

https://www.researchgate.net/publication/339170645_Malicious_Account_Detection_on_Twitter_Based_on_Tweet_Account_Features_using_Machine_Learning

https://www.sciencedirect.com/science/article/pii/S1568494621002830

https://www.mdpi.com/2076-3417/10/22/8160

Related Work:
Supervised machine learning - feature extraction, traditional machine learning classifiers
Two feature vectors - Global Outlier Standard Score and Local Outlier Standard Score.
Machine learning algorithm - K-NN, Decision Tree, Gaussian Naive Bayes, Support Vector Machine, and Random Forest.
Bot-or-Not's classification system -
Supervised machine learning - Deep Neural Networks
CNN for detecting spams on Twitter, text based classifier and combined classifier to calculate F1 Score. Feature extraction to group the dataset into three categories: tweet-based, user profile-based, and social graph-based features.
Graph convolutional networks (GCNs)
Graph Attention network
GraphSAGE
Graph isomorphic network
Feature-based detection and propagation-based detection, it applies machine learning or graph mining algorithms to identify malicious accounts on social networks.
Semi-supervised graph - based on graph attention network for spam bot detection
Advertising, malicious links and fake news are part of malicious activities
Spambot detection: feature based method and propagation based method
MRF model on the similarity graph

Literature Review: Supervised Learning and GNN

Bot detection on social media platforms has become a critical area of research as malicious bots are increasingly used for spreading misinformation, advertising, and other nefarious activities. Various machine learning approaches have been applied to this problem, ranging from traditional supervised classifiers to cutting-edge graph-based methods. This review provides an overview of existing work, focusing on both supervised learning and advanced graph-based techniques for bot detection.

1. Supervised Machine Learning for Bot Detection

A significant portion of early bot detection research has utilized **supervised machine learning** techniques. These methods often rely on **feature extraction** from user data, tweet content, and network activity to classify accounts as bots or legitimate users. The feature sets used typically fall into three categories: **tweet-based features** (e.g., the content of the tweet), **user profile-based features** (e.g., account age, number of followers), and **social graph-based features** (e.g., follower network structure).

Several traditional machine learning algorithms have been widely used in this domain, including:

- **K-Nearest Neighbors (K-NN)**
- **Decision Trees**
- **Gaussian Naive Bayes**
- **Support Vector Machines (SVMs)**
- **Random Forests**

In these approaches, **two feature vectors**—the **Global Outlier Standard Score** and **Local Outlier Standard Score**—are often employed to identify anomalous user behaviors, both at a global and local level. These scores are particularly useful for detecting outliers that might signal the presence of bot activity.

Bot-or-Not's classification system is one example of a well-known bot detection system, which applies supervised learning techniques, including **Deep Neural Networks (DNNs)**. Bot-or-Not uses **Convolutional Neural Networks (CNNs)** for detecting spam on Twitter, primarily relying on **text-based classifiers**. By extracting features from tweets and combining them with user profile and social graph features, this system can achieve high classification accuracy, as measured by the **F1 Score**.

2. Graph-Based Machine Learning for Bot Detection

While traditional machine learning methods have shown effectiveness, they often fail to capture the complex relationships between users in a social network. This has led to the development of

graph-based learning approaches, which are better suited to detecting bots in a networked environment.

Some of the most prominent graph-based approaches include:

- **Graph Convolutional Networks (GCNs)**: GCNs have been used to classify users in a social graph by aggregating information from their neighbors. This method allows the model to consider the network structure and connections between users when making predictions.
- **Graph Attention Networks (GATs)**: GATs improve on GCNs by incorporating attention mechanisms, allowing the model to focus more on the most relevant neighbors in the graph. This has proven useful in spam detection, where a bot's neighbors may provide important clues about its behavior.
- **GraphSAGE**: Unlike traditional GCNs, which rely on pre-computed node features, GraphSAGE generates embeddings by sampling and aggregating features from a node's local neighborhood. This approach allows for more scalable and efficient bot detection in large social networks.
- **Graph Isomorphic Networks (GINs)**: GINs are designed to capture more complex structural patterns in graphs and can distinguish between different graph structures more effectively than GCNs or GraphSAGE. This makes GINs suitable for identifying bots that may exhibit sophisticated or unusual connection patterns.

