Application-based spam detection with machine learning algorithms

Ali Erbey[1], Necaattin Barışçı[2]

[1] *Uşak University, Distance Education Vocational School, Department of Computer Programming, Uşak, Turkey*
[2] *Gazi University, Faculty of Technology, Department of Computer Engineering, Ankara, Turkey*

Abstract

Today, the use of social media sites such as Facebook, Instagram, Twitter is increasing day by day. Shares on social media can even change the agenda by reaching huge masses. On Twitter, a social media site, the agenda topics can be followed through the section called Trend-Topic. This Trend- Topic section may be manipulated by spammers from time to time. In order to avoid such unwanted situations, it is necessary to determine whether the user is spam or not. Machine learning algorithms can classify whether a user is spam or not. With machine learning algorithms, successful results are also obtained in situations such as image processing, speech, voice recognition and malware detection. In this study, machine learning algorithms Naive Bayes, K Nearest Neighbors, Random Forest, j48, Multilayer Perceptron were used to classify users. As a result of the evaluations, Random Forest algorithm, one of the machine learning algorithms used, made the most successful classification with an accuracy rate of 88%.

Keywords

Twitter, spam detection, machine learning

1. Introduction

Today, with the widespread use of the internet and the increase in the use of mobile devices, online social networking sites, social networks such as Facebook, Twitter and LinkedIn are becoming more and more popular [1]. These sites are followed by millions of people; In addition to being sites where friends, family or acquaintances can be contacted, they are also used as microblogging services, recommendation services, real-time news sources and content sharing places [2]. Users can share by creating status messages on Twitter, one of these sites. In Twitter, which is a popular microblogging site in terms of sharing, these status messages created are called tweets [3]. With these tweets sent by the users, the Trend Topic section, which constitutes the existing agenda topics, is formed.

Trend Topic section can be directed to the agenda in an undesirable way with messages sent from time to time, out of purpose. The heavy use of social media has facilitated the neglect of these environments by malicious people [4]. Wanting to change the agenda is also a method that can be neglected by malicious people. In order to prevent such omissions, many studies have been carried out in areas such as natural language processing and data mining with the data collected from Twitter [5].

When we look at the existing studies in detecting spam with Twitter posts, we see that there is a lot of work. These studies are clustered in certain areas. Twitter spam detection studies are mainly handled in three groups. These are: a) those who only examine the tweets by text mining, b) those who analyze the tweet text by associating with the user who sent the tweet, c) those who examine the relations of users with spam.

Text mining-based research mostly focuses on tweet text. In these studies, researchers first extract features and then classify them with algorithms such as Naive Bayes and j48. Feature

IVUS 2022: 27th International Conference on Information Technology
EMAIL: alierbey@gmail.com (A. Erbey); nbarisci@gazi.edu.tr (N. Barışçı)
ORCID: 0000-0002-0930-4081 (A. Erbey); 0000-0002-8762-5091 (N. Barışçı)

extraction sometimes follows feature selection to improve classification accuracy and reduce training time. Gupta and Kumar [6] used multiple linear regression to select important features. Other features such as the tweet's character count, word count, like or retweet count are used in most research.

Some research takes user characteristics into account when deciding whether a user is spam. These features can be account age, number of followers/followers, follower/followers' rate, format of the profile page. However, since these features can be easily changed by the user, they are considered to be less reliable.

Because user characteristics can change easily, some researchers have studied the relationships between spammers and real users. By examining their following / following relationships, they created a network for each user. Setting up these networks can be costly in terms of computation time, power and data collection time.

In the following sections of this study, obtaining the spammy dataset, classification, detection of spammy users and feature selection processes are carried out. In the last section, the obtained results are evaluated.

2. Material and method

We collect the dataset before making the classification. The data collection process has an important role in the classification process.

2.1. Spamming Twitter dataset

In this study, user characteristics and the method of evaluating users' tweet attributes were chosen in order to classify spam. The reason for choosing this method is that it is less costly in terms of data collection and it is seen to give better results regardless of the tweet content, as it depends on user characteristics.

