Analyzing cyber security related tweets and subgroups using Machine Learning

Fatima Liaqat
Faculty of Science , University of Calgary
Calgary, Canada
fatima.liaqat1@ucalgary.ca

Amna Hassan
Faculty of Science , University of Calgary
Calgary, Canada
amna.hassan@ucalgary.ca

Dwight Pittman
Faculty of Science , University of Calgary
Calgary, Canada
dwight.pittman1@ucalgary

Abstract---This paper presents a study on the effectiveness of various machine learning models and feature selection techniques for detecting spam tweets. The study explores the impact of different features such as tweet text, location, action, and retweet status on the accuracy of spam detection models. The results indicate that using "tweet and location" as base features leads to the highest accuracy score, while the "is_retweet" feature has low accuracy and can be considered as noise. The study also highlights the importance of balancing precision and recall when building spam detection models, as predicting false positives can lead to unintended consequences such as censorship of legitimate content. The findings suggest that a combination of different techniques and features is necessary to achieve the best results in detecting spam tweets. The study recommends using Support Vector Machine (SVM) and Logistic Regression (LR) models for detecting spam tweets over Naive Bayes (NB).

Keywords— machine learning, social network, cyber security

I. INTRODUCTION

Critical social phenomena including elections, disease outbreaks, and cyberattacks have all been predicted, detected and/or studied using online social media platforms like Twitter. Due to the widespread use of social media platforms like Twitter to disseminate information regarding cyber security dangers and occurrences, both accurate and misleading, it is necessary and important to identify what tweets hold quality information, and which ones qualify as spam. Tweets can be classified as spam when they consist of automatically generated content, meaningless content and click bait.

Spam tweets often use phishing techniques or contain suspicious links or URLs which could lead to the malware being downloaded onto one's device. It is very important to detect and identify these types of tweets before they can be spread online. Another type of spam is fake news, also known as information pollution. These sorts of tweets are often combined with click bait to quickly pull a reader in and spread misinformation. The phenomenon of information pollution seems of little consequence, but in reality, it can easily be used to shift many people's opinions and decision making on a topic without consequence. Combined with malicious intent, these strategies can be used for political gain, can greatly affect society, and can even sway elections. Once again, detecting these tweets is essential to prevent fake news from hitting Twitter and the web before it can spread like wildfire.

Our research could help with detecting and preventing cybercrime, such as internet scams and data theft, for intelligence and law enforcement agencies. By identifying and tracking the connections that these offenders use, agencies can take action to disrupt and halt the operations of cybercriminals. Furthermore, a deeper comprehension of the current cyber threat scenario will help firms strengthen their overall security posture . Our study utilizes a hybrid methodology and uses a dataset that has not been used in prior research publications. Furthermore, by finding outliers, streamlining the analysis, improving visualization, and creating more accurate results, our grouping data approach can offer various advantages in data analysis and help to form better conclusions that will support the goal of our research.

This paper will discuss the previous work done in this area in section II. After that we will discuss the methodologies adopted in this paper in section III which will be followed by the Findings sections where we will discuss the results of each model in detail. Finally, we will summarize and conclude our paper's main findings in section IV. The last part of the report will be the references.

II. RELATED WORK

In [Mahaini, 2021] paper, they used a systematic approach to construct a dataset of cyber related accounts using a cyber security taxonomy, real-time sampling of tweets, and crowdsourcing. They used a multi-classifier approach to detect general and three subsets of cyber security related accounts with rich set of features and different ML models to compare models accuracy. Overall, they demonstrated the potential usefulness of detecting cyber security related accounts on social media for different purposes such as cyber threat intelligence and evaluating the effectiveness of cyber security awareness activities. Following are the strength and weaknesses of the paper.

- *Strengths*: Thoughtful sub-database classification with clear explanation.
- Detailed and easy-to-follow description of data collection process.
- ***Weaknesses:*** Requires individuals with good cyber security knowledge and time-consuming labeling task.
- Difficulty in obtaining a reliable dataset, resorting to random selection of some non-related accounts.

In [Dionísio, 2019], the authors focus on the detection of cyberthreats on Twitter using Deep Neural Networks. Their approach begins with keyword-based filtering on their dataset to distinguish between relevant and irrelevant information in the tweets. To do this, they deploy a Convolutional Neural network which is often used for Natural Language Processing. Tweets that are deemed to be potential threats are then run through a Named Entity Recognition model by using a Bidirectional Long Short-Term Memory neural network. This second neural network takes a tweet, and categorize the words from the tweet to allow for easy access to relevant information. As for the dataset used, the authors gathered tweets from two sets of accounts. The first set was used for training, validation and testing, while the second was only used in the testing portion. Tweets were deemed to be relevant when they posed security risks to one of the three private organizations that were specified in the paper which were a worldwide travel services provider, a global-company cybersecurity department, and a nation-wide power utility.

- *Strengths*: Keyword based filtering to detect relevant information from tweets.
- The authors also ran their dataset against previous systems that were designed to tackle the same problem to compare the systems.
- *Weaknesses:* Gathered their own tweets and labeled them themselves.

