IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

An Investigation of Class Imbalance Nature on Twitter Spam Detection

Mr. Syed Aamiruddin
Teaching cum Research Fellow
Netaji Subhash University of Technology New Delhi Teaching cum Research Fellow

ABSTRACT

In recent years, the ever-increasing popularity of social networks like Twitter have become an important source for real-time information which offers unprecedented opportunities to aggregate people and news dissemination, since it is so rapidly updating which makes it easy to fall into the trap of believing everything as truth which opens new modalities for cyber-crime perpetrations and people become a prime target of spammers. The paper proposes the various classification learning techniques methods for the detection of the Twitter- spammers by using Support Vector Machine, Naïve Bayes, extreme gradient boosting, Random Forest and neural network. The paper mainly focusses on the crispy set and not on the fuzzy set because the problem is to detect whether a tweet is a spam or not, so the annotation which is used to classify the tweets, "0" for the non-spam and "1" for spam thus it is a binary classification problem which is included in the crispy set. Initially, the tweets were first extracted manually and extracting features for classification. After selection of the features, five machine learning algorithms were cross-validated to determine the best base classifier for spam detection. Followed by handling class imbalance problem in which the proportion of the positive class is very small in comparison to the negative class which is very large. This problem is dealt with techniques which helps in improving the performance of the classifier even with the imbalanced dataset. Results showed that the proposed approach has Random Forest base classifier as the best traditional algorithm along with this data sampling technique Random undersampling showed the best sensitivity value of 93.4%.

Keywords – Twitter Spam, Class imbalance, Machine learning, social network security

1.INTRODUCTION

Since the very moment the first computer was created, spamming became a possibility as this is the only prerequisite for the spamming to exist, it does not matter what type of network it is being used on, because what spam looks like and how it works depends on its environment.

In twitter users usually, communicate with each other by sending and receiving tweets that can be viewed and are retweeted by the users' followers. There are two types of friends exist in the Twitter: "followers" and "followings". Along with this sometimes Twitter is being used by big companies to keep a check on the satisfaction of their customers but this popularity attracts attention of the spammers as well, even the trending topic on which the maximum number of hashtags are associated is consider easy to use by spammers to add malicious content with the trending topics and most of the users open the links which are shared by their friends in spite of the fact that the user has never met the user who has shared

the link. Twitter spam is usually referred to as the unsolicited tweets that contain the malicious links directing victims which is highly irrelevant and useless to the user it is being tweeted to floods the users' page with malicious content without permission. In simple terms the word spam signifies everything repetitive, inattentive, and vexing Thus in order to understand how spammer users operate in the Twitter community, it is useful to

recall the social network behavior. The traditional approach which is being followed by the Twitter to detect and filter spam users' is based on blacklists, twitter also implements a blacklists filtering module which is their anti-spam system "BotMaker". Nonetheless, the approach which is used to protect victims through the blacklist based schemes failed due to the time lag, as the victims used to already click on the spam message before the message is being blocked by the "BotMaker", therefore, the process adopted by the Twitter to protect the victims failed.

