iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** II **Month of publication:** February 2022

DOI: https://doi.org/10.22214/ijraset.2022.40397

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue II Feb 2022- Available at www.ijraset.com

A Semi-Supervised Learning Approach for Tackling Twitter Spam Drift

Priyanka .R[1], Dr. J Bhuvana[2]
[1]*Student, Department of Master of Computer Applications School of Computer Science IT, Jain Deemed to Be University, Jayanagar 9th Block, Bengaluru, Karnataka– 560041, India.*
[2]*MCA Coordinator, Department of Master of Computer Applications School of Computer Science & IT, Jain Deemed to Be University, Jayanagar 9 th Block, Bengaluru, Karnataka– 560041, India.*

Abstract: ***Twitter play an important role in accelerating the spread of spam. In order to protect the users, Twitter and the research community have been developing different spam detection systems by applying different machine-learning techniques. However, a recent study showed that the current machine learning-based detection systems are not able to detect spam accurately because spam tweet characteristics vary over time. This issue is called "Twitter Spam Drift". In the proposed system a semi-supervised learning approach (SSLA) has been proposed to tackle this. The new approach uses the unlabeled data to learn the structure of the domain. To handle the drift, live twitter stream of data is taken for the study. The pre-processing of live-downloaded data is labeled and machine learning is applied to detect spam and non-spam users. The data is stored in cloud storage, which can be accessed by user from anywhere. Experimental results were conducted on more than one machine learning algorithm and finds the better for the proposed problem, in-terms of accuracy.***

I. INTRODUCTION

The online social networks (OSNs) such as Facebook, WhatsApp, and Twitter have become a very important part of daily lives nowadays. People use them to make friends, communicate with each other, read the news, and share their stories. Twitter, which was founded in 2006, has become one of the most popular microblogging platforms since then. A Twitter generic profile consists of three components: the account's tweets, followers, and friends. In addition to the account components, there are several Twitter-specific features, such as mentions, hashtags, and retweets. Users can only post messages (tweets) up to 140 characters. These tweets can contain text, hashtags, mentions, and shortened URLs. Unfortunately, due to the high popularity of Twitter, it has become very attractive to spammers. In Twitter, spammers tweet for several goals, such as to spread advertisement, disseminate pornography, spread viruses, phishing, or simply just to compromise a system's reputation. Spammers use trending hashtags to direct users to unrelated topics. Also, spammers use mentions to spread spam tweets. The most important part of spam tweets is the shortened URLs, which enable spammers to deceive users. Different studies showed that about 5% to 6% of messages in Twitter are spams.9,10 Consequently, the research community and Twitter have proposed several spam detection systems to protect users. Twitter has applied rules against spammers or those who behave abnormally. For instance, users who are frequently sending friend requests, sending duplicate content, mentioning others, or posting tweets containing only a URL are considered spammers. Also, Twitter provides different options to its users to report spammers, such as selecting report @username, clicking on the report icon, or clicking on report conversation. However, spammers are using different ways to evade detection by buying followers or mixing spam tweets with normal tweets. This motivates the research communities to develop new, innovative mechanisms.

Twitter is becoming increasingly popular in the last few years. It has been rated the most popular social network among teenagers according to a recent study. However, the popularity of Twitter also makes it an attractive platform for spamming activities. Twitter spam, which is referred as unsolicited tweets containing malicious links that directs victims to external sites containing malware downloads, phishing, drug sales, or scams, etc., has already polluted the platform.

Consequently, security companies, as well as Twitter itself, have devoted themselves to make Twitter a spam-free platform. For instance, Trend Micro uses a blacklisting service called Web Reputation Technology system to filter spam URLs for users who have their products installed. Twitter also applies blacklist as a component in their detection system called BotMaker. However, blacklists' lagging time fails to protect victims from new spam. Research shows that, more than 90% victims may visit a new spam link before it is blocked by blacklists. In order to tackle the limitation of blacklists, researchers have proposed some detection schemes which can make use of user activity patterns to detect spam without checking the URLs.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue II Feb 2022- Available at www.ijraset.com

II. PROBLEM STATEMENT

The current machine learning approaches used for detecting Twitter spam cannot overcome the problem of drifted Twitter spam because the statistical features of spam tweets are changing over time. As a result, the accuracy of the traditional machine learning algorithm is decreasing gradually as time goes on. This is the most considerable problem in existing twitter spam detection. In labeled large dataset, we have observed this problem and carried out a thorough analysis of it.

