Comparing Common Supervised Machine Learning Algorithms For Twitter Spam Detection in Scikit-Learn

Anıl Düzgün, Fecir Duran, Atilla Özgür

Gazi University, Ankara/Turkey, anil.duzgun@gazi.edu.tr
Gazi University, Ankara/Turkey, fduran@gazi.edu.tr
Jacobs University, Bremen/Germany, a.oezguer@jacobs-university.de

***Abstract* - Twitter is one of the most widely used social networks today. Because of its wide usage, it is also the target of various spam attacks. In recent years, Spam Detection on Twitter using artificial intelligence methods became quite popular. Twitter Spam Detection Approaches are generally categorized into following types as as User Based, Content Based, Social Network Based Spam Detection. In this paper, a user based features based spam detection approach is proposed. Using a publicly available recent baseline dataset, 11 lightweight user based features are selected for model creation. These features selected for ease of computing and rapid processing since they are numeric or boolean. The advantage of user based spam detection approach is that the results are obtained more rapidly since they do not contain complex features. Selected Features are verified, default profile, default profile image, favorites count, followers count, friends count, statuses count, geo enabled, listed count, profile background tile, profile use background image. Feature verified is used as a class label to measure success of the model. After the feature selection, the dataset is divided into test and training data. Following 10 common supervised machine learning algorithms are selected for the experiments: (1) Support Vector Classification, (2) K Nearest Neighbor, (3) Naive Bayes, (4) Decision Tree, (5) Bagging, (6) Random Forest, (7) Extra Trees, (8) AdaBoost, (9) Multi Layer Perceptron, and (10) Logistic Regression. Success of the algorithms are measured using following 9 metrics: (1) Accuracy, (2) precision, (3) recall, (4) True Positive, (5) True Negative, (6) False Positive, (7) False Negative, (8) Training Time, (9) Testing Time. The results were compared according to the metrics above.**

***Keywords* - Supervised Machine Learning, Scikit-Learn,Twitter Spam Detection.**

I. INTRODUCTION

Twitter is one of the most widely used social networks today. It is used for different purposes such as news, discussion, information sharing, questionnaire and etc. For Twitter users, after relationships are built, they can receive tweets, usually something interesting or recent activities shared by their friends. Nowadays, Twitter has largely shortened the distance between people, and reshaped the way they communicate with each other [1].

According to 2018 statistics, Twitter has 336 million users and 157 million active users [2] .Recently, banks and financial institutions in the USA have started to analyze Twitter and Facebook accounts of loan applicants before actually granting the loan [3].

A versatility and spread of use have made Twitter the ideal arena for proliferation of anomalous accounts, that behave in unconventional ways. These malicious accounts, commonly known as bots, often tries to mimic real users. Recently, media reported that the accounts of politicians, celebrities, and popular brands featured a suspicious inflation of followers. As a first example, during the 2012 US election campaign, the Twitter account of challenger Romney experienced a sudden jump in the number of followers. Later, majority of these followers had been claimed as fake users. As a second example, before the last general Italian elections (February 2013), online blogs and newspapers had reported statistical data over a supposed percentage of fake followers of major candidates [3]. As a final example, malicious bots and misinformation networks on Twitter may have been used in the 2016 US presidential elections [4] .

Due to above reasons, spam detection and fake user detection on Twitter has been an important matter [3]. Twitter Spam Detection Approaches are generally categorized into following types as User Based, Content Based, Social Network Based Spam Detection [5] [6] [7] [8] [9].

Chen at al. used Random Forest, Decision Trees (C4.5), Bayes Network, Naive Bayes, K Nearest Neighbor, Support Vector Machine algorithms for Twitter spam classification in two different studies [5], [6]. They compared these algorithms using True Positive, False Positive, F-Measure metrics. Zheng et al. [7] used Support Vector Machines, Decision Tree, Naive Bayes, Bayes Network algorithms for Twitter Spam classification. They compared these algorithms using Precision, Recall, F-measure metrics. Jeong et al. [8] used Decision Trees (J48) and Random Forests algorithms for Twitter spam classification. They compared their results using True Positive and False Positive metrics. Miller at al. [9] preferred clustering methods instead of classification and used

DenStream and a modified version of StreamKM called StreakKM++. They compared their results using Specificity, False Positive, Accuracy, Balanced Accuracy, Precision, F-measure and, Recall metrics.

As can be seen from above examples, most of the studies in the literature used limited number of machine learning algorithms and limited number of performance metrics for comparison purposes. This study aims to fill this gap using 10 different machine learning algorithms and compare algorithms using 9 performance metrics in Sci-kit learn machine learning toolbox [11]. Similar studies that compares machine learning algorithms using different metrics exists in other domains [10], [13].

II. METHODS

Firstly the Twitter dataset was obtained from the source referenced by study called “Fame for sale: efficient detection of fake Twitter followers [3]. This publicly available baseline dataset was created to help studies that wants to detect fake Twitter followers [3] . This dataset consists of 1806 Human and 3495 Fake accounts, for a total of 5301 accounts. From this dataset, 11 lightweight user based features are selected for Twitter spam detection model [12]. These features and their explanation can be seen in Table 1.

These features are selected for ease of computing, rapid processing since they are either numeric or boolean data types. The advantage of user based spam detection approach is the results are obtained more rapidly since text based complex features are less.

After the feature selection process, the dataset is divided into training (75%) and test (25%) datasets. Using Scikit-Learn Machine Learning Toolbox [11], 10 common supervised machine learning algorithms are trained for Twitter Spam Classification. These machine learning algorithms are: (1) Support Vector Classification, (2) K Nearest Neighbor, (3) Naive Bayes, (4) Decision Tree, (5) Bagging, (6) Random Forest, (7) Extra Trees, (8) AdaBoost, (9) Multi Layer Perceptron, and (10) Logistic Regression.

Performance of these algorithms are compared using following 9 metrics : (1) Accuracy, (2) precision, (3) recall, (4) True Positive, (5) True Negative, (6) False Positive, (7) False Negative, (8) Training Time, (9) Testing Time.

III. RESULTS

The results of the experiments are given on Table 2.
Experiments are conducted on only one computer. Its configuration is following :

1. Windows 10 64 Bit
2. Intel Core I i7-2670QM CPU @ 2.20 GHz
3. 4 GB RAM
4. Python 3.6.1. 64 Bit
5. Scikit Learn version is 0.19.1

Usually, the most important metric for machine learning systems are accuracy. From this point of view, except for Support vector machines (0.845), Naïve Bayes (0.786), Extra Trees (0.785), Multi Layer Perceptron (0.643), other classifiers have accuracy of 0.95 and above.

If implemented system is a low memory and low CPU power system like an embedded system, then training and testing time would be most important metrics, see Duran at al. [14]. According to Table 2, most of the algorithms are trained and tested below 100 milli seconds. These are very good results but training time of Support Vector Machines (20 s) is very long compared to others. Similarly, AdaBoost and MultiLayer Perception classifiers are not suitable to use in embedded systems. Since Decision Trees and Naïve Bayes classifiers are both fast to train and fast to detect (1ms), they can also be used as pre classifiers in systems which has a dynamic modelling specially in real time systems. If we look at the classifiers results with all the metrics are in our mind, best classifiers are Random Forests, Decision Tree, K Nearest Neighbor and Bagging. Interestingly, three of these four classifiers are tree based classifiers. Two of these four classifiers are ensemble classifiers (Random Forests, Bagging).

IV. CONCLUSION

Using Scikit-Learn machine learning toolbox 10 common machine learning algorithms are trained for Twitter spam classification on a benchmark dataset in this paper. Finally algorithms compared using 9 different metrics. Best results belongs to following four classifiers: Random Forests, K-nearest Neighbor, Decision Tree and Bagging.

In a future study, we aim to work on larger datasets and use additional lightweight and complex features such as content based features to improve experiment results.

Table1. Feature Description

	Feature Name	Feature Description
1	Verified	it is a class label which will measure model's success
2	default profile image	When default profile image is true, indicates that the user has not altered the theme or background of their user profile.
3	favourites count	the number of Tweets this user has liked in the account's lifetime.
4	followers count	the number of followers this account currently has. Under certain conditions of duress, this field will temporarily indicate "0".
5	friends count	the number of users this account is following. Under certain conditions of duress, this field will temporarily indicate "0".
6	statuses count	the number of Tweets (including retweets) issued by the user.
7	geo enabled	When geo enabled true, indicates that the user has enabled the possibility of geotagging their Tweets. This field must be true for the current user to attach geographic data when using POST statuses / update
8	listed count	listed count is the number of public lists that this user is a member of.
9	profile background tile	When profile background tile is true, indicates that the user's profile_background_image_url should be tiled when displayed.
10	profile use background image	When profile use background image true, indicates the user wants their uploaded background image to be used.

Table 2: Results of The Experiments

	Classifier	Accuracy	Precision	Recall	F1	True Positive	True Negative	False Positive	False Negative	Training Time (ms)	Testing Time (ms)
1	**Support Vector Machines**	0.845	0.697	1.000	0.821	0.488	0.155	0	0.155	20059	1147
2	**K Nearest Neighbor**	0.976	0.955	0.979	0.967	0.627	0.017	0.008	0.017	13	45
3	**Naive Bayes**	0.786	0.928	0.433	0.591	0.631	0.012	0.202	0.012	12	1
4	**Decision Tree**	0.977	0.953	0.985	0.969	0.626	0.017	0.005	0.017	19	1
5	**Bagging**	0.974	0.949	0.981	0.965	0.624	0.019	0.007	0.019	85	6
6	**Random Forests**	0.980	0.955	0.992	0.973	0.627	0.017	0.003	0.017	70	5
7	**Extra Trees**	0.785	0.805	0.524	0.635	0.598	0.045	0.170	0.045	33	6
8	**AdaBoost**	0.973	0.947	0.979	0.963	0.624	0.020	0.008	0.020	930	114
9	**Multi Layer Perceptron**	0.643	0.000	0.000	0.000	0.643	0.000	0.357	0.000	165	2
10	**Logistic Regression**	0.943	0.954	0.884	0.918	0.628	0.015	0.041	0.015	46	26

References

[1] C. Chen, S. Wen, J. Zhang, Y. Xiang, J. Oliver, A. Alelaiwi and M. M. Hassan, "Investigating the deceptive information in Twitter spam," *Future Generation Computer Systems,* vol. 72, pp. 319-326, 2017.

[2] C. Smith, *400 Interesting Twitter Statistics (July 2018) | By the Numbers,* https://expandedramblings.com/index.php/march-2013-by-the-numbers-a-few-amazing-twitter-stats, 2018.

[3] S. Cresci, R. D. Pietro, M. Petrocchi, A. Spognardi and M. Tesconi, "Fame for sale: Efficient detection of fake Twitter followers," *Decision Support Systems,* vol. 80, pp. 56-71, 2015.

[4] T. P. Policy, *Update: Russian interference in the 2016 US presidential election,* https://blog.twitter.com/official/en_us/topics/company/2017/Update-Russian-Interference-in-2016--Election-Bots-and-Misinformation.html, 2017.

[5] 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 Transactions on Computational Social Systems,* vol. 2, pp. 65-76, 9 2015.

[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 International Conference on Communications (ICC)*, 2015.

[7] X. Zheng, Z. Zeng, Z. Chen, Y. Yu and C. Rong, "Detecting spammers

on social networks," *Neurocomputing,* vol. 159, pp. 27-34, 2015.

[8] S. Jeong, G. Noh, H. Oh and C.-k. Kim, "Follow spam detection based on cascaded social information," *Information Sciences,* vol. 369, pp. 481-499, 2016.

[9] Z. Miller, B. Dickinson, W. Deitrick, W. Hu and A. H. Wang, "Twitter spammer detection using data stream clustering," *Information Sciences,* vol. 260, pp. 64-73, 2014.

[10] A. Ozgur, H. Erdem and A. Özgür, "Saldırı Tespit Sistemlerinde Kullanılan Kolay Erişilen Makine Öğrenme Algoritmalarının Karşılaştırılması," vol. 5, pp. 41-48, 1 2012.

[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research,* vol. 12, pp. 2825-2830, 2011.

[12] "https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/user-object," Twitter. [Online].

[13] A. Özgür and H. Erdem, "The impact of using large training data set KDD99 on classification accuracy," 3 2017.

[14] F. Duran and H. Tıraşlıoğlu, "A Multi-Purpose Mobile Application for Embedded Systems," pp. 50-55, 2017.

[15] S. Liu, Y. Wang, J. Zhang, C. Chen and Y. Xiang, "Addressing the class imbalance problem in Twitter spam detection using ensemble learning," *Computers & Security,* vol. 69, pp. 35-49, 2017.