3. Feature-Based and Propagation-Based Bot Detection

Beyond standard graph-based learning techniques, bot detection has also employed **feature-based** and **propagation-based methods**. These approaches aim to leverage both user characteristics (features) and the way that information spreads across a social network.

In **feature-based methods**, machine learning or graph mining algorithms are applied to identify bots based on characteristics such as user activity, tweet content, and profile information. Propagation-based methods, on the other hand, focus on how influence and information spread through the network. This method is especially useful for detecting **spam bots**, which often coordinate to amplify certain messages or content.

One notable approach uses **Markov Random Fields (MRF)** on similarity graphs to detect spambots. In these methods, the bot detection problem is modeled as a similarity graph, where edges represent relationships between users. The MRF model helps propagate labels (e.g., bot or non-bot) through the graph based on the similarity of connected nodes.

4. Semi-Supervised Learning and Malicious Bot Detection

In more recent work, **semi-supervised learning** techniques, particularly those based on **Graph Attention Networks (GATs)**, have been applied to **spam bot detection**. These models are trained using a combination of labeled and unlabeled data, making them effective in situations where ground truth labels are scarce.

An important application of these methods is in the detection of bots involved in **malicious activities**, such as spreading **advertisements**, **malicious links**, and **fake news**. By focusing on the propagation of harmful content and the relationships between bots and legitimate users, these models can identify not only the presence of bots but also whether they are engaged in harmful or benign activities.

5. Integration of Outlier Detection and Graph-Based Approaches

A promising avenue for bot detection involves the integration of **outlier detection methods** with **graph-based approaches**. By incorporating **Global and Local Outlier Standard Scores** as node features in graph models, researchers can enhance the model's ability to detect **anomalous behaviors** that signal bot activity. This fusion of traditional machine learning techniques with graph neural networks is still an emerging field, with significant potential for improving bot detection accuracy.

Using Propagation-Based Detection in Graph Neural Networks:

- **Existing Work**: Propagation-based detection methods, such as those using **Markov Random Fields (MRFs)**, have been used to identify malicious accounts based on the **spread of influence** or interactions in a network. These are typically feature-based or semi-supervised graph models.
- **Novel Contribution**: You can propose a new GNN architecture that leverages **propagation-based methods like MRF models** to detect spambot behavior. Specifically, a **Graph Convolutional Network (GCN)** or **GraphSAGE** could be enhanced with a **propagation layer** that models how influence spreads through a network, highlighting **coordinated bot activity**. In this way, GNNs would be tuned to **learn not just from local neighbors but from influence propagation across the graph**.
    - **Why novel**: Combining **graph mining methods** with **neural network propagation mechanisms** in this way could lead to new insights, allowing the model to detect bots based on **how information or influence spreads** through bot networks.

Attention-Based Semi-Supervised Learning for Malicious Bots:

- **Existing Work**: There has been prior work using **Graph Attention Networks (GATs)** in semi-supervised spam detection on social networks, but these methods often focus solely on bot detection, without addressing the different types of bots (benign vs. malicious).
- **Novel Contribution**: Extend existing methods by proposing a **GAT-based semi-supervised model** that not only detects bots but also **classifies them as benign or malicious**. This could involve training the GAT to pay attention to specific structural or behavioral signals, like those linked to **advertising, malicious links, or fake news**, which are common among malicious bots. By focusing the attention mechanism on certain high-risk signals, the GNN would become more effective at identifying malicious behavior in real time.
    - **Why novel**: Introducing an **attention mechanism** that is fine-tuned to spot **specific types of malicious activity** (like malicious links, fake news, etc.) provides an added layer of functionality to bot detection systems, moving beyond binary bot identification toward more granular classification.