III. PROPOSED SYSTEM

A semi-supervised learning approach (SSLA) has been proposed to tackle the Twitter spam drift where we used unlabeled data to learn the structure of the domain. Different experiments were performed on live tweet dataset and evaluate the proposed approach and the results show that the proposed SSLA can reduce the effect of Twitter spam drift and outperform the existing techniques.

SSLA is a type of ML technique that is useful when the number of labeled instances is limited. The aim of SSLA is to combine labeled and unlabeled data to create better learners.

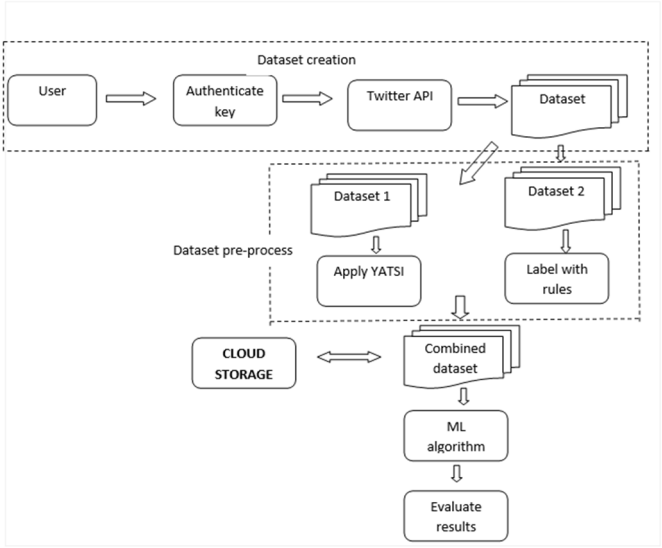
The SSLA has been used to evaluate different applications, such as software fault detection, text classification, spam email detection, quantitative structure-activity modeling, and so forth.

The SSLA was chosen to solve the Twitter spam drift problem for several reasons. First, the SSLA uses a combination of labeled and unlabeled data at the same time.

Second, SSLA is more applicable than supervised learning approaches when the amount of unlabeled data is huge. This attribute is very important when dealing with changed spam tweets.

Third, the SSLA lowers the effort of labeling a large dataset, which is very expensive and time-consuming, while maintaining high accuracy rates.

IV. METHODOLGY



A. Data Collection

Live data is streamed from twitter through API. The dataset is collected with particular hashtag. The imported data is stored in json file format for further processing. The hashtag can be changed by user for any specific handle to retrieve the dataset.

B. Data Preprocessing

The dataset imported in json format is considered for further processing. The dataset is extracted only few important columns for the study. The dataset is split into D1 and D2, where is marked to be labeled dataset with the following rules. The user is considered to be genuine, if Followers count >30, Number of Tweets posted >50 and Number of tweets liked >0, else the user is considered to Spam user. With this labeled dataset D1, we used KNN algorithm to create the labeled dataset of D2. Then D1 and D2 is combined to get the final dataset.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue II Feb 2022- Available at www.ijraset.com

C. Label Data Using YATSI

The preprocessed dataset from module 2, dataset D2 is considered to apply YATSI. YATSI is a semi-supervised classification algorithm that can be built on top of any supervised classification algorithm and the nearest neighborhood algorithm. YATSI consists of two stages. In the first stage, an initial prediction model, which generated on the training set and prediction for unlabeled instances are determined by using a supervised classifier. In the second stage, the actual predictions for unlabeled instances are determined by using the nearest neighborhood algorithm. It has been proven that changing the value of the weighting parameter F, which gives as much weight to the training set as to testing set, improved the YATSI performance.

The weight is used to reduce the influence of the test set, and it can be calculated by the following equation: weight = p *(training instances/test instances), where p is a parameter that can be defined by a user to raise or lower the importance of the test-set.

D. Cloud Storage

The resultant dataset from the pre-processing is stored in cloud storage. We used drivehq.com for storing the dataset. Dataset is stored as CSV (Comma separated values) file and can be accessed through file transfer protocol for further processing and spam detection modules.

E. Twitter Spam Prediction

Dataset prepared from the above module was take for study, and the dataset was split into 70% training dataset, 30% test dataset. Machine learning algorithm such as KNN, logistic regression, Naive Bayes, and Support Vector machine is considered for learning and spam detection. The accuracy and error values such as mean absolute error (MAE), mean squared error (MSE) , root mean squared error (RMSE) and R-squared error was considered for study.