In [Jeong-Ha Park, 2022], the authors provide a unique methodology for identifying cyberattacks in this work. To determine the community that is most pertinent to cyberattacks, firstly the research conducts community detection on Twitter users in relation to cyberattacks. The second method overcomes the limitations of a lexical analysis of tweets, such as term-based filtering and keyword frequency, by doing textual similarity analysis between the tweet and the cyberattack-related keywords. Lastly, by combining text-based and graph-based models, the research suggests an unique cyberattack detection approach. Data analysis consists of graph analysis and text analysis. In graph analysis, classify the entire cyberattack-related users into groups according to community detection. In text analysis, analyze the similarity between tweets and the cyberattack-related keywords. Finally, construct a cyberattack detection model by connecting the analyzed results with attack information extracted from the Data Set. Following are the strengths and weaknesses of the paper:

- *Strengths*: incorporate the semantics in Tweets to evaluate the relevance with cyberattacks and employ community detection to identify the most relevant group to the cyberattacks.
- Weaknesses: New approach (Novel approach) - never been tested, may have its flaws.
- Community detection : The Louvian algorithm yields communities that may be arbitrarily badly connected

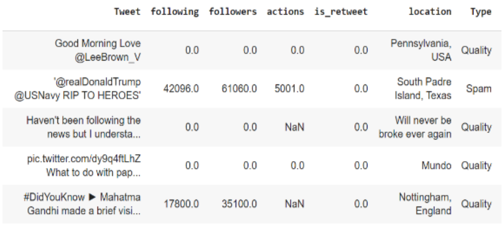
III. METHODOLOGY

This section discusses the methodology this paper.

A. Dataset

The research used a dataset from Kaggle “the UtkMI’s Twitter Spam Detection Competition”. For our purpose, we needed pre-labeled data, so we only used the original train.csv. The file contains 14,900 individual data points with seven columns: tweet, following, followers, actions, is_retweet, location and Type. After initial cleaning, we only found two inaccurate labels and then divided the train.csv into further two files based on an 80-20 ratio. We named the files train80.csv and test20.csv, and the division gave a new set of data with 11,918 data points to train and 2,981 data points for testing the model. During further cleaning, we found some null values in the dataset that were replaced with -1 or unknown depending on column data format. We had to standardize the location data to country names and replace incorrect and null values with unknown. Each feature is described below:

- Tweet: Holds the content of what the account has published to Twitter in this particular instance. This feature is a string with a maximum of 280 characters.
- Following: A numerical value which represents the number of accounts that the account who made the tweet follows. An account will generally follow accounts that post information that they are interested in, or accounts that they associate themselves with.
- Followers: A numerical value which represents the number of accounts that follow the account who made the tweet. Similarly, a follower generally indicates some sort of interest or association with the account who made the tweet.
- Actions: A numerical value which is an accumulation of the favorites, replies and retweets that the tweet has received. These terms represent the total number of engagements made with the tweet. A favorites represents another account’s enjoyment or agreeing with the contents, whilst a retweet is when another account shares the tweet with their followers. While most features only had a couple of null values in the data set, nearly quarter of the values for the actions feature were null.
- Is_retweet: This is a binary feature. A 0 indicates that this is the original tweet, made by the account being described in the data. A 1 indicates that the tweet was made by a separate account, and the account being described in the data shared the tweet.
- Location: The location feature represents where an account is located and where they are tweeting from. Unfortunately, this field is entered by the user of the account, which allows for inaccurate information. These could be incorrect locations, locations that are too broad, or something that isn’t even a place. Therefore, we decided to only use locations that included a country, and set other locations to “Unknown”.
- Type: This feature holds the label information of the data set, and is used as the ground truth. Data is labeled as either “Spam”, or “Quality”. Spam tweets are described as politically motivated, automatically generated content, meaningless content, and/or click bait. In these cases, a tweet has an alternative agenda, and is not being used to spread credible or relevant information. A tweet labeled as quality is any other kind of tweet.



Tweet	following	followers	actions	is_retweet	location	Type
Good Morning Love @LeeBrown_V	0.0	0.0	0.0	0.0	Pennsylvania, USA	Quality
'@realDonaldTrump @USNavy RIP TO HEROES'	42096.0	61060.0	5001.0	0.0	South Padre Island, Texas	Spam
Haven't been following the news but I understa...	0.0	0.0	NaN	0.0	Will never be broke ever again	Quality
pic.twitter.com/dy9q4ftLhZ What to do with pap...	0.0	0.0	0.0	0.0	Mundo	Quality
#DidYouKnow ► Mahatma Gandhi made a brief visi...	17800.0	35100.0	NaN	0.0	Nottingham, England	Quality

Fig. 1. Dataset Image

B. Approach

We used single features and pair features appraoch with combinations of various supervised machine learning algorithms like Support Vector Classification (SVC), Naives Bayes (NB) and Logistic Regression (LR) models that were paired with Count, Tfid vectorizers to see the performance of models and discover which technique gives us the highest accuracy in detecting spam tweets.

