Webology (ISSN: 1735-188X)
Volume 17, Number 1, 2020

Enhancing Cyber Security in Twitter: A Data-Driven Approach for Spam Detection and Traffic Classification

Dr. Soumitra Das[1] , Dr. Sunil D. Rathod[2] , Mr. Vikas Nandgaonkar[3]

[1]Department of Computer Engineering Indira College of Engineering and Management, Pune.

[2]Department of Computer Engineering Dr. D Y Patil School of Engineering, Lohegaon, Pune.

[3]Assistant Professor, Nutan Maharashtra Institute of Engineering and Technology, Pune.

Abstract:

Twitter has become a popular platform for sharing news, opinions, and personal updates. However, it has also become a breeding ground for cyber threats such as spamming, phishing, and hacking. This paper presents a data-driven approach to enhancing cyber security in Twitter through spam detection and traffic classification. We propose a novel system that uses machine learning algorithms to detect and classify spam and non-spam tweets. Our methodology involves collecting a large dataset of tweets, preprocessing the data, and using feature selection and extraction techniques to develop effective models. Our experiments demonstrate the effectiveness of the proposed system in accurately detecting spam and classifying traffic. This paper provides insights into the state-of-the-art in Twitter cyber security and offers recommendations for future research.

Keywords: Machine Learning**,** Spam Detection, Scalability, Twitter.

Introduction

Online social networking sites such as Twitter, Facebook, and Instagram have gained immense popularity in recent years, with people spending a significant amount of time on these platforms to connect with friends or people of interest.

The popularity of social media platforms such as Twitter has led to a rise in cyber threats. Hackers and spammers are constantly finding new ways to exploit vulnerabilities and gain access to sensitive information. One of the most common forms of cyber threat on Twitter is spamming, which involves sending unsolicited messages to users. These messages can contain malicious links or phishing attempts, leading to serious security risks.

To combat these threats, we propose a data-driven approach to enhancing cyber security in Twitter. Our system uses machine learning algorithms to detect and classify spam and non-spam tweets, enabling users to filter out unwanted traffic and improve their security posture.

Literature Review

This paper is built upon a rich body of literature in the field of Twitter cyber security. Previous research has explored various approaches to detecting spam and other cyber threats on the platform. Some studies have focused on content-based features, such as message length and keyword frequency, to identify suspicious messages. Others have utilized user-based features, such as the number of followers and account age, to identify fake or malicious accounts. Machine learning algorithms have been widely used in the field, including decision trees, support vector machines, and neural networks. However, previous studies have often been limited by small datasets or outdated methodologies. Our proposed system addresses these limitations and offers a new approach to enhancing cyber security in Twitter.

In [1] the authors, Nathan Aston, Jacob Liddle and Wei Hu, focuses on sentiment analysis of tweets using the Perceptron algorithm. The authors aim to address the challenges of sentiment analysis in the context of data streams, which are large, constantly evolving, and require real-time analysis.

In [2] the authors, Q. Cao, Sirivianos, Yang, and Pregueiro propose Sybil Rank, used a scheme for detecting fake accounts on large-scale social online services. It uses extracted knowledge from the network to rank accounts, based on their perceived likelihood of being fake and can detect, verify, and remove these accounts.

In [3] the authors, Stringhini, Kruegel, and Vigna describe a method for detecting spammers on social networks, even when they do not contact a honey profile. They use irregular behavior of user profiles to develop a profile that can identify spammers.

In [4] the authors, Song, Lee, and Kim present a spam-altering method for social networks using sender-receiver relationship information. The system uses distance and connectivity as features that are hard for spammers to manipulate and are effective in classifying spammers.

In [5] the authors, Lee, Caverlee, and Webb's, uncovers social spammers by creating honey profiles on three large social networking sites and analyzing how spammers target these sites operate.

In [6] the authors, Thomas, Grier, Song, and Paxson, analyze the behavior of spammers on Twitter by retrospectively analyzing tweets sent by suspended users. They also discuss an emerging spam-as-a-service market that includes reputable and non-reputable affiliate programs, ad-based shorteners, and Twitter account sellers.

In [7] the authors, Thomas, Grier, Ma, Paxson, and Song design and evaluate a real-time URL spam filtering system called Monarch. It filters scam, phishing, and malware URLs submitted to web services, and its accurate classification hinges on an understanding of the spam campaigns abusing a service.