V. YATSI ALGORITHM

Algorithm: High level pseudo code for the two-stage YATSI algorithm

Input: a set of labeled data D_l and a set of unlabeled data D_u, an of-the-shelf classifier C and a nearest neighbor number K; let $N = |D_l|$ and $M = |D_u|$

Step 1:
Train the classifier C using D_l to produce the model M_l
Using the model M_l to "pre-label" all the examples from D_u
Assign weights of 1.0 to every example in D_l
and of $F \times (N/M)$ to all examples in D_u
Merge the two sets D_l and D_u into D

Step 2:
For every example that needs a prediction:
Fined the K-nearest neighbors to the example from D to produce set NN
For each class:
Sum the weights of the examples from NN that belong to that class
Predict the class with the largest sum of weights.

Fig. 1. YATSI Algorithm (Driessens et al. 2006)

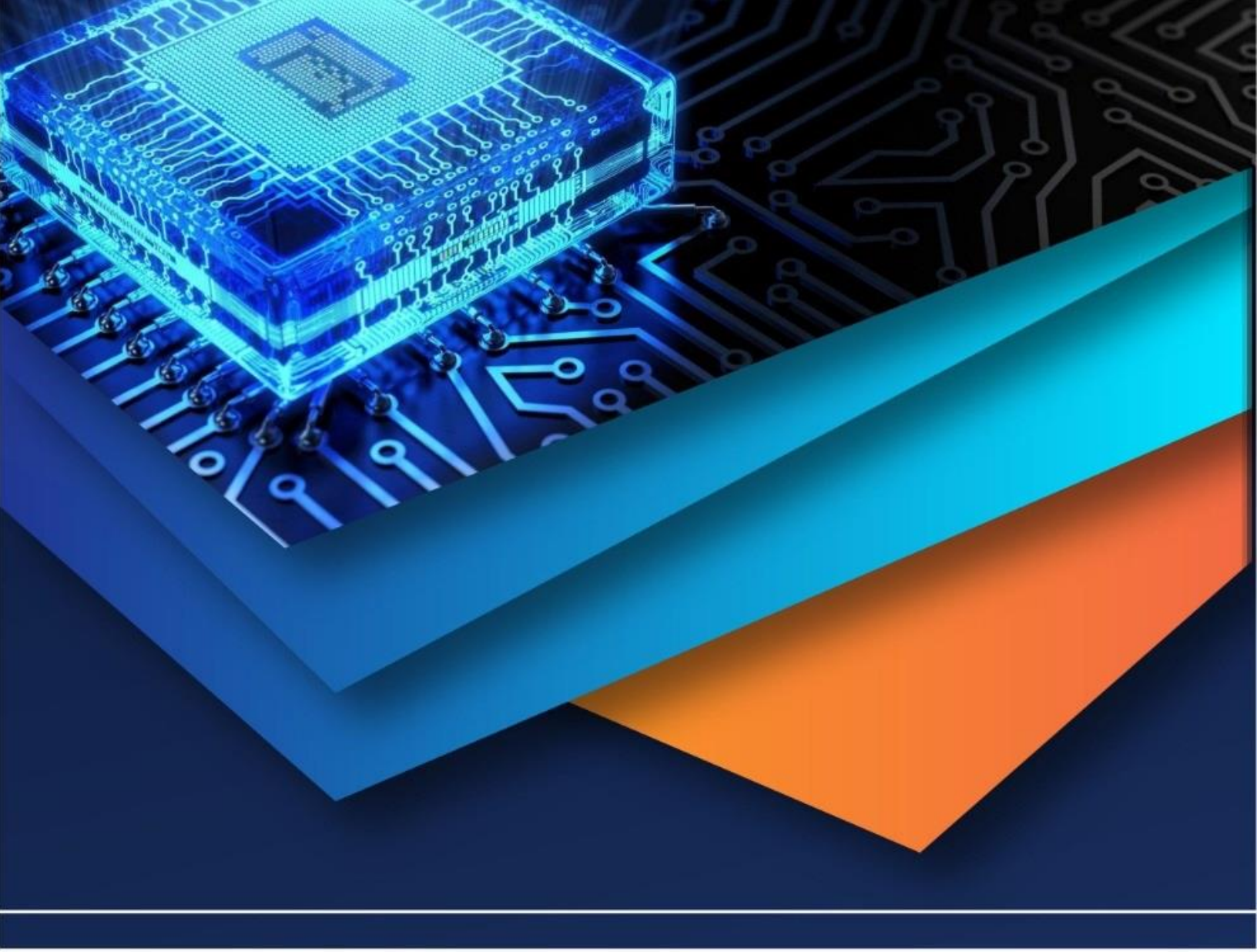
VI. CONCLUSION

Twitter popularity has not only attracted more users, but it also makes it a very attractive platform for spammers. As the number of spammers is growing rapidly, the research community and Twitter have been developing different spam detection systems to protect users. Various machine learning approaches have been proposed to detect spam tweets. However, a recent study pointed out a new problem in Twitter detection systems called drifted twitter spam. The study shows that spam tweet characteristics are changing over time, which affect the performance of the traditional ML algorithms. Consequently, the proposed system introduced a new approach that can reduce the effect of Twitter spam drift while detecting spam tweets. The proposed approach, SSLA, is a semi-supervised ML technique that uses the unlabeled data to learn the structure of the domain. Thus, it can detect spam tweets with high accuracy even when the Twitter spam drift problem occurs. SSLA uses the YATSI algorithm, which can be built on top of any supervised machine learning algorithms. One of the advantages of using YATSI is that it usually improves the predictive performance of the base classifier. Experiments were carried out, and the results showed that high accuracy of spam detection is achieved.

As future work, we are interested to analyze using some deep learning models such as neural networks and Convolutional Neural networks (CNN).

REFERENCES