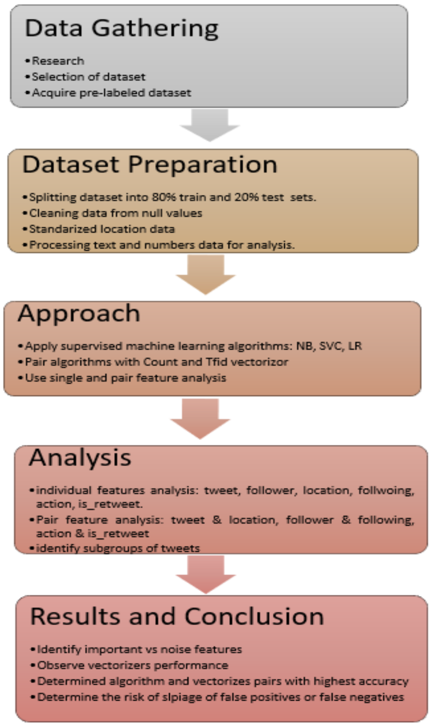

Data Gathering
•Research
•Selection of dataset
•Acquire pre-labeled dataset
Dataset Preparation
•Splitting dataset into 80% train and 20% test sets.
•Cleaning data from null values
•Standarized location data
•Processing text and numbers data for analysis.
Approach
•Apply supervised machine learning algorithms: NB, SVC, LR
•Pair algorithms with Count and Tfid vectorizor
•Use single and pair feature analysis
Analysis
•individual features analysis: tweet, follower, location, follwoing, action, is_retweet.
•Pair feature analysis: tweet & location, follower & following, action & is_retweet
•identify subgroups of tweets
Results and Conclusion
•Identify important vs noise features
•Observe vectorizers performance
•Determined algorithm and vectorizes pairs with highest accuracy
•Determine the risk of slpiage of false positives or false negatives

Fig. 2. Research Methodology for Framework Development

Individual feature Analysis

This analysis of the dataset based on specific attributes revealed information on many aspects of twitter user behavior. Here are a few ways to interpret the characteristics we discovered:

- Follower/Following count: The amount of followers/following a user has on Twitter can be a sign of their influence there or what topics they might be interested in. It can also help us determine whether a tweet is quality or spam. In section 4.3, we went into more detail about how follower count relates to other features like retweets and location to look for trends in how users and tweets with various levels of influence interact with content.
- Is retweet: The number of retweets indicates how many times a tweet has been shared on Twitter. Retweet analysis allowed us to determine which tweets have the greatest chance of going viral or reaching a larger audience, which had an influence on our analysis of tweets containing cybersecurity topics.
- Location: Analyzing the locations where the tweets were tweeted from can reveal regional variations in user activity or interests. Our data indicates that the United States is an extremely popular location. Section 4.3 presents additional examination of the location attribute in relation to other attributes.

1) Pair feature Analysis

We examined the connection between pairs of features in the dataset by pair feature analysis. Here are a few of the pairs we examined at:

- Content and location: We will be able to spot trends in the kinds of cybersecurity subjects that are most popular in various countries by looking at the distribution of content types across different places. This will enable us to better identify regional variations in the way that people talk about cybersecurity.
- Following and Followers: Users who follow accounts whose tweets have been classified as spam are much more, however the total number of followers for quality tweets is noticeably larger than that for spam.
- actions and is_retweet: Tweets that have not been retweeted have more actions than the ones which have not been retweeted. A tweet that has gotten a lot of likes and comments but hasn't been retweeted may indicate that the content of the tweet has attracted a lot of user interest but hasn't yet been widely shared. This can imply that a specific audience or community would find the tweet's content to be extremely relevant or appealing, and threat actors might use this information to target or influence that audience.

2) Subgroups

We categorized these tweets into subgroups for our final report. We will use a combination of manual and automated techniques to identify tweets that contain relevant keywords or phrases linked to cybersecurity in order to categorize tweets with cybersecurity subjects.Here is the approach we used to do this:

- We defined a set of keywords or phrases that are related to cybersecurity. We identified 2 subgroups: data breach and malware.
- Used our current dataset and selected some representative tweets from our training set.
- Checked the sample of tweets manually to make sure there are no false positives and that they are truly relevant to cybersecurity issues.
- These tweets were then further categorized into subgroups depending on the distinctive characteristics of their content, location, or action
- Finally, based on the properties of our dataset, we examined the subgroups to learn more about various facets of Twitter cybersecurity, such as the most

prevalent risks, the geographic distribution of cybersecurity talks, or the amount of participation around certain cybersecurity themes.

C. *Improvements:*

After attempting to add additional features to our best pair (Tweet and Location) to further increase the accuracy of our models, we saw similar or worse results. This indicated to us that with our methodology, we had reached the limit of how successful we could be with the available features in the data set. Despite this, we believe that the addition of more features with informative content about the account that made the tweet could be the key to finding more success. An example of some features that we believe would be useful are:

- Account's tweets per day: This statistic could indicate when an account is spewing information rather than cultivating informative tweets.
- Account's interactions per day: Another statistic which shows how often an account is interacting with other tweets. This could be relevant as we believe spam accounts would not be interested in doing this, as they only care about interactions on their own tweets.
- Account creation date: The date of the creation of the account could help in identifying spam accounts. We believe that long-time users would be less likely to tweet spam related content and an account that is relatively new might be a more likely candidate.
- Tweet's favorite to comment ratio: This ratio can be informative to the overall reception that the people who are interacting with the tweet have towards it. For example, a tweet with a low number of favorites, and a high number of comments will generally indicate that users disagree with the contents of the tweet, as they have taken time to respond to the tweet but have often not favorited it.

Furthermore, the combination of these features, as well as using the existing features in the data set, could allow for better predicting when a tweet is spam or quality.