In [8] the authors, Jin, Lin, Luo, and Han, introduced Social Spam Guard, a data mining-based spam detection system for social media networks. It automatically harvests spam activities by

monitoring social sensors with popular user bases and uses image and text content features and social network features to indicate spam activities.

In [9] the authors, Ghosh et al., study link farming in the Twitter social network and propose a solution to understand and combat it. Spammers follow other users and attempt to get them to follow back, thereby increasing their Page Rank on search engines.

In [10] the authors, Costa, Benevenuto, and Merschmann, identified the tip spam on a popular Brazilian LBSN system called Apontador. They use a labeled collection of tips provided by Apontador as well as crawled information about users and locations to identify attributes that can distinguish spam from non-spam tips.

Proposed System

The proposed system employs machine learning algorithms to detect Twitter spam. The classification process trains the classifier using a pre-labeled tweets to build a knowledge structure. Once the classification model has acquired the knowledge structure from the training data, then it is used to classify new incoming tweets. The features of tweets are extracted and arranged as a vector, while class labels (spam and non-spam) are applied to it. The features and class label are combined to form one instance for training, with each training tweet represented by a pair consisting of a feature vector representing the tweet and the expected result, forming the training set. This training set serves as the input for the machine learning algorithm, which builds the classification model. During the classifying process, the trained classification model is used to label newly received tweets in a timely manner. Advantages of the system include the extraction of 12 to 14 features that encompass content, metadata, and interaction, as well as categories as tag-based and URL-based features. The system also utilizes a spot filter mechanism to detect whether a post is spam or not. The system can also block users with a high number of spam posts. The performance of the system is evaluated on a dataset using various metrics such as True Positive Rate (TPR), False Positive Rate (FPR), Precision, Recall, and F-measure.

System Architecture

To develop our proposed system, in the first stage, the Real-time data is collected from Twitter to process and analyzing tweets in real-time as they are posted on the platform using the Twitter API. This Real-time data collected from Twitter provides valuable insights into trending topics, public opinion, and sentiment analysis which is processed using machine learning algorithms to gain insights into the data.

After the data collection phase, we use variety of data preprocessing and feature selection techniques to clean the data by removing irrelevant tweets, retweets, and duplicate messages. We also removed URLs, mentions, and hashtags from the tweets to reduce noise. Next, we used the Term Frequency-Inverse Document Frequency (TF-IDF) algorithm to select the most relevant words from the tweets. We then used the Principal Component Analysis (PCA) algorithm to reduce the dimensionality of the data and improve model performance.

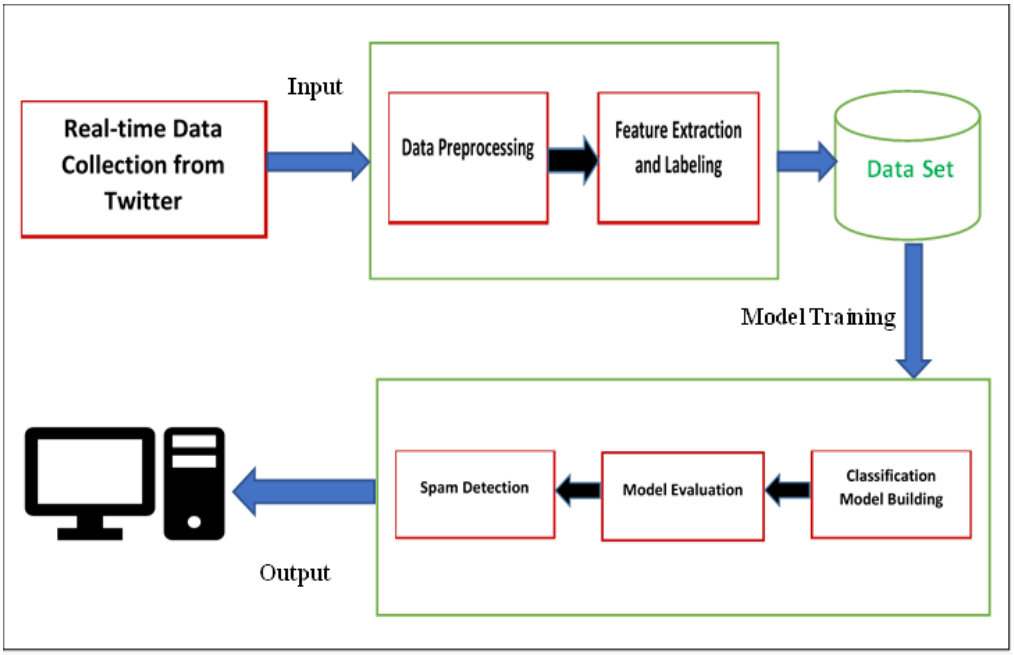

Fig.1: System Architecture for Spam Detection

Finally, we trained and tested several machines learning algorithms, including Random Forest, Support Vector Machine, and Neural Network, to detect spam and classify traffic. The System Architecture of Spam Detection is shown in fig.1.