For training, a topic was selected from the Turkey Trend topic list, since a data set with spam users should be obtained. 15000 tweets were collected from this trending topic with the Twitter public API. Then, repetitive data, news content, tweets containing URL only were removed from this dataset and the remaining 3798 tweets were classified as spam and not spam. As a result, 3798 tweets were classified as 1666 spam and 2132 non-spam users.

In order to evaluate which users can send spam tweets after classification, the last maximum of 100 tweets of each user were collected. Since some users did not have 100 tweets, a total of 305.604 tweets were reached. Then, these tweets were processed for each user, and some unnecessary data such as smileys were removed from the text of the tweet. Because tweets are unofficial texts, some autocorrect libraries were used and typos were corrected. Then, as a more complex process, some new features are obtained from this tweet data, such as how often the user tweets, the average number of characters of the user's tweets, or the unique words tweeted in those 100 tweets, represented as columns in the final dataset has been done.

As a result of the data collection and preprocessing stage, a dataset consisting of 3798 rows representing each user and 22 columns in total, including features such as the age of the user, whether he is a verified account, whether he entered a URL on the profile page, was obtained.

2.2. Classification

The next part after data collection is to determine whether the user is a spam user with different machine learning algorithms.

Weka software was used to classify the obtained data as spam or not spam. Weka is a program developed for machine learning and text mining intended to assist in the application of machine learning techniques [7].

In this study, Naive Bayes algorithm, k Nearest Neighbor algorithm, Random Forest algorithm, j48 algorithm and Multilayer Perceptron classification algorithms were used in Weka software.

Considering the studies in the literature, the algorithms used in other studies are shown in Table 1.

Table 1
Algorithms used in other studies

Authors	Algorithms
Diale et al.[8]	SVM, RF, c4.5
Wang[1]	DT, NN, SVM, NB
Aydın et al.[4]	DT, LR, SVM
McCord et al.[9]	RF, SVM, NB, KNN

The reason for choosing the NB, KNN, RF , j48 and MLP algorithms used in this study is to try to create a combination of algorithms that are widely used in the literature and in addition to them, less used algorithms. The reason for

choosing the most used algorithms is to make comparisons with previous studies. The reason for choosing the less used algorithms is to create an alternative to the frequently used algorithms.

2.2.1. Naive Bayes

The Naive Bayes (NB) algorithm is a simple probabilistic classifier that calculates a probability set by counting the frequency and combinations of values in a given data set. It is a classification algorithm that classifies data by calculating it with probability principles. Naive Bayes is a popular algorithm used commercially or open source for email spam filtering [11].

2.2.2. *k* Nearest Neighbors

In 1968, Cover and Hart proposed the k Nearest Neighbor (KNN) algorithm, which they have been working on for a long time [12]. The intuition underlying the K Nearest Neighbor Classification is quite simple, samples are classified according to the class of their nearest neighbors [13]. Having an efficient algorithm for performing nearest neighbor operations on large datasets can provide rapid improvements for many applications [14]. KNN is one of the useful algorithms in terms of speed.

2.2.3. Random Forest

Breiman [15] developed the Random Forest (RF) method as an extension of classification trees. In the RF algorithm, each node has a random feature selection [16]. It is an algorithm that aims to increase the classification value by using more than one decision tree.

2.2.4. j48

The purpose of the Decision Tree Algorithm is to determine how the feature vector behaves for a few samples [17]. In the WEKA data mining tool, J48 is an open-source Java implementation of the C4.5 algorithm [17].

2.2.5. Multilayer Perceptron

Multilayer perceptron (MLP) is an algorithm that can be effectively used for classification purposes and has been used a lot recently. In general, back propagation algorithm learning technique based on slope drop method is used in MLP. With this technique, the error between the desired output and the produced output is minimized [18].

2.3. Determination of spam users

The data obtained during the classification of the data was divided into 80% training data and 20% test data and evaluated in NB, KNN, RB, j48 and MLP algorithms. It is known that 1666 of 3798 users are spam users and 2132 of them are non-spam users in the dataset. The number of users to test is 760 people, which is 20% of the data. Looking at whether users are spam with the Naive Bayes algorithm, the Naive Bayes algorithm classified users with an accuracy rate of 76%.

When the complexity matrix of the NB algorithm is examined, the data are shown in Table 2.