We also believe that the introduction of further techniques could improve the classification of tweets. For example, the addition of keyword-based filtering could be beneficial to identify the subject of a tweet. Another area to explore would be to further classify spam tweets into their type of spam, such as phishing, malware, suspicious links, etc. Keyword-based filtering would also be a good place to start in this area.

IV. FINDINGS

A. *Models and their Performances*

As discussed in section III, we ran our model on individual and combinations of features using various combinations of techniques and compared these results to find the best accuracy score. The results showed that the combination of "tweet & location" as a feature with the tfid vectorizer and SVC model predicted the labels with highest accuracy score of 94.8993% (see TABLE I) and (see Figure 3). The accuracy scores of all techniques of this feature combination are higher than any individual or other features combination's score. This indicates that a combination of "tweet & location" is overall the best feature to detect spam vs quality tweets. The combination of "following & follower" score also went up in comparison to their individual scores. We also noticed that "is_retweet & action" scores collectively and individually are the least effective features to detect spams.. Especially, the "is_retweet" feature is not effective as its accuracy score remains below 50% regardless of which technique we used. We also noted that regardless of the model or vectorizer used, when looking at "is_retweet" on its own, the model would classify every tweet as quality. Even when "is_retweet" was combined with "action", it show no significant improvement.

Further we see that in single features, "tweet" as a feature with the combination of the a tfid vectorizer and SVC model has the highest accuracy score in detecting spam vs quality tweets (see TABLE 1) and (see Figure 4). Also, the difference between the scores of "tweet" and "tweet & location" using Tfid and SVC model, shows less than a 4% decrease. This indicates that "tweet" can be used as a sole feature in detecting spams when the location information is absent. Also, the "tweet" feature with tfid vectorizer and LR model accuracy is also very high and could be used as an alternative approach to detect spam tweets.

We observe that in most cases the tfid vectorizer with the SVC model is overall the best combination in detecting spams vs quality tweets. The exception of this is the case of "location", where we found that the Count vectorizer with the NB model proved to be the better approach. We also observed that in most cases the model's accuracy was affected by predicting false positives, meaning that models were often predicting spam tweets as "Quality". This was especially true when it came to the "followers" feature.

Moreover, we found that adding additional features to the "tweet & location" pairing did not improve the models. Therefore, we decided that further research into more combinations of features would provide similar results.