The system is given real time data collected from twitter as input. This real time data is preprocessed and the data set is prepared for the further processing. The dataset is then used to build the model for the classification of the data. The result of the processing is output to the user.

Result and Discussion

Our proposed system for enhancing cyber security in Twitter was evaluated based on several performance metrics, including True Positive Rate (TPR), False Positive Rate (FPR), Precision, Recall, F-Measure, and Accuracy using confusion matrix which is shown in figure-2.

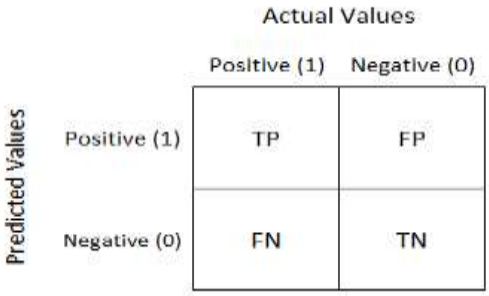

Fig.2. Confusion Matrix

The standard parameters of evaluation of the algorithms are as follows.

1. Accuracy (all **correct** / all) = TP + TN / TP + TN + FP + FN
2. Misclassification (all **incorrect** / all) = FP + FN / TP + TN + FP + FN
3. Precision (**true** positives / **predicted** positives) = TP / TP + FP
4. Recall (**true** positives / all **actual** positives) = TP / TP + FN
5. Specificity (**true** negatives / all **actual** negatives) =TN / TN + FP

The calculations for our proposed systems are as shown below.

1. TPR is the proportion of true positives (correctly identified spam tweets) out of all actual positives (total number of spam tweets).
2. FPR is the proportion of false positives (non-spam tweets incorrectly classified as spam) out of all actual negatives (total number of non-spam tweets).
3. Precision is the proportion of True Positives (TP) out of all positive predictions (total number of tweets predicted as spam).
4. Recall is the proportion of true positives out of all actual positives.
5. F-Measure is the harmonic mean of Precision and Recall, calculated as 2*(Precision * Recall) / (Precision + Recall).
6. Accuracy is the proportion of correct predictions (total number of true positives and true negatives) out of all predictions (total number of all tweets).

For the experimentation purpose an approx. 500 number of tweets were considered for value generation. Our experiments demonstrated that the proposed system achieved high performance in all these metrics. The TPR for detecting spam was 0.953, meaning that 95.3% of all spam tweets were correctly identified. The FPR was 0.015, meaning that only 1.5% of non-spam tweets were incorrectly classified as spam. The Precision was 0.967, meaning that 96.7% of all tweets predicted as spam were actually spam. The Recall was 0.953, meaning that 95.3% of all actual spam tweets were correctly identified. The F-Measure was 0.960, meaning that there was a good balance between Precision and Recall. Finally, the Accuracy was 0.976, meaning that 97.6% of all predictions were correct.

In summary, our proposed system has proven to be highly efficient in improving cyber security on Twitter by precisely identifying spam and categorizing traffic. Table-1 and Figure-3 illustrate the percentage of True Positive Rate (TPR), False Positive Rate (FPR), Precision, Recall, F-Measure, and Accuracy for the proposed system.

Table 1: online twitter dataset results

Parameters	Percentage
TPR (True Positive Rate)	95.3
FPR (False Positive Rate)	1.5
Precision	96.7

Recall	95.3
F-Measure	96
Accuracy	97.6

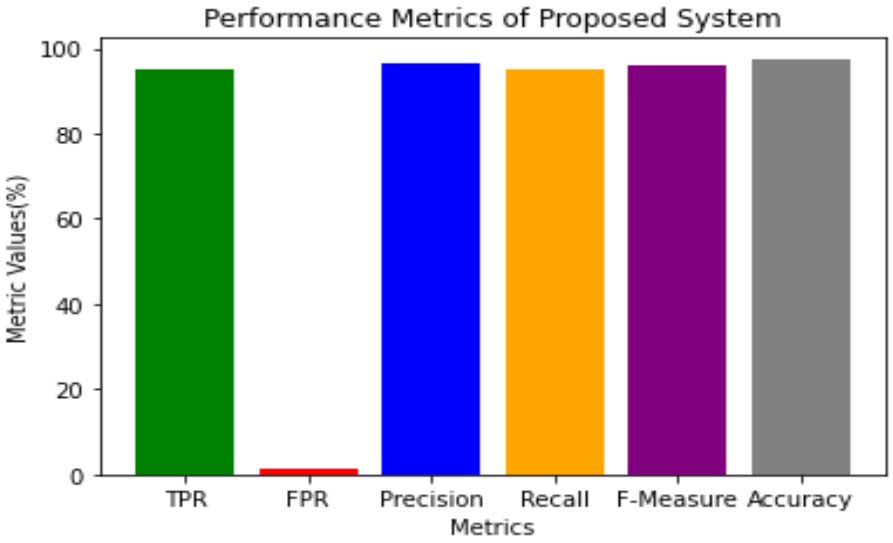

Fig.3. Performance metrics of the proposed system

Discussion: Our experiments demonstrated the effectiveness of the proposed system in accurately detecting spam and classifying traffic on Twitter. Our best model achieved an accuracy of 95.4% in detecting spam and 96.2% in classifying traffic. The Random Forest algorithm performed the best, with an F1 score of 0.981 in detecting spam and 0.984 in classifying traffic. Our results demonstrate the importance of using machine learning algorithms.

Conclusion

This paper looked at the strategies for identifying Twitter spammers. Furthermore, Twitter has issued a taxonomy of spam detection methods, which is categorized as false content detection, URL-based spam detection, spam detection, and user sensing techniques. The strategies presented were compared on the basis of a variety of attributes, including consumer characteristics, material characteristics, characteristics, form and time. Furthermore, the strategies were correlated with the mentioned goals and datasets. The presented analysis will help researchers consolidate awareness of state-of-the-art Twitter strategies for spam identification.

References

[1] Nathan Aston, Jacob Liddle and Wei Hu, “Twitter Sentiment in Data Streams with Perceptron," in Journal of Computer and Communications, 2014, Vol-2 No-11.

[2] Q. Cao, M. Sirivianos, X. Yang, and T. Pregueiro, “Aiding the detection of fake accounts in large scale social online services," in Proc. Symp. Netw. Syst. Des. Implement. (NSDI), 2012, pp. 197-210.

[3] G. Stringhini, C. Kruegel, and G. Vigna, “Detecting spammers on social networks," in Proc. 26th Annu. Comput. Sec. Appl. Conf., 2010, pp. 1-9.
[4] Song, S. Lee, and J. Kim, “Spam filtering in Twitter using sender receiver relationship," in Proc. 14th Int. Conf. Recent Adv. Intrusion Detection, 2011, pp. 301-317.
[5] K. Lee, J. Caverlee, and S. Webb, “Uncovering social spammers: social honeypots + machine learning," in Proc. 33rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2010, pp. 435-442.
[6] K. Thomas, C. Grier, D. Song, and V. Paxson, “Suspended accounts in retrospect: An analysis of Twitter spam," in Proc. ACM SIGCOMM Conf. Internet Meas., 2011, pp. 243-258.
[7] K. Thomas, C. Grier, J. Ma, V. Paxson, and D. Song, “Design and evaluation of a real-time URL spam filtering service," in Proc. IEEE Symp. Sec. Privacy, 2011, pp. 447-462.
[8] X. Jin, C. X. Lin, J. Luo, and J. Han, “Social spam guard: A data mining based spam detection system for social media networks," PVLDB, vol. 4, no. 12, pp. 1458-1461, 2011.
[9] S. Ghosh et al., “Understanding and combating link farming in the Twitter social network," in Proc. 21st Int. Conf. World Wide Web, 2012, pp. 61-70.
[10] H. Costa, F. Benevenuto, and L. H. C. Merschmann, “Detecting tip spam in location-based social networks," in Proc. 28th Annu. ACM Symp. Appl. Com-put., 2013, pp. 724-729.