In order to overcome the limitation of blacklisting, various machine learning based schemes that detect Twitter Spam by mining the spammers' and spam tweets. This paper presents a methodology to analyze users' behaviors and the content of the tweets, in order to find the best features to identify twitter spammers. The research follows the number of steps; the very first step includes the collection of the tweets using the tweeter API in which the user builds a handshake between the client and the server, till the time the connection is established the user can extract the tweets and save the collected tweets in the CSV (Comma separated values) file in which the data is stored in the tabulated form. The second step includes the features that can differentiate spam from non-spam are selected and extracted from the tweets collection. Third, the tweets are labeled (as spam or non-spam) based on the features selected along with this a small set of training dataset is being selected from the entire dataset. Finally, various machine learning algorithms are applied, algorithms like Support Vector Machine, Naïve Bayes, Random Forest, XgBoost and neural network to develop classification models, which can then be deployed to detect spam in the real-time. SVM is a supervised machine learning algorithm, where we plot each data item as a point in n-dimensional space (where n is a number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyperplane
that differentiates the two classes very well. Sumaiya Pathan and R. H. Goudar in [14] giving description about the Support Vector Machine model for detection for spam in twitter. NB algorithm is a classification algorithm whose most important characteristic is its assumption of independence among the attributes present in the dataset. The maximum posterior probability principle is used to select the best value for the class [3]. RF is a meta-learning approach that uses multiple random decision trees as base learner. The main characteristic of RF is that it contributes to reducing the influence of the variance in the total error. This is due to its selection method among different decision trees, each one of which is trained on a random sample with replacement based on the original data, and the fact that it optimizes on a subset of the set of attributes in every step, rather than the entire set [3]. Xgboost is the technique where the word boosting stand for refers to a family of algorithms which converts weak learner to strong learners. In this algorithm, multiple rules are present to classify a target variable into one to of the two class in the case of a categorical problem. Vina Ayumi in [26] explained about the Xgboost model interpretation for the classification problem. NN is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. NN used in this paper is multilayer feedforward network in which there is input layer, hidden layer for the computation purposes and the output layer, in this network the direction of motion is from the input layer towards the hidden layer and computed result is forwarded to the output layer. Yoshua Benjio, Rejean Ducharme, Pascal Vincent and Christian Jauvin in [13] describing about the neural model where computation process is both in forward and backward phases.

This paper also takes into consideration about the class imbalance problem that widely exists in the real world Twitter data. Imbalance class distribution is that the proportion of the spam tweets in the data set is much smaller than that of the non- spam tweets in reality. In the machine learning algorithms, the smaller set may be considered as noise and the model only takes into consideration the larger dataset. To address the problem methods are present to deal with the imbalance classification, the methods are Random under-sampling, Random oversampling and Synthetic data generation. Random under-sampling it reduces the number of observations from majority class to make the data set balanced. Random oversampling replicates the observations from minority class to balance the data. Synthetic data generation instead of replicating and adding the observations from the minority class, it overcomes imbalances by generates artificial data. All these techniques generate a number of balanced data sets from the original imbalanced training data set by adjusting the class distribution. The data used for classifying the tweets as spam and non-spam does not focus on the uncertain nature of the data (tweets) thereby not using a fuzzy oversampling techniques to deal with the imbalance classification problem. The next step aims to train a classification model from each of the redistributed data sets. The final steps combine the predictions from all the classification models to reach a final decision using a majority voting scheme.

Section 2 presents background and related work regarding spammers' profile, Section 3 presents the evaluation of the base machine learning algorithms, Section 4 presents about the class imbalance problem, Section 5 compares between ensemble learning algorithms with the non-ensemble learning algorithms, Section 6 presents the conclusion and the future work.