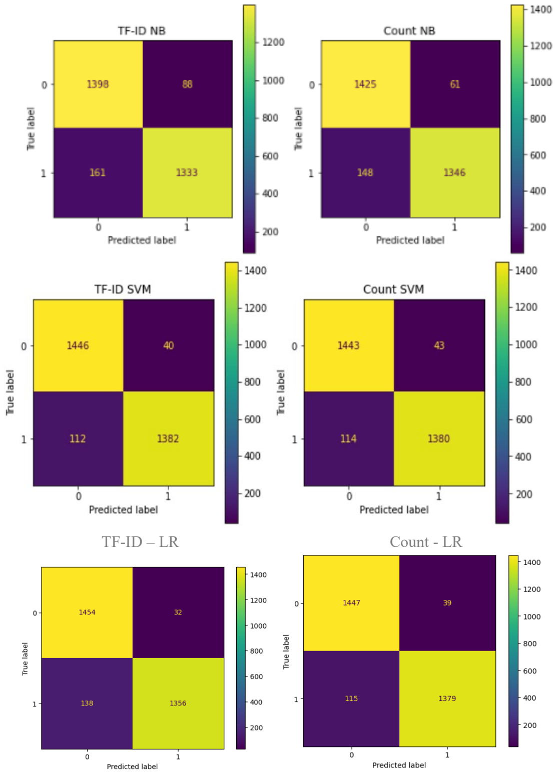


Fig. 3. Accuracy matrices of "tweet & location" feature

Fig. 4. Accuracy matrices of “tweet” feature

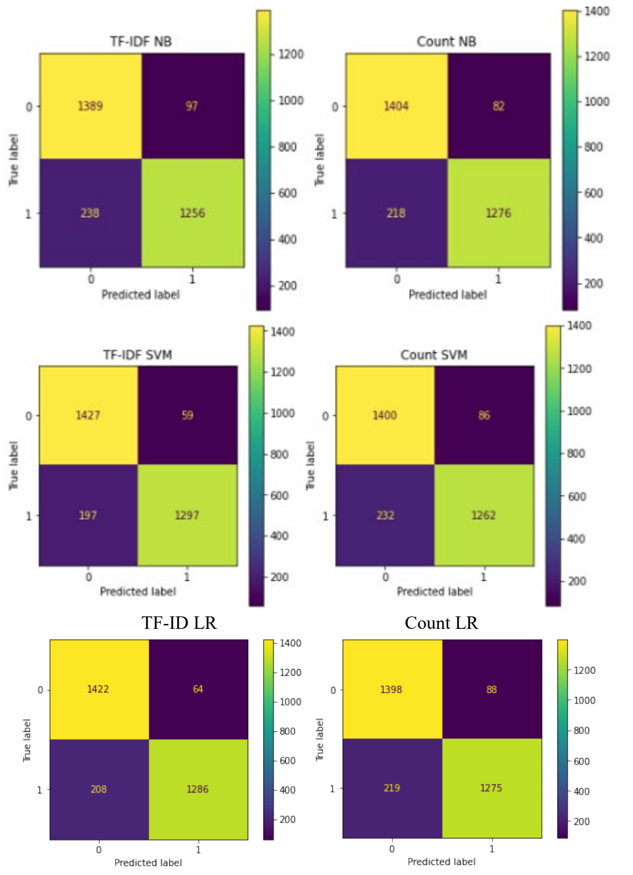


TABLE I. COMPARISON OF ACURRAY SCORE OF ALL MODELS, VECTORIZERS AND FEATURES

Feature	Accuracy Scores in Percentages					
	Tfid & NB	***Count & NB***	***Tfid & SVC***	***Count & SVC***	***Tfid & LR***	***Count & LR***
tweet	88.7584	89.9329	91.4094	89.3289	90.8725	89.6980
follower	76.0403	61.7785	75.8054	61.7114	70.5369	61.6779
location	72.9866	75.5369	72.9530	75.2685	73.2215	75.4026
following	77.3826	64.8658	77.3826	64.8658	75.9060	64.8658
is_retweet	49.8658	49.8658	49.8658	49.8658	49.8658	49.8658
action	61.9799	53.1208	62.2148	53.1208	61.9128	53.1208
tweet & location	91.6443	92.9866	**94.8993**	94.7315	94.2953	94.8322
following & Follower	84.1275	70.9396	89.2617	70.4362	84.0268	70.9060
Is_retweet & action	61.8121	53.1208	62.2148	53.1208	61.9463	53.1208

B. Sub-grouping

It will be possible to identify patterns and insights that might not be immediately obvious when looking at the data by conducting subgroup analyses on the data. It might be possible to spot regional variations in cybersecurity concerns or trends, for instance, by categorizing tweets by location. The types of actions (like, comment, etc.) users take to a particular tweet may be able to assist us to learn which subgroup subjects are more popular. In summary, using subgroups will be an effective technique to examine cybersecurity data and discover information that can guide us to come up with a more detailed analysis.Our approach is mentioned in section III above. Due to time constraints this section was not done in depth and we would have categorized the tweets into more subgroups and would have done further analysis of the dataset for subgrouping.

- From our Train80.csv data set we looked at 180 tweets for our subgroup analysis.
- We identified 2 subgroups: data breach and malware.
- Keyword set we used for data breach was: [Data leak,Stolen data,Exposed data,Breached data,Compromised data,Unauthorized access to data,Data security breach,Data privacy violation,Personal information stolen,Sensitive information exposed,Credit card information stolen,Passwords compromised,Hackers stole data,Cybersecurity breach].
- Keyword set we used for malware was
-

  [Virus alert,Malware infection,Suspicious activity on my computer,Strange pop-ups on my computer,Computer running slow,Unauthorized access to my computer,Malicious software detected,Malware attack,Trojan horse,Adware,Ransomware,Spyware,Keylogger,Botnet,Phishing malware,Exploit kit,Malicious link,Malware removal,Malware scanner,Malware analysis].

1) Final results

- According to our analysis there is a larger number of people following the malware subgroup category than data breach.
- Data breach subgroup has more actions (total number of favorites , replies, and retweets of said tweet).

On a larger scale, our research can help in identifying trends and patterns in conversations around cybersecurity on Twitter by performing subgroup analysis. This can assist us or researchers using our methodology in better understanding the subjects that users find most interesting, the risks that are most common, and the reactions that users have to them.

C. Visualizations of Data Analysis

After cleaning our data, we examined it and identified more connections that we may use to further examine the data. For the time being, we have conducted a general analysis of the dataset using a few attributes and features. The visualizations are as follows:

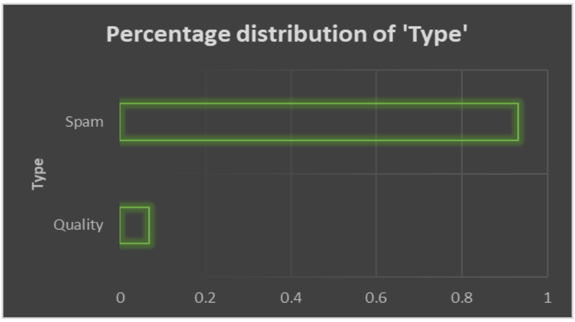


Fig. 5. The percentage of distribution of Quality vs Spam

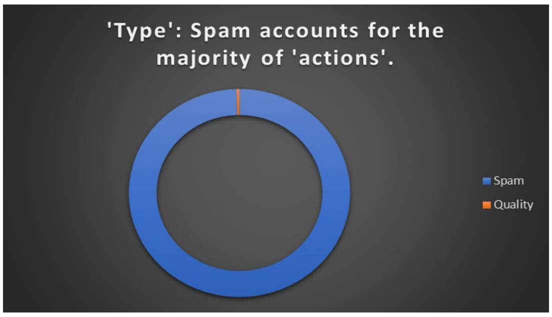


Fig. 6. Sum of followers for the two types: Quality has more followers

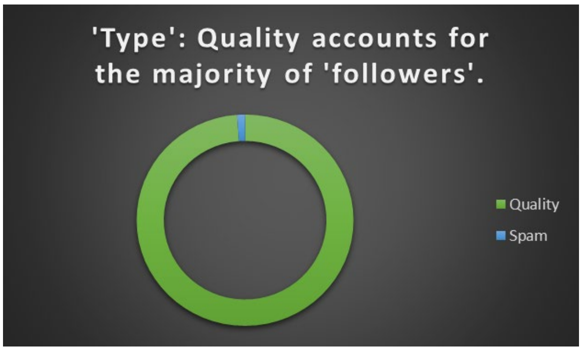


Fig. 7. Type and actions: The type spam accounts for majority of the actions.

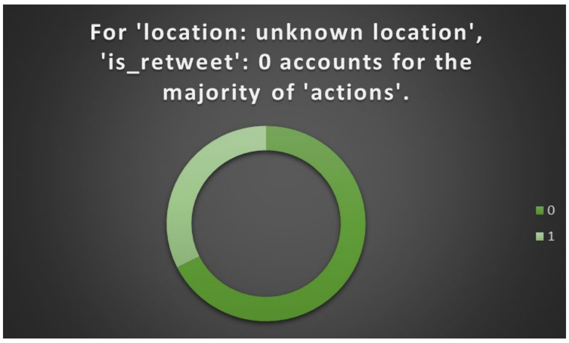


Fig. 8. Analysis of number of following, classified by is_retweet and its type.

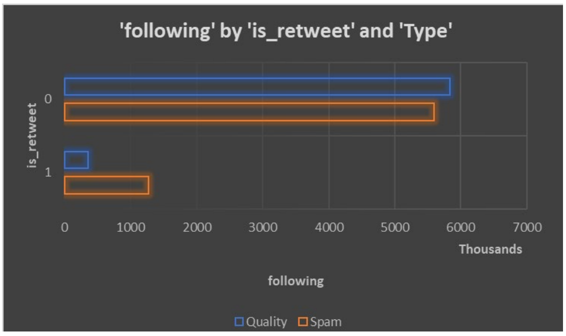


Fig. 9. Tweets with unknown location that have not been retweeted account for majority of actions.

V. SUMMARY AND CONCLUSION

As discussed in section IV, using "tweet & location" as our base feature to detect spam has given us the best accuracy score overall. Even using "tweet" as a single feature could be very beneficial when location details are not available. This is a good indicator that we can filter most spam tweets. We also found that SVC model combined with tfid vectorizer proved to be the most effective technique in detecting spam across any feature. We also noticed that is_retweet feature had very low accuracy score even when combined with action. This result confirms that we can considered is_retweet more or less as a "noise" in data.

Our research showed us that the accuracy of the models we have used has been affected by predicting many false positives, the problem of predicting spam tweets as quality (false positives) can lead to several negative consequences. If these models are used individually to automatically flag tweets for moderation, then poor assessments can lead to legitimate content being censored, which can infringe on free speech or lead to complaints from users. It can also lead to many spam tweets seeping through and being passed as quality. Therefore, it is important to carefully balance the trade-offs between precision (correctly identifying high-quality tweets) and recall (minimizing false negatives) when building these types of models. From this we can say that using one technique is not the best choice when checking for accuracy. Using a combination of different techniques and features would be the best way to detect the spam tweets.

Overall we can say that the "tweet" is the most important feature followed by "location" to determine whether a tweet is spam or not, and "action" and "is_retweet" features can be classified as "noise. Furthermore, Tfid vectorizer showed better accuracy score as compare to count vectorizer when combine with different models. Lastly, SVC is the best model followed by LR and are more preferred models over NB when detecting Spam vs Quality tweets.

REFERENCES

[1] Mahaini, M.I. and Li, S. (n.d.). Institute of Cyber Security for Society (iCSS) & School of Computing, University of Kent, Canterbury, UK. In Proceeding of IEEE/ACM International Conference on Advance in Socila Network Analysis and Mining, 2021. DOI: 10.1145/3487351.3492716

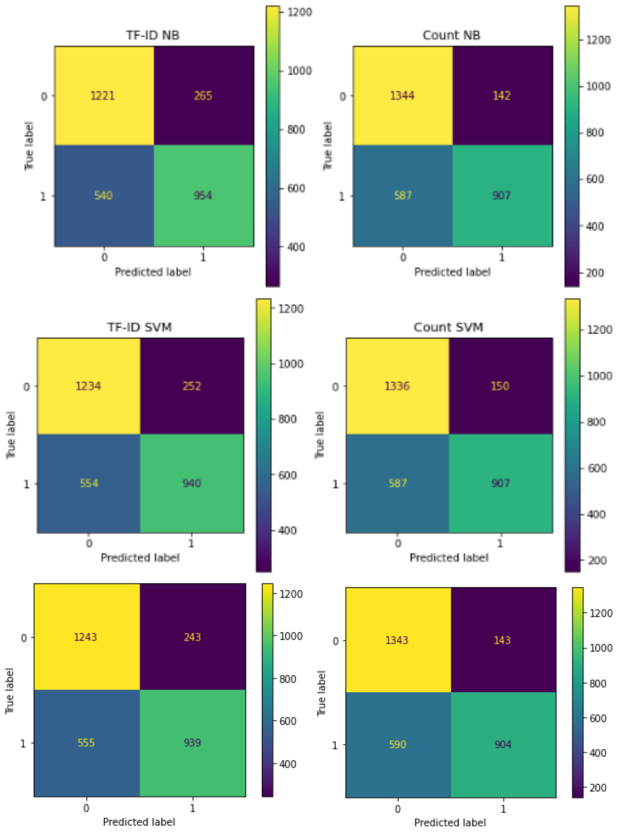
[2] Kaggle 2019. Retrieved from kaggle, 2023 from https://www.kaggle.com/competitions/utkmls-twitter-spam-detection-competition/data

[3] Dionísio, N., Alves, F., Ferreira, P.M. and Bessani, A., 2019, July. Cyberthreat detection from twitter using deep neural networks. *In 2019 international joint conference on neural networks (IJCNN)* (pp. 1-8). IEEE.