[1] C. Chen, J. Zhang, Y. Xie, Y. Xiang, W. Zhou, M. M. Hassan, A. AlElaiwi and M. Alrubaian, A performance evaluation of machine learning-based streaming spam tweets detection, IEEE Trans. Comput. Soc. Syst. 2(3) (2015) 65–76.
[2] M. Mateen, M. Iqbal, M. Aleem and M. Islam, A hybrid approach for spam detection for Twitter, 2017 14th Int. Bhurban Conf. Applied Sciences and Technology (IBCAST) (Islamabad, Pakistan, 2017), p. 466.
[3] C. Grier, K. Thomas, V. Paxson and M. Zhang, @ spam: The underground on 140 characters or less, in Proc. 17th ACM Conf. Computer and Communications Security, ACM, New York, NY, USA, 2010, pp. 27–37.
[4] F. Benevenuto, G. Magno, T. Rodrigues and V. Almeida, Detecting spammers on twitter, Collaboration, Electronic Messaging, Anti-Abuse and Spam Conf. (CEAS), 13–14 July 2010, Redmond, Washington, US, pp. 1–10.
[5] N. El-Mawass and S. Alaboodi, Detecting Arabic spammers and content polluters on Twitter, 2016 Sixth Int. Conf. Digital Information Processing and Communications (ICDIPC), 2016, pp. 53.
[6] C. Chen, J. Zhang, X. Chen, Y. Xiang and W. Zhou, 6 million spam tweets: A large ground truth for timely Twitter spam detection, in 2015 IEEE Int. Conf. Communications (ICC), 2015, pp. 7065–7070.
[7] A. Al-Zoubi, J. Alqatawna and H. Faris, Spam pro¯le detection in social networks based on public features, 2017 8th Int. Conf. Information and Communication Systems (ICICS), April 4–6 2017, Irbid, Jordan, pp. 130–135.
[8] Z. Miller, B. Dickinson, W. Deitrick, W. Hu and A. H. Wang, Twitter spammer detection using data stream clustering, Inf. Sci. 260 (2014) 64–73.
[9] N. Eshraqi, M. Jalali and M. Moattar, Detecting spam tweets in Twitter using a data stream clustering algorithm, 2015 Int. Congress on Technology, Communication and Knowledge (ICTCK) (Mashhad, Iran, 2015), pp. 347–351.
[10] C. Chen, J. Zhang, Y. Xiang and W. Zhou, Asymmetric self-learning for tackling twitter spam drift, 2015 IEEE Conf. Computer Communications Workshops (INFOCOM WKSHPS), 26 April – 1 May 2015, Hong Kong, China, pp. 208–213.
[11] Twitter Help Center, Reporting spam on Twitter, https://support.twitter.com/articles/64986.
[12] C. Chen, Y. Wang, J. Zhang, Y. Xiang, W. Zhou and G. Min, \Statistical features-based real-time detection of drifted Twitter spam, IEEE Trans. Inf. Forensics and Sec. 12(4) (2017) 914–925.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY