Distinguishing between Benign Viral Content and Malicious Content on X

Adenutsi Andy Eleos

Abstract

The rise of online communication in recent times has meant that social media platforms have become an important part of communication. These platforms provide an avenue for information exchange and interaction. However, they are susceptible to the spread of malicious content which can have repercussions on users on such platforms. Detecting anomalies is vital in identifying and mitigating such threats. Machine learning models that detect anomalies, usually fail to differentiate between benign anomalies and malicious anomalies. The aim of this research is to develop a model that uses the techniques of semi-supervised machine learning to accurately distinguish between benign and malicious anomalies on X. The approach is driven by the need for verifiable and accurate information, as well as the protection of users from potential harm on social platforms. In this study, we aim to improve on anomaly detection and ensure a safe online environment.

Keywords

Anomaly Detection, Social Networks, Benign Anomalies, Malicious Anomalies, Viral Content, Semi-Supervised, Machine Learning.

Introduction

Social media networks have made communication and sharing information easier, creating an interconnected global community. However, the open nature of these platforms makes it easier for the dissemination of malicious content such fake news and scams, sexual predators, online fraudsters(Savage et al., 2016). This is a vulnerability of most social media platforms posing a challenge to integrity of social networks and user trust.

Anomalies simply means deviation from normal patterns of behavior. A research work describes anomalies as observations that appear to be inconsistent with the remainder of the set of data and stem from the fact that the magnitude of the deviation will depend on the specific problem domain(Savage et al., 2016). The phrase “with the remainder of the set of data” and “domain specific problem domain” paints a picture that in anomaly detection, understanding of what

should be observed is much needed as finding anomalies varies depending on the problem one wants to solve and the dataset that is available. This knowledge will be incorporated in our work because benign anomalies and malicious anomalies exhibit very similar characteristics. An example of such trait is popularity.

While some anomalies are benign, such as viral posts that gain attention due to their popularity and novelty, others are malicious with the intent of misleading or harming users. Malicious anomalies include the spread of fake news that threaten the security of users and the integrity of social media platforms.

Traditional machine learning models have been effective in detecting these deviations, but they perform poorly when it comes to differentiating between benign anomalies, popular or viral posts and malicious ones intended to cause harm. In the context of social media, finding the difference between the two anomalies is crucial because not all viral post is harmful. The term when a machine learning model does not accurately classify an anomaly is called a false positive. A false positive is an outcome where a model incorrectly predicts a positive class or to put it in simpler terms when the model predicts a benign post as malicious or vice versa. Reducing false positives or alerts in anomaly detection has remained an unsolved problem (Omar et al., 2013). Because promoting safe and verifiable content is prioritized on social media platforms it is crucial that there is a way malicious content can be detected and mitigated, and this paper seeks to address this problem.

Incorporating insights from recent studies, like the work: classifying social media bots as malicious or benign by (Mbona & Eloff, 2023), provides a foundational understanding of behavioral patterns and the characteristics that differentiate these anomalies. Their work also shows the importance of using machine learning algorithms or techniques to improve the

accuracy of model classification. Drawing from their findings, the approach in this paper also uses behavioral indicators and post engagement analysis to refine anomaly detection.

Detecting anomalies just like all data projects go through similar stages: parameterization, training and detection (Omar et al., 2013). Parameterization involves data collection, cleaning the data and identifying the variables or features. The training stage involves feeding a model with the data to make predictions.

An anomaly model can be built mainly in two ways: Supervised, Unsupervised. Supervised Anomaly Detection requires a labelled training dataset that has both normal and anomalous samples to construct the model. However, there is a downside to Supervised Anomaly Detection models: shortage of training data, obtaining accurate labels, and noise which results in higher false alarm rates (Omar et al., 2013). Unsupervised Anomaly detection does not require training data. They are based on two basic assumptions. The first is they presume that a large group of the dataset is normal and small group is abnormal. Based on these assumptions, clustering these data groups of similar instances that appear frequently are assumed to be normal while infrequent instances are regarded malicious. The drawback of the Unsupervised technique is that it is computationally expensive. A research work: Anomaly Detection Using Data Mining Technologies in Social Networks by (Rajesh & Kiran, 2018) discussed Semi-Supervised as a class of machine learning that leverages on the strengths of the unsupervised and supervised machine learning techniques. This research work will combine all the knowledge obtained from the research work to build a model that classifies benign and malicious tweets. The rest of this paper is as follows:

Research questions

What are the characteristics that differentiate between benign and malicious anomalies on social networks?