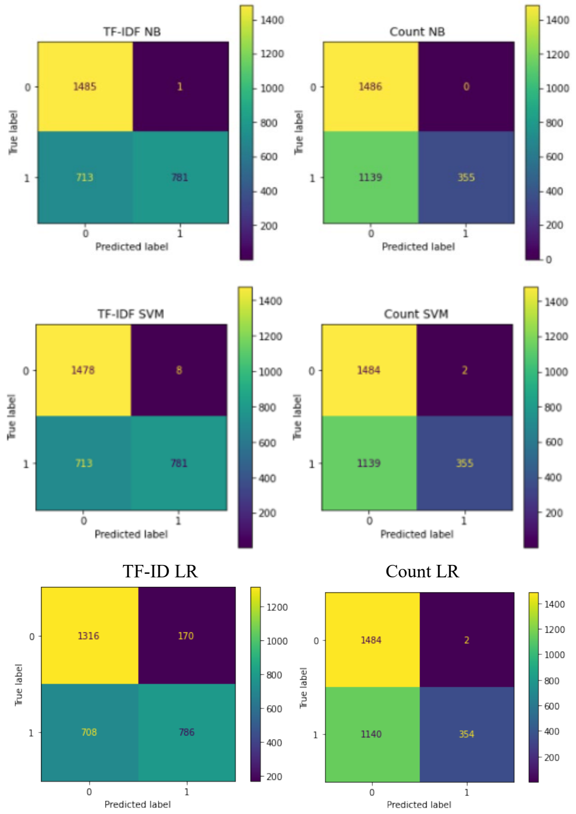
[4] Jeong-Ha Park and Hyuk-Yoon Kwon. 2022. Cyberattack detection model using community detection and text analysis on social media. Retrieved January 24, 2023 from https://www.researcher-app.com/paper/10025686, DOI: 10.1016/j.icte.2021.12.003

APPENDIX

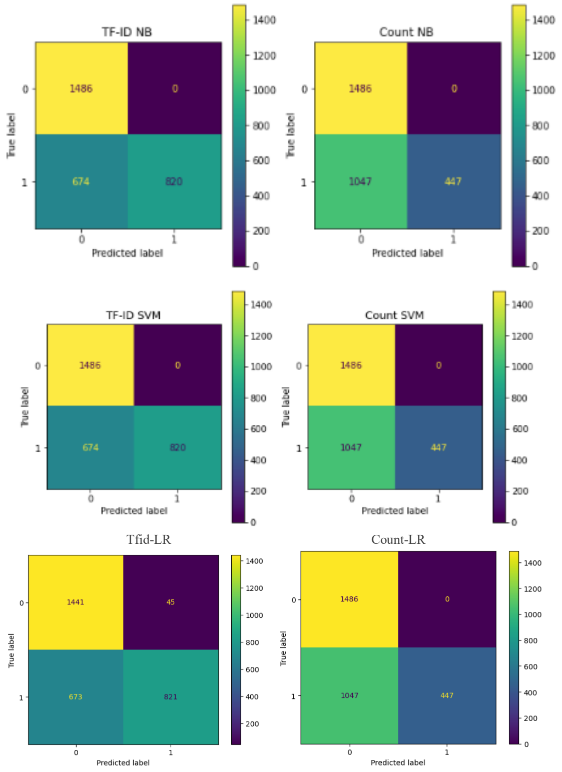
This section provides all the confusion matrix of remaining features.



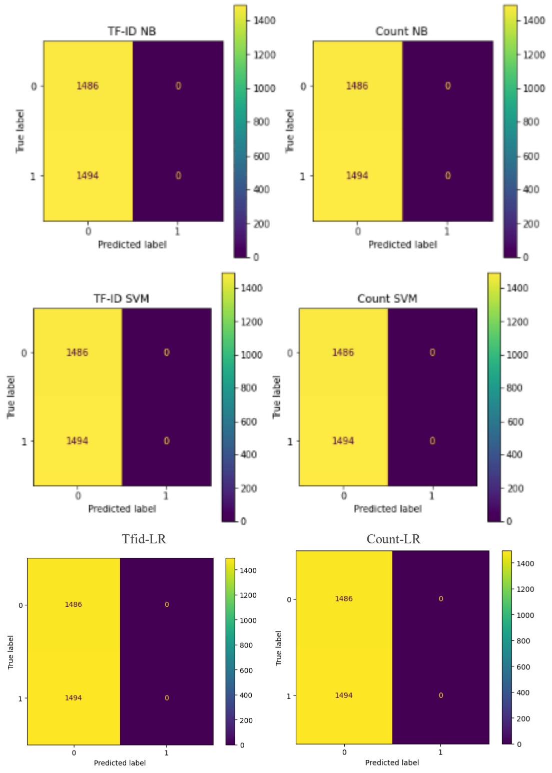
Accuracy matrices of “Location” feature



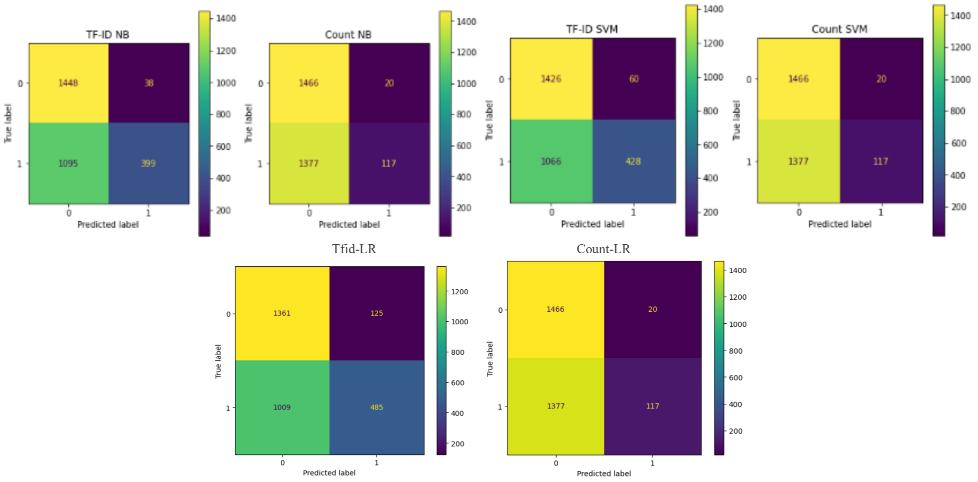
Accuracy matrices of “Followers” feature



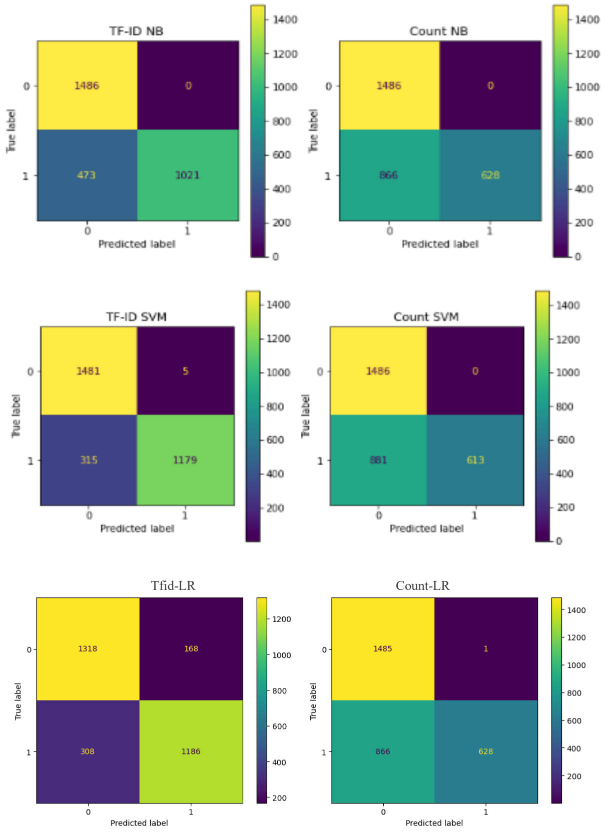
Accuracy matrices of “Action” feature



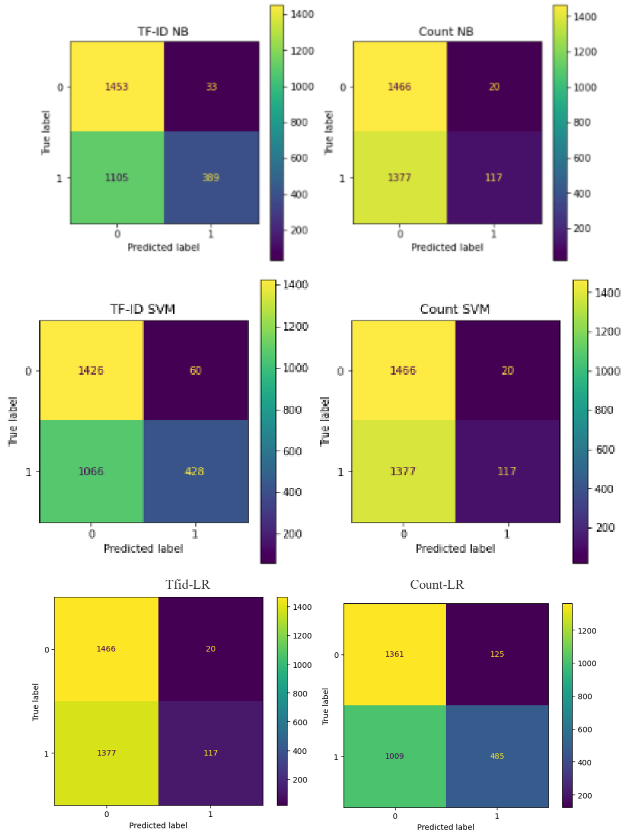
Accuracy matrices of “is_Retweet” feature.



Accuracy matrices of “Action” feature



Accuracy matrices of “Following & Followers” feature



Accuracy matrices of “is_Retweet & Action” feature.