Table 2
NB Complexity Matrix

		Predicted Class	
True Class	Class	Positive	Negative
	Positive	382	60
	Negative	122	196

According to the complexity matrix of the algorithm in Table 2; Of the 760 people in the 20% test data, 382 people who were spam were classified as spam, and 60 people who were spam were classified as non-spam. 122 non-spam were classified as spam, while 196 non-spam were classified as non-spam. When we look at the KNN algorithm, one of the machine learning algorithms, it is seen that it classifies users with an accuracy rate of 74%.

The complexity matrix of the KNN algorithm is as shown in Table 3.

Table 3
KNN Complexity Matrix

		Predicted Class	
True Class	Class	Positive	Negative
	Positive	357	85
	Negative	107	211

According to the complexity matrix of the KNN algorithm in Table 3; Of the 760 people in

the 20% test data, 357 people who were spam were classified as spam, and 85 people who were spam were classified as non-spam. 107 non-spam were classified as spam, while 211 non-spam were classified as non-spam.

When the Random Forest algorithm was used in the study, an accuracy rate of 88% was achieved.

Table 4
RF Complexity Matrix

		Predicted Class	
True Class	Class	Positive	Negative
	Positive	406	36
	Negative	53	265

According to the RF algorithm complexity matrix in Table 4; Of the 760 people in the 20% test data, 406 people who were spam were classified as spam, and 36 people who were spam were classified as non-spam. 53 non-spam classified as spam, 265 non-spam classified as non-spam.

85% accuracy rate was observed with the J48 algorithm. The complexity matrix of the J48 algorithm is as shown in Table 5.

Table 5
J48 Complexity Matrix

		Predicted Class	
True Class	Class	Positive	Negative
	Positive	382	60
	Negative	122	196

According to the complexity matrix of the j48 algorithm in Table 5; Of the 760 people in the 20% test data, 382 people who were spam were classified as spam, and 60 people who were spam were classified as non-spam. 50 non-spam classified as spam, 268 non-spam classified as non-spam.

The MLP algorithm found an accuracy rate of 82%. The complexity matrix of the MLP algorithm is as shown in Table 6.

Table 6
MLP Complexity Matrix

		Predicted Class	
True Class	Class	Positive	Negative
	Positive	390	52
	Negative	82	236

According to the complexity matrix of the MLP algorithm in Table 6; Of the 760 people in the 20% test data, 390 people who were spam were classified as spam, and 52 people who were spam were classified as non-spam. 82 non-spam people were classified as spam, while 236 non-spam were classified as not spam.

2.4. Feature selection

Feature selection is one of the important steps of pattern recognition, machine learning and data mining. Its purpose is to eliminate irrelevant and redundant variables in order to understand the data, reduce the computational requirement, reduce the dimensionality effect, and improve the performance of the predictor [19]. The sections selected by the Weka software after feature selection are shown in Table 7.

Table 7
Remaining sections after feature selection

Column Name	
Friend_count	Favourite_freq
Account_age	Reply_freq
Tweet_freq	Unique_freq
Hashtag_freg	Spam

When the data is re-evaluated after the feature selection, the performances of the algorithms are seen in Figure 1.

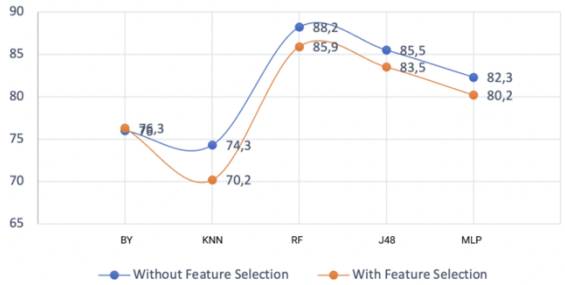

Figure 1: Performance of algorithm

As shown in Figure 1, it has obtained similar results with feature selection and without feature selection.

3. Conclusion and discussion

In this study, SPAM users on Twitter were tried to be detected. In this study, the data set consists of 3798 user information and different machine learning algorithms are used to classify

users. The Random Forest algorithm achieved the highest accuracy rate of 88%. The NB algorithm achieved 76%, KNN 74%, j48 83% and MLP 80% correct classification rates. When the algorithms were applied again after the feature selection was made, it was observed that the accuracy rate decreased in other algorithms except the NB algorithm.