How can machine learning techniques be effectively used to detect and classify anomalies in social media content?

What metrics should be used to evaluate the performance of the proposed model?

Methodology

- *Data Exploration*

This research utilized two separate unlabeled datasets of all trending tweets for the months of June and July. Exploratory Data Analysis was performed on these datasets to generate their summary statistics, inspecting missing values.

- *Data Preprocessing*

The two datasets were concatenated into a single dataset to ensure that our model has a lot of data to train on (32,789 observations). The data preprocessing steps include filtering for the dataset only English tweets. Irrelevant columns such as 'tweet IDs' and timestamps were dropped while keeping relevant columns such as the 'user_followers_count', 'retweet_count' and 'tweet_text' to be used as features for the model. The boxplots in figure 1, figure 2 shows the

distribution and outliers or anomalies user_followers_count and retweet_count respectively.

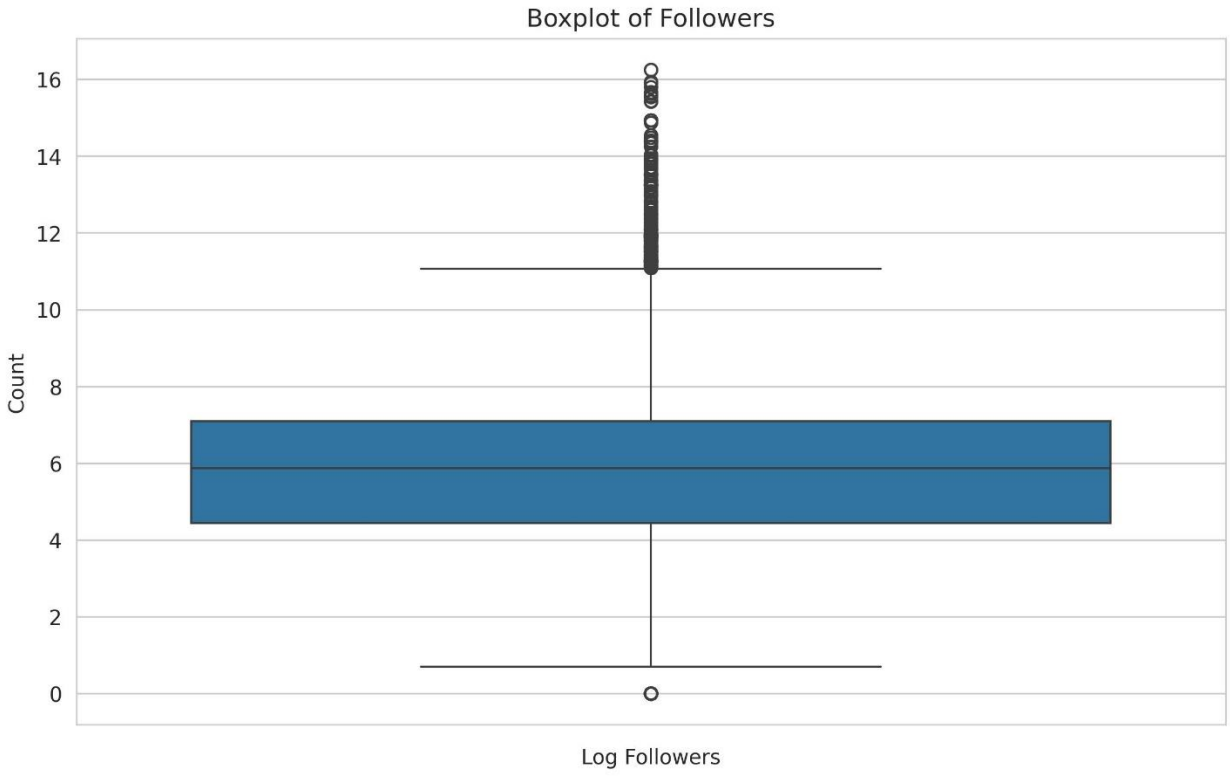

Figure 1Boxplot of Followers Count

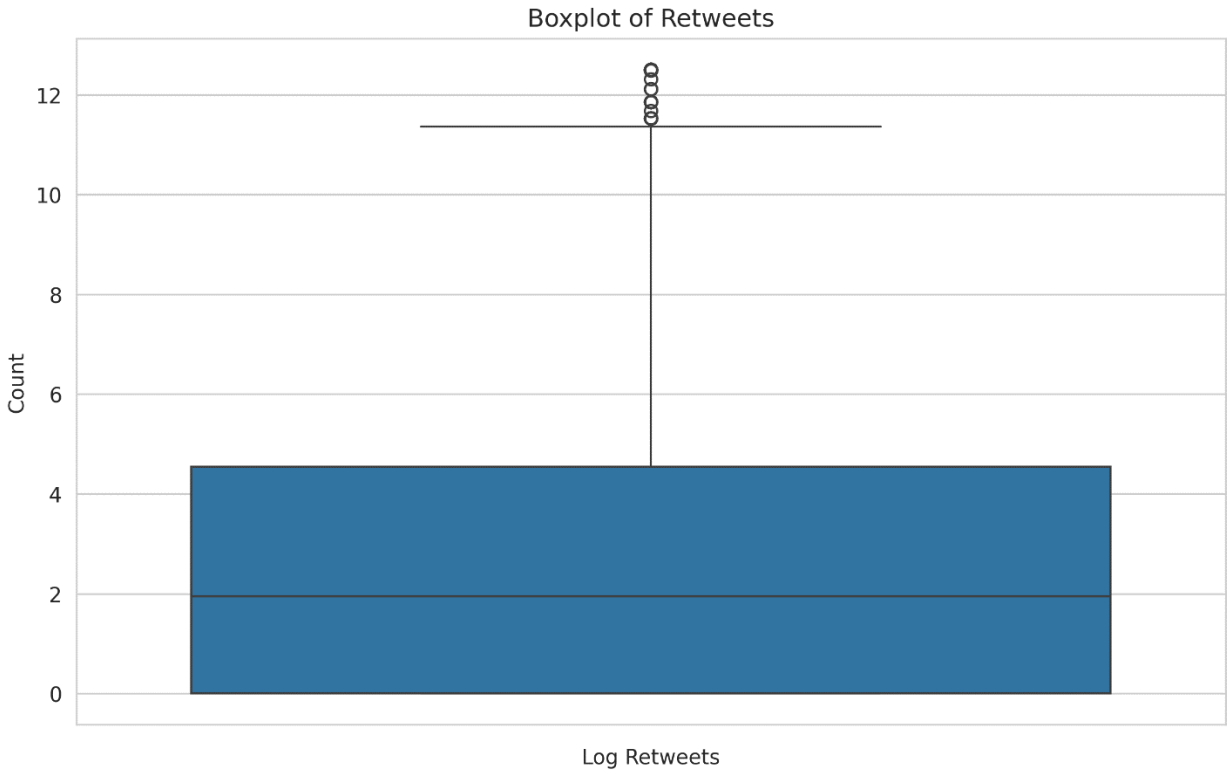

Figure 2 Boxplot of Retweet Count

Applying behavioral analysis from (Mbona & Eloff, 2023), A hypothesis was generated stating A high 'retweet_count' from an account with high 'user_followers_count' likely means the account is an influencer or a celebrity which in turn could mean a harmless or benign content. A high retweet count from an account with lower 'user_followers_count' could mean the account is a likely a bot or new account that is likely spreading misinformation. The hypothesis is later refined as it is observed that is biased favoring accounts with a high 'user_followers_count'. Another hypothesis was generated: If the sentiment is negative, the tweets is likely malicious and vice versa. Table 1 combines the two hypotheses. Text preprocessing methods- Tokenization and Lemmatization was performed on the text data to make it easier for text classification.

Table 1 Table for Tweet Classification

Sentiment	Retweet Count	User Follower Count	Label
Positive	High	High	Benign
Positive	High	Low	Malicious
Positive	Low	High	Benign
Positive	Low	Low	Benign
Neutral	High	High	Benign
Neutral	High	Low	Benign
Neutral	Low	High	Benign
Neutral	Low	Low	Benign
Negative	High	High	Malicious
Negative	High	Low	Malicious
Negative	Low	High	Benign
Negative	Low	Low	Benign

- *Feature Engineering and Unsupervised Machine Learning*
  New Features were generated to use for the model's prediction. Sentiment Analysis was conducted on the tweets categorizing each tweet as positive, negative or neutral sentiment. Since the dataset was not labelled, unsupervised learning or clustering techniques using IsolationForest model from scikit learn was applied to engineer a new feature: 'is_anomaly' to classify the tweets as an anomaly or not anomaly based on the sentiment features, retweet

count and user followers count. (Omar et al., 2013) work on unsupervised learning in the introduction of this paper, which states that unsupervised learning model presume many the observations are normal and a rather small number are the outliers. As unsupervised methods were used to cluster or group the tweets, this justifies why the benign or normal tweets were more represented than the malicious ones in our dataset in Figure 3. In this part the tweets were only clustered or grouped but have not been labeled yet.

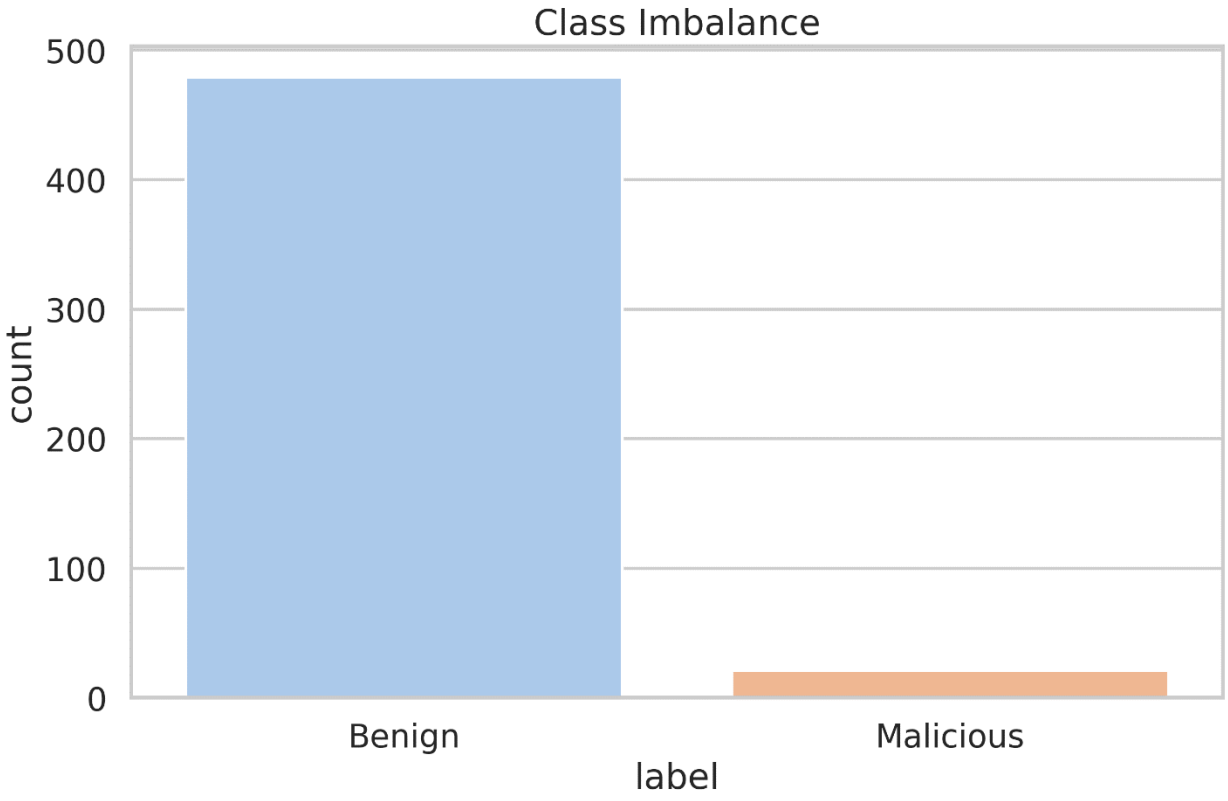

Figure 3 Distribution of Imbalanced labels

- *Semi-Supervised Machine Learning*
  Applying findings from the research work from (Omar et al., 2013). Machine Learning Techniques for Anomaly Detection, part of the unlabeled dataset was sampled (500 observations) for manual labelling. It was important to label the dataset since semi-supervised methods require a labeled training set containing both normal and anomalous

samples to construct the model (Omar et al., 2013). The tweet was labelled as either benign or malicious based on the hypothesis made in the data-preprocessing subsection of the methodology section of this research work. Since viral or anomaly tweets is the scope of this research work, the features 'is anomaly', 'sentiment', 'retweet_count', 'user_followers_count' was used to support the outcome of the labels. After classifying all the anomaly tweets, the rest of the sample dataset with was imputed with the label 'Benign' basing this on another hypothesis: if a tweet has a low retweet, that tweet is normal or the tweet is not an anomaly. The downside of Semi-supervised methods is there is shortage of train data and this training set usually contain some noise (Omar et al., 2013). Figure 3 also shows imbalance in our sample dataset which could lead to bias in the model's prediction. Figure 4 shows equal distribution of each class after the have been balanced.

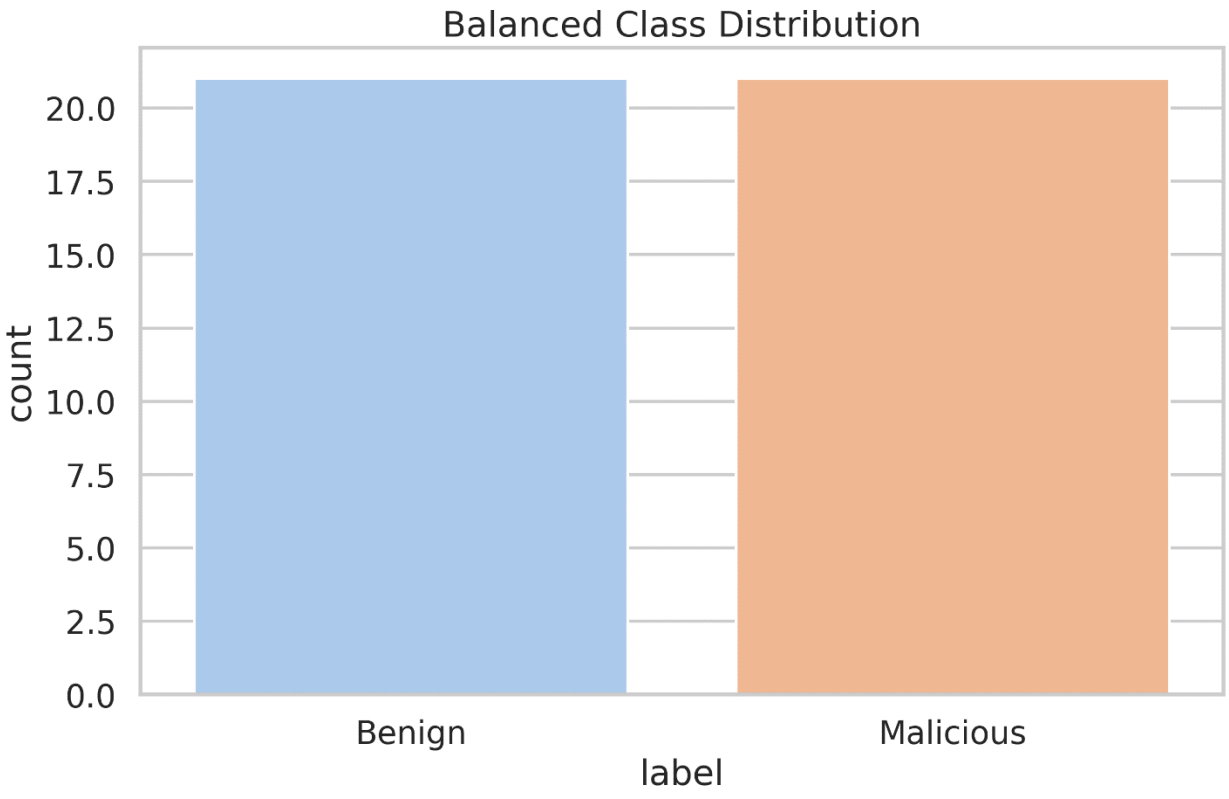

Figure 4 Distribution of Balanced Classes

Balancing the sample dataset helps each class to be equally represented which leads to fairness and the model making accurate predictions. Semi-Supervised methods make utilization of both named and unlabeled information (Rajesh & Kiran, 2018). Semi-Supervised methods use a little measure of the named information(in this case the manually labeled train set) with a lot of unlabeled information (the clustered set). Applying this knowledge the semi-supervised technique involving the use of the pretrained model, RandomForestClassifier from scikitlearn was used to predict the labels in the sample dataset and the rest of the actual dataset.

- *Results*

The metrics used for the model's evaluation were accuracy, precision, recall. The model achieved an accuracy of about 78 percent, a precision of 100 percent and a fairly good score of 66 percent in recall. Spot checks were also conducted to see if the model predicted the tweets well. The confidence interval of the model was also used to evaluate how sure the model was making these predictions. Figure 5 shows the count distribution of the predicted

classes with about 30 percent being malicious tweets.

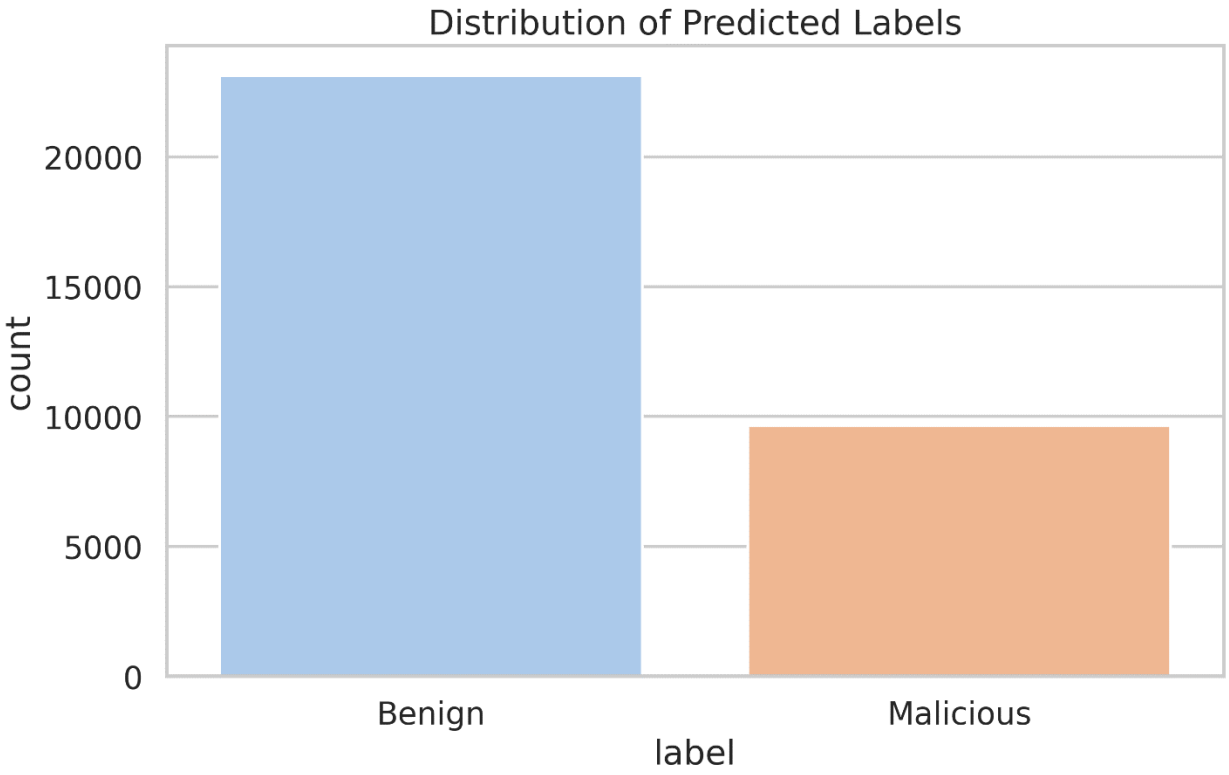

Figure 5 Distribution of Predicted Classes

Discussions

- *Comparison with Existing Research*

  Social media have become targets for malicious individuals and anomaly detection is used to identify malicious individuals (Savage et al., 2016). This research work goes in-depth with anomaly detection, distinguishing between harmless or benign anomalies and malicious anomalies specifically on social media posts. A tweet qualifies as an anomaly in this research if it has a high number of likes and a high retweet count that makes it stand out from other tweets. Such tweets are commonly referred to as trending in the context of social media. It is difficult to classify benign and malicious anomaly posts because they have similar characteristics in terms user engagement. Both have a high like count and retweet count. Drawing attention to another research work: Classifying Social Media Bots as Benign or Malicious by (Mbona & Eloff, 2023) discussed the characteristic of malicious social media bots: Malicious bots engage in activities such as spamming, dissemination of fake news, web scraping to steal user information (Mbona & Eloff, 2023). The insight gained from this paper shaped the hypothesis used in this research: For a post to be classified as benign or malicious, characteristics such as the retweet count, number of likes, tweet sentiment should be considered. The hypothesis is explained using the following scenarios: A post with high user engagement (high likes and retweet count) from an account with low followers could mean the post or content is not credible, the account could be a bot or the account engaging with such post are fake accounts promoting the post with the intent of misleading other accounts or users into thinking the post is credible. A post with high user engagement from an account with high followers could mean the account is likely an influencer or a celebrity and could imply that such post gained popularity through its novelty. A malicious post will have

malicious URLs, fake news or spam content which depicts a negative sentiment. A benign post on the other hand could have a positive sentiment.

- *Limitations*

  There are some gaps in this research work's hypothesis. A tweet with a negative sentiment does not always mean the post is malicious. An example to back this claim: there could be a post on news about earthquakes from a credible account such as CNN or BBC which could have a negative sentiment and a high user engagement (likes and retweets), per the hypothesis this qualifies as a malicious tweet, but this is not the case as the news is actually accurate information from an authentic source. Another gap in the hypothesis is tweets with neutral sentiment are considered benign but could be a post intended to cause harm or be part of a coordinated disinformation campaign. Here is a scenario supporting this statement, A series of tweets about political issues can be posted at the same time from low-follower accounts with similar wording, generic and lack specific details. These tweets could seem benign but could be part of a disinformation campaign or could be indicative that the tweet was bot generated. Another is that a high engagement from low-follower accounts does not always default the posts as malicious. A tweet from a lesser-known account about an opinion on a topic that is already viral could have high retweets and likes count but it does not mean the account is spreading misinformation.

- *Future Work*

  Using sentiment and engagement metrics solely to classify tweets as benign or malicious may not be sufficient and accurate given the complexity of online communications. This research work can be refined and improved. Negative sentiment does not mean the content is malicious it could be a legitimate report on adverse events. Also, high engagement from an account with low followers could mean viral trend rather than harmful content. Similarly, a

tweet with neutral sentiment could have a potential to be malicious. To address these challenges, other techniques such as topic modelling to help discern the topic and context of the tweet, network analysis to determine coordinated behavior among accounts as well as using other pretrained machine learning models such as K-means, Support Vector Machines, or Ensemble Methods. Using these approaches can produce a more robust and accurate model that uses not only sentiment and engagement metrics but also context and network behavior to make predictions. This strategy could be a step further in effectively detecting harmful content and behavior in the increasingly complex digital space.

- *Use of Language Models*
  In the preparation of this research work, a Large Language Model (LLM) developed by OpenAI, specifically ChatGPT, was utilized to assist in refining the structure of the content, generating ideas for expanding certain sections, and improving overall clarity. The model was used as a tool for editorial support only, and all intellectual contributions, critical analyses, and final content decisions were made by the author. The LLM's involvement does not satisfy the authorship criteria as it does not possess the capacity for accountability or authorship responsibility. The authors acknowledge that the ultimate responsibility for the content remains with them.

Acknowledgements

I would like to express my gratitude to my college professor, Mr. Mark Attah Mensah who gave me the nudge I needed to find passion in python, data analysis and data science as well as providing me with a datacamp membership. Mr. Mark not only imparted knowledge but also gave me the motivation I needed to pursue a career in data analysis and data science. He also provided constructive criticism helping me refine my ideas and scope for this research. This research would not have been possible without his contribution to my academic growth. I am thankful for Mr. Mark's mentorship and the opportunity to learn under his guidance.

Link to Github Repository:

https://github.com/leoisqualified/Benign-and-Malicious-Post-Detection-on-Twitter

References

Mbona, I., & Eloff, J. H. P. (2023). Classifying social media bots as malicious or benign using semi-supervised machine learning. Journal of Cybersecurity, 9(1). https://doi.org/10.1093/cybsec/tyac015

Omar, S., Ngadi, A., & Jebur, H. H. (2013). Machine Learning Techniques for Anomaly Detection: An Overview. In International Journal of Computer Applications (Vol. 79, Issue 2).

Rajesh, A., & Kiran, S. (2018). Anomaly Detection Using Data Mining Techniques in Social Networking. In International Journal for Research in Applied Science & Engineering Technology (IJRASET) (Vol. 887). www.ijraset.com1268

Savage, D., Zhang, X., Yu, X., Chou, P., & Wang, Q. (2016). Anomaly detection in online social networks. http://arxiv.org/abs/1608.00301