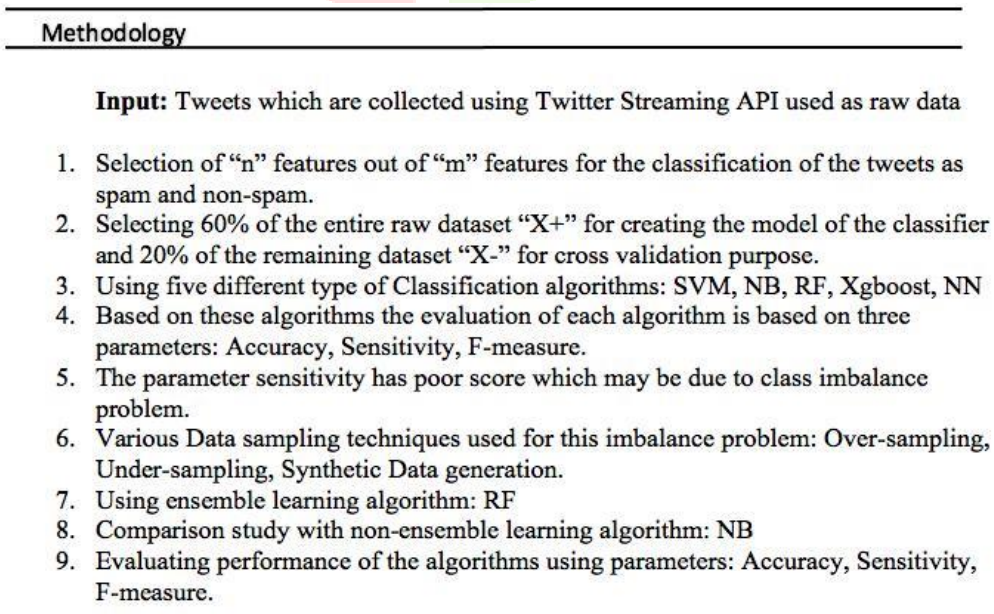
2. BACKGROUND AND RELATED WORK

Social networks are a popular mean to share data and ideas. Peoples spends a lot of time on the social networking sites to stay updated and also to make new friends and submitting messages, information. Nowadays social networking sites have become a huge source of big data but spam mitigation is one of the key security challenges in online social networks. Twitter on the other side has become the popular site for breaking news and entertainment to sports and politics. Since Twitter is all about the trust network thus it is one of the social sites where it is easy to understand the spread of spammer profiles. Twitter reports seven behavior that defines who a spammer is [4]:

- Posting harmful malicious links
- Aggressive following behavior
- Abusing the @ reply or @ mention function to post unwanted messages to users
- Creating multiple accounts
- Posting repeatedly about trending topics to try to grab attention
- Repeatedly posting duplicate updates
- Posting links to unrelated tweets

Initially, various heuristic rule-based methods for filtering tweets have been developed to overcome the limitations of blacklisting which include three rules, which are suspicious URL search, username pattern matching and the keyword detection. The proposed idea was to remove all the tweets that are suspicious so as to eliminate the impact of spam but this approach was also not suitable enough as a large number of tweets which were not even spams were removed thus this approach also failed. Therefor this paper proposes various machine learning techniques for identifying tweet as spam or non-spam based on the range of features.

A number of related research projects, such as [20], [14], [3] and [6] were found. G. Edward in [20] aimed at using base classifier Random Forest that would allow users to detect the tweets as spam or non-spam. The result of the G. Edward approach is having ACC of 81.2% and SEN of 78%. Sumaiya Pathan and R.H. Goudar in [14] aimed at using Support Vector Machine model for detection of spam messages by classifying tweets into two categories, a good message, and a bad message. Abdulla Talha Kabakus and Resul Kara in [3] conducted a survey based on the features of Twitter Spam detection, the features used in the paper are account-based features, tweet based features, graph-based features and hybrid-based features. Georvic Tur and Masun Nabhan Homsi in [6] used cost-sensitive classifiers for Twitter Spam detection on news media account, the paper also discusses about the class imbalance problem and solved imbalance classification using resampling method.
Twitter spam detection is a critical problem due to two main issues: tweet messages are long up to 140 characters and it is difficult to find a good representation for both tweets and user [2]. The workflow employed to construct the new classifier:


Methodology

Input: Tweets which are collected using Twitter Streaming API used as raw data

1. Selection of "n" features out of "m" features for the classification of the tweets as spam and non-spam.
2. Selecting 60% of the entire raw dataset "X+" for creating the model of the classifier and 20% of the remaining dataset "X-" for cross validation purpose.
3. Using five different type of Classification algorithms: SVM, NB, RF, Xgboost, NN
4. Based on these algorithms the evaluation of each algorithm is based on three parameters: Accuracy, Sensitivity, F-measure.
5. The parameter sensitivity has poor score which may be due to class imbalance problem.
6. Various Data sampling techniques used for this imbalance problem: Over-sampling, Under-sampling, Synthetic Data generation.
7. Using ensemble learning algorithm: RF
8. Comparison study with non-ensemble learning algorithm: NB
9. Evaluating performance of the algorithms using parameters: Accuracy, Sensitivity, F-measure.

Fig. 1 Workflow employed to construct the new spam classifier.

A. Dataset

Collection of 10032 Tweets from Twitter using Twitter Streaming API and based on the features manually ascribing the tweets into two categories, spammers, and non-spammers. From the dataset 6019 tweets were used as a training set in experimentation while 2008 tweets were used as a testing dataset. Training phase aims at the construction of a model in classifier.

B. Pre-Processing phase

The preprocessing phase is the most important phase in classifying a tweet as spam or non-spam since it has a very crucial impact on the computational performance of any machine technique. In this phase, the tasks applied in this phase are Feature Extraction and the labeling of the tweet as one of the two categories

1. Feature Extraction:
   - Number of URLs: URLs included in tweets contain link to a web address on the world wide web. This feature is used by spammer as "URLs shortening". This is a technique that allows spammer using fewer characters inside a tweet but hides the presence of malicious URLs. In this spammer includes a shortened dangerous URL in his tweets to entice unaware users to access to unsafe websites thus the first feature is to check the presence of URL or not in a tweet [4].
   - Character Length: Since a tweet can be at max 140 characters' long, the second feature is to check the length of the tweet by counting the number of characters
   - Friends count and followers count: Friends count is the number of friends the user is following and the followers' count is the number of different users is following the user. If a user has a large number of friends and fewer followers, the user can be a spammer. The third feature is to keep count of the followers and friends
   - The reputation of user: It is defined as a ratio of the number of followers divided by the sum of the number of followers and number of friends.
   - Status count: The status count depicts the number of tweets including the retweet issued by the user, this last feature will tell about the activities and about the active time of the user.

2. Labeling: Based on the features selected the tweets are classified as one of the two categories as spam or non-spam based on the range which is being imposed on each of the features which is selected for the classification followed by annotating each tweet for spam or non-spam by "1" and "0" respectively.

C. Classification Phase

Five algorithms are used in the classification phase. These classification algorithms are trained and tested separately on training dataset, using cross-validations. These algorithms were Support Vector Machine (SVM), Naïve Bayes (NB), Random Forest (RF), Extreme Gradient Boosting (Xgboost), Neural network

D. Evaluation Phase

Machine learning algorithms' performances were evaluated by using three parameters, accuracy (ACC), sensitivity (SEN) and F-Measure. The reason for using sensitivity as one of the performance parameters is because although the accuracy of many model configurations may be fairly good, they could still be useless if the SEN is not high enough. This problem is aggravated by the class imbalance present in the data, given the fact that genuine instances of tweets considered spam are less in number than the healthy ones. The confusion matrix for evaluating the parameters is as follows:

		Prediction	
		Spam	Not Spam
True	Spam	a	b
	Not Spam	c	d

where "a" represents the number of spams that were correctly classified which is TP (true positive), "b" represents the number of spams that were falsely classified as non-spam which is FN (False negative), "c" represents the number of non-spam messages that were falsely classified as spam which is FP (False positive), and "d" represents the number of

non-spam users that were correctly classified which is TN (True negative).

Accuracy refers to the percentage of correctly classified tweets and it is given by [4]

$$ACC = \frac{TP+TN}{TP+TN+FN+FP}$$

SEN is equivalent to the true positive rate, and it represents the proportion of positives which are correctly classified [3]

$$SEN = \frac{TP}{TP+FN}$$

F-measure is a ratio of twice the product of accuracy and sensitivity divided by summation of accuracy and sensitivity

$$\text{F-measure} = \frac{2(\text{Accuracy}*\text{Sensitivity})}{\text{Accuracy} + \text{Sensitivity}}$$

3. EVALUATION OF BASE LEARNING ALGORITHMS

3.1 Dataset:

The distribution of tweets over a class for both training and testing sets are illustrated in TABLES 1 and 2 respectively. Out of the entire dataset, the training set constitutes 60% which is 6019 tweets, it can be noticed that 293 tweets out of 6019 in training set were tagged manually as spam on the basis of features selection while the remaining as non-spam. Similarly, for testing dataset 83 tweets are tagged as spam while remaining as non-spam.

TABLE 1

Spam	Non-Spam	Total
293	5726	6019

TABLE 2

Spam	Non-Spam	Total
83	1925	2008

3.2 Evaluation:

In this subsection, five different type of experiments is employed based on five different machine learning algorithms for Twitter spam detection. The goal of these experiments is to determine the best traditional algorithm based on the two parameters, accuracy and sensitivity for each classifier. The following points are observed using TABLE 3:

TABLE 3

Classifier	Accuracy	Sensitivity	F-measure
SVM	95.8	62.3	75.5
NB	98.7	81.3	89.1
RF	95.9	84.5	89.3
Xgboost	97.4	77.6	87.8
NN	95.7	25	39.8

- RF classifier yielded the highest rates of sensitivity 84.5% and F-measure for the spam detection in comparison to the other classifiers.
- The result of SVM classifiers to detect tweets as spam is not up to the mark to become the best classifier in spite of having high accuracy because the sensitivity value is low and also the F-measure value is low when compared with other classifiers.
- Although the NB has the highest accuracy but the sensitivity and F-measure value of the classifier is low therefore NB is not appropriate to be used in spam classification because accuracy is not the parameter in the selection of the best classifier.
- Xgboost classifier showed the moderate result with 97.4% accuracy (second highest accuracy), with 77.6% sensitivity (third highest accuracy) and F- measure 87.8%.
- The best results were 95.9% of ACC, 98.5% of SEN, 89.3% of F-measure, exhibited by RF classifier in the experiment. Therefore, the best base classifier is Random Forest.

From the above observations, the conclusion that can be drawn is that the poor score of the SEN may be due to the class imbalance problem. Therefore, experiment for the class imbalance is performed.

4. CLASS IMBALANCE IN TWITTER SPAM DETECTION

As in many data mining application domains, class imbalance problem exists in real-time, also in the case of Twitter spam detection faces the problem of class imbalance as in a data set which is collected randomly from the Twitter consist of the number spam tweets which are usually very less as compared to the number of non-spam tweets (i.e., x<<y). In this work, we define the class imbalance rate in the dataset as:

$$= \overline{x\ y}$$

The class imbalance rate can be varying depending on the activeness of spammers during the data collection period. Dataset of 10032 tweets showing that 4.57% of the collected data tweets are spam. This yields a class imbalance rate over 19. However, most of the existing base classifier for spam detection is carried out without class imbalance problem in mind, so that their results are obtained from relatively balanced data sets. Thus the result of the base classifier Random Forest may not the actual performance of the spam classifier.

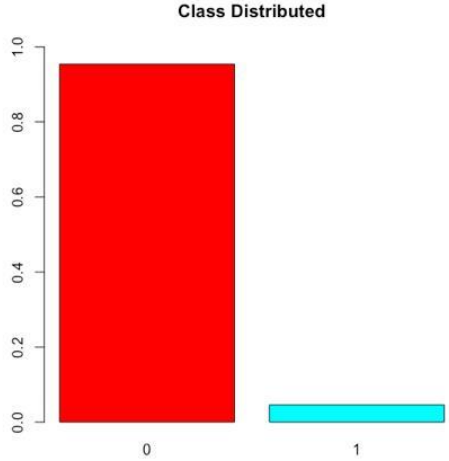


TABLE 4

SPAM(%)	NON-SPAM(%)
4.57	95.4

The plot of the class distribution where “0” represents non-spam and “1” represents spam and TABLE 4 is the percentage distribution of the spam and non-spam.

4.1 Data Sampling Techniques

In order to address the problem of imbalance classification in Twitter spam detection, an ensemble learning approach that incorporated a majority voting scheme to combine various base classifiers models [5]. In this approach, each model is created independently upon a data set that is re-balanced from given imbalanced dataset using the following type of data sampling techniques random oversampling, random undersampling and synthetic data generation.

Random oversampling (ROS): This method works with minority class. Random oversampling balances the data by randomly oversampling the minority class by replicating the observations from minority class to balance the data. In this method, it does not lead to the information loss but it ends up adding multiple observations of several types.

Random undersampling (RUS): This method works with majority class. Random under-sampling method randomly chooses observations from majority class which is eliminated until the data set gets balanced. This method reduces the number of training samples which helps to improve runtime and storage troubles.

Synthetic Data generation: This method, instead of replicating and adding the observations from the minority class, it overcomes imbalances by generates artificial data. In regards to synthetic data generation, synthetic minority oversampling technique (SMOTE). SMOTE algorithm creates artificial data based on feature space (rather than data space) similarities from minority samples.

4.2 Dataset:

Two steps are carried out to mitigate the problem of overfitting by balancing classes and applying ensemble learning algorithm which is Random forest algorithm. The first step is resampling with data sampling techniques. The second step includes using ensemble learning rule and evaluating the performance on the basis of the Accuracy(ACC) and sensitivity(SEN).

TABLES 5, 6,7 and 8 depicts the imbalance dataset with only 4.8% of the training data as spammers, the random oversampling dataset, random undersampling dataset and synthetic data generation respectively.

TABLE 5:

Spam	Non Spam	Total
5726	5726	11542

TABLE 6:

Spam	Non spam	Total
293	5726	6019

TABLE 7:

Spam	Non-Spam	Total
293	293	586

TABLE 8:

Spam	Non-spam	Total
3942	4058	8000

4.3 Evaluation:

In this subsection, three different type of experiments employing ensemble learning algorithms for spam detection and classification using random Forest technique. The goal of these experiments is to determine the best data sampling technique with ensemble leaning algorithm. The following points are observed are using TABLE 9

TABLE 9

Classifier	Accuracy	Sensitivity	F-measure
ROS (RF)	95.81	89.3	92.4
RUS (RF)	95.97	93.4	94.7
Synthetic Data (RF)	95.82	89.7	92.2

Data sampling technique evaluation using ensemble learning algorithm (Random Forest)

- RUS (random undersampling) has yielded the highest sensitivity of 93.4 and F-measure of 94.7 among all the other data sampling techniques.
- ROS and Synthetic data has yielded almost same sensitivity value and F-measure value, both the classifier is working similarly in the twitter spam detection
- The best results are 95.97% Accuracy, 93.4% sensitivity, 94.7% F-measure exhibited by RF Random undersampling technique.

5. ENSEMBLE LEARNING VS NON ENSEMBLE LEARNING

In this section, two different type of experiments employing as comparative study for spam detection and classification using random Forest technique which is ensemble learning algorithms other using Naïve Bayes which is not ensemble learning algorithm. The following observation is made using TABLE 9 and TABLE 10:

TABLE 10

Classifier	Accuracy	Sensitivity	F-measure
ROS (NB)	53.6	13.67	21.7
RUS (NB)	52.0	11.14	18.3
Synthetic Data (NB)	41.9	7.12	12.1

Data sampling technique evaluation using non ensemble learning algorithm (Naive Bayes)

- Using non-ensemble learning algorithm, the accuracy has declined so sharply from 95.97 to 52.0 this implied that the data sampling techniques are only useful using ensemble learning algorithm.
- Not only the Accuracy has declined sharply but also the sensitivity has declined from 93.4 to 13.67 which is a very sharp fall in the sensitivity, therefore, it is not at all suitable to use non-ensemble learning algorithm after balancing the data using data sampling techniques.
- F-measure parameter value has also declined by using Naïve Bayes algorithm which is not ensemble learning algorithms as the value of F-measure has declined from 94.7 to 18.3 therefore, the data sampling techniques should be applied only with the ensemble learning algorithm.

6. CONCLUSION AND FUTURE WORK

In order to alleviate the security threat caused by the Twitter spam, this paper investigated that it was conjectured that the types of attributes which are selected or to be included in order to assign a tweet as spam or non-spam would have the major influence over the Machine Learning classifier. Based on the usefulness of the features in spammer detection for the base classifiers like Support Vector Machine, Naïve Bayes, Random Forest, Extreme Gradient Boosting, Neural Network schemes using the Twitter dataset which has been collected using Twitter Streaming API. The result showed that the Random Forest classifier gives the best performance, using Random Forest classifier achieved 95.9% Accuracy and 84.5% sensitivity.

This paper also investigates the class imbalance problem in the machine learning based classifiers. It showed that the effectiveness of twitter spam detection is affected by the imbalance classification of the data which implies imbalance distribution of the spam tweets and the non-spam tweets which is seen in the real-time data very often. Experiments have been conducted using the data sampling techniques which balances the data followed by using ensemble learning algorithm, the result shows that the proposed approach can improve the twitter spam detection performance on imbalanced Twitter datasets. The Random undersampling data sampling technique showed the best result with Random Forest with 95.97% Accuracy and 93.5% sensitivity.

As a future direction for this project, a specialization of spam detection that could take into account features like distribution of tweets over the 24-hour period, replies/retweets, may also take into account the writing style of each tweet in order classify the tweet in the category of spam and non-spam. The reason for proposing this is that each one of these accounts has a user base that agrees with the own style of writing [5] thus spammer would be identified with a particular way of reporting, but this will only take into consideration with those traits that make the tweets irrelevant for the users. In addition to this fuzzy based data sampling technique could also be used with the cost-sensitive classifier, along with this sentiment analysis [5] could also be performed on the tweets in order to categorize the tweets as the subjective or objective which could also be used as one of the feature to label the tweets as spam or non-spam.

REFERENCES

[1] Liu S,Wang Y, Zhang J,ChenC, Xiang Y. "Addressing the class imbalance problem in twitter spam detection using ensemble learning", Computer Security 2016; 69:35-49

[2] Claudia Meda, Federica Bisio, Paolo Gastaldo and Rodolfo Zunino: "A Machine Learning Approach for Twitter Spam Detection", Security Technology (ICCST), 2014
International Carahan Conference, Rome, Italy (2014).

[3] Abdullah Talha Kabakus and Resul Kara: "A Survey of Spam Detection Methods on Twitter", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 8, No. 3, 2017.

[4] McCord M., Chuah M."Spam Detection on Twitter Using Traditional Classifiers. In: Calero J.M.A., Yang L.T., Mármol F.G., García Villalba L.J., Li A.X., Wang Y. (eds) Autonomic and Trusted Computing. ATC 2011. Lecture Notes in Computer Science, vol. 6906. Springer, Berlin, Heidelberg

[5] M. Rajdev and K. Le,"Fake and Spam Messages: Detecting Misinformation During
Natural Disasters on Social Media", 2015 -IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), 2015.

[6] G. Tur and M. N. Homsi, "Cost-sensitive classifier for spam detection on news media Twitter accounts," 2017 XLIII Latin American Computer Conference (CLEI), Cordoba, 2017, pp. 1-6.

[7] F. Brunton, "Spam", 1st ed. Cambridge, MIT Press Ltd, 2013.

[8] K. Murphy, **"**Machine learning**"**, 1st ed. Cambridge**,** Massachusetts: The MIT Press, 2012.

[9] M. Nabhan Homsi, N. Medina, M. Hern´andez, N. Quintero, G. Perpi˜nan, A. Quintana
and P. Warrick,"Automatic heart sound recording classification using a nested set of ensemble algorithms", Computing in Cardiology Conference,Vancouver,BC Canada(2016).

[10] Chaoliang Li and Shigang Liu**: "**A comparative study of the class imbalance problem in Twitter spam detection", Concurrency Computation: Practice and Experience, Wiley 2017; e428.

[11] Mikel Galar, Alberto Fern´andez, Edurne Barrenechea, Humberto Bustince and Francisco Herrera: "A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches", IEEE Transaction on Systems, Man, and Cybernetics –Part C: Applications and Reviews.

[12] Adnan Amin, Sajid Anwar, Awais Adnan, Muhammad Nawaz, Newton Howard, Junaid Qadir, Ahmad Hawalah and Amir Hussain: "Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study", JOURNAL OF IEEEACCESS.

[13] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, Christian Jauvin : "A Neural Probabilistic Language Model", Journal of Machine Learning Research 3 (2003) 1137–1155, Published 2/03.

[14] Sumaiya Pathan and R. H. Goudar: "Detection of Spam Messages in Social Networks based on SVM", International Journal of Computer Applications (0975 – 8887) Volume 145
– No.10, July 2016.

[15] Agarwal S, Jain K "Hybrid Approach for Spam Detection using Support Vector
Machine and Artfical Immune System", First International Conference on Network and Soft Computing", Aug 2014, pg no: 05-09.

[16] "A Hybrid Approach for Spam Filtering using Local Concentration and K- means clustering", 2014, 5th International Conference, pg no: 194- 199.

[17] D. Baker and A. McCallum: "Distributional clustering of words for text classification", In SIGIR'98,1998.

[18] Y. Bengio and J-S. Senécal. Quick training of probabilistic neural nets by importance sampling. In AISTATS, 2003.

[19] Yanping Xu, ChunhuaWu, Kangfeng Zheng, Xinxin Niu and Yixian Yang: "Fuzzy– synthetic minority oversampling technique: Oversampling based on fuzzy set theory for Android malware detection in imbalanced datasets", International Journal of Distributed Sensor Networks 2017, Vol. 13(4)

[20] Chawla NV, Japkowicz N, Kotcz A. Editorial: "Special issue on learning from imbalanced data sets", ACMSigkdd Explor Newsl. 2004;6(1):1-6.
[21] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. "Detecting spammer on twitter", in: CEAS 2010 - Seventh Annu. Collab. Electron. Message. Anti-Abuse Spam Conf., Redmond, Washington, USA, 2010
[22] C. Grier, K. Thomas, V. Paxson, and M. Zhang. "@spam: the under- ground on 140 characters or less", In Proceedings of the 17th ACM conference on Computer and communications security, CCS '10, pages 27–37, New York, NY, USA, 2010. ACM.
[23] X. Zhang, S. Zhu, and W. Liang. "Detecting spam and promoting campaigns in the twitter social network", In Data Mining. IEEE ICDM2012, pages 1194–1199, 2012.
[24] L. Breiman "Random Forests. Machine Learning", 45(1):5-32 (2001).
[25] V. Martha, W. Zhao, X. Xu, "A study on Twitter User-Follower Network", ASONAM'13, August 25-29, 2013, Niagara, Ontario, CAN.
[26] Ayumi Vina "Pose-based human action recognition with extreme gradient Boosting", Research and development (SCOReD), IEEE, Kuala Lumpur, Malaysia, 2016.
[27] C. Yang, R. Harkreader, J. Zhang, S. Shin, G. Gu, "Analyzing Spammers' Social Networks for fun and profit", – Texas A&M University, College Station, Texas, WWW 2012 Session Security and Fraud in Social Networks Akbani R, Kwek S, Japkowicz N. Applying support vector machines to imbalanced datasets. In: European Conference on Machine Learning. Pisa, Italy:Springer; 2004:39-50.
[28] S. M. A. Elrahman and A. Abraham, "A Review of Class Imbalance Problem," Journal of Network and Innovative Computing, vol. 1, pp. 332- 340, 2013.

[29] K. Ghosh, A. Banerjee, S. Chatterjee and S. Sen, "Imbalanced Twitter Sentiment Analysis using Minority Oversampling," 2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST), Morioka, Japan, 2019, pp. 1-5, doi: 10.1109/ICAwST.2019.8923218.