In addition to the accuracy rates in the algorithms, the number of users whose real class is negative but classified as positive in the complexity matrix is also important. Since these users are classified as spam even though they are not spam, they will suffer if the algorithm is trusted. This will create an undesirable situation. When we look at the results, it is seen that the j48 algorithm gives the lowest rate with 50 users.

As a suggestion for future research, different optimizations of feature extraction and different machine learning algorithms methods can be tried on the collected data and more successful classification results can be achieved.

4. References

[1] A. H. Wang, "Detecting spam bots in online social networking sites: a machine learning approach," in IFIP Annual Conference on Data and Applications Security and Privacy. Springer. Berlin, Heidelberg, 2010. pp. 335-342 doi:10.1007/978-3-642-13739-6_25

[2] M. Pennacchiotti and A.M. Popescu, A machine learning approach to twitter user classification. in Fifth International AAAI Conference on Weblogs and Social Media. 2011.

[3] A. Go, R. Bhayani and L. Huang, Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 2009. 1(12): pp. 2009.

[4] İ. Aydın, M. Sevi and M.U. Salur, "Detection of Fake Twitter Accounts with Machine Learning Algorithms," 2018 International Conference on Artificial Intelligence and Data Processing (IDAP), 2018, pp. 1-4, doi: 10.1109/IDAP.2018.8620830.

[5] E.S Akgül, C. Ertano and B. Diri, Twitter verileri ile duygu analizi. Pamukkale University Journal of Engineering Sciences. 2016, Vol. 22 Issue 2, p106-110. 5p.

[6] D.K Gupta and A. Kumar. Spam And Sentiment Analysis Model For Twitter Data Using Statistical Learning. in Proceedings of the Third International Symposium on Computer Vision and the Internet. 2016. ACM.

[7] G. Holmes, A. Donkin and I.H. Witten, Weka: A machine learning workbench. 1994.

[8] M. Diale, T. Celik, and C. Van Der Walt, Unsupervised feature learning for spam email filtering. Computers & Electrical Engineering, 2019. 74: pp. 89-104.

[9] M. Mccord and M. Chuah. Spam detection on twitter using traditional classifiers. in international conference on Autonomic and trusted computing. 2011. Springer.

[10] T.R. Patil, and S. Sherekar, Performance analysis of Naive Bayes and J48 classification algorithm for data classification. International journal of computer science and applications, 2013. 6(2): pp. 256-261.

[11] V. Metsis, I. Androutsopoulos, and G. Paliouras. Spam filtering with naive bayes-which naive bayes? in CEAS. 2006. Mountain View, CA.

[12] A. Kataria, and M. Singh, A review of data classification using k- nearest neighbour algorithm. International Journal of Emerging Technology and Advanced Engineering, 2013. 3(6): pp. 354-360.

[13] P. Cunningham, and S.J. Delany, k-Nearest neighbour classifiers. Multiple Classifier Systems, 2007. 34(8): pp. 1-17.

[14] M.Mujaand D.G.Lowe, Scalable nearest neighbor algorithms for high dimensional data. IEEE transactions on pattern analysis and machine intelligence, 2014. 36(11): pp. 2227-2240.

[15] L.Breiman, Random forests. Machine learning, 2001. 45(1): pp. 5-32.

[16] K.J. Archer, and R.V. Kimes, Empirical characterization of random forest variable importance measures. Computational Statistics & Data Analysis, 2008. 52(4): pp. 2249-2260.

[17] G. Kaur and A. Chhabra, Improved J48 classification algorithm for the prediction of diabetes. International Journal of Computer Applications, 2014. 98(22).

[18] A.R. Yılmaz, O. Yavuz, and B. Erkmen. Training multilayer perceptron using differential evolution algorithm for signature recognition application. in 2013 21st Signal Processing and Communications Applications Conference (SIU). 2013. IEEE.

[19] F.Asdaghi and A.Soleimani, An effective feature selection method for web spam

detection. Knowledge-Based Systems, 2019.
166: pp. 198-206.