International Journal of Pure and Applied Mathematics
Volume 117 No. 20 2017, 895-909
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

ijpam.eu

Analyzing behavior of spam tweetsin machine learning environment

Rahul Katarya
Department of Information Technology
Delhi Technological University, Delhi, India
rahulkatarya@dtu.ac.in

Abstract

There has been a rapid growth in the popularity of online social networking communities. However, online services have also attracted infiltration by anenormous number of spams. Spam is becoming a growingissue on social networking sites such as Facebook, Twitter, and Instagram. These spams pose great security threats to online users. In this paper, we have focused on the Twitter dataset, to study possible techniques to detect spam. An initial study of the Twitter dataset revealed that user-based and content-based features could be used to differentiate efficiently between spammers and legitimate users. These features form the basis of classifying a particular tweet as spam or non-spam. This research paper discusses the application of supervised machine learning algorithms to carry out spam detection. We have proposed a novel framework, which utilized well-known classifiers such as Naïve Bayes, KNN, and Meta-bagging and named as NKMFSD (Naïve Bayes-KNN-Meta-bagging for spam detection) framework for study and analysis of Twitter behavior. Toestimate the effciency of our proposed NKMFSDsystem, we calculated various parameters like probabilities of different categories of the attribute, distance to calculate then earest neighbor, false positive rate (FPR) and true positive rate (TPR). We estimated results from six different strategies[1], and compared results of those with our proposed system and our experimental studies reveal that our proposed systemdelivered the best result.

Keywords: Spam; Naïve Bayes; KNN; Meta-bagging; Twitter

1. Introduction

Online social networking[2–6] sites like Facebook, Twitter, Instagram, and My space allows millions of users across the world to unite with each other andcreate professional relationships. Online social networks are also embedded with recommender systems which have a high impact as they recommend useful items and offers to the users[7–12]. Users having a Twitter account can post their messages which are known as tweets[13–15]. These tweets may also include HTTP links and images. Although tweets allow the users to express their thoughts to a broadaudience,they are also becoming a medium of spreading unwanted and potentially malicious content over the social network. This highlights the need for an efficient spam detection model, which is capable of detecting spam as well as the account from where such spamoriginated before they reach vulnerable users sitting online.Any user can inform a spam tweet by clicking on "report as spam" link available on their respective homepage. Twitter investigates such reports and in case they are found spam, the reported accounts are suspended. Another mechanism for Twitter users to report a spam is to post a tweet in a "@spam@username" format. In this format, @username is the username of the account being reported which gets suspended if the spam is detected. However, the effectiveness of these methods is limited because spammers tend to exploit and abuse these methods. Some spammers deliberately report legitimate messages as spam while others report random tweets and usernames, which are not spamming accounts. Twitter puts efforts to flag such spammers, closing doubtful accounts, deleting fake and unverified IDs and filtering out malicious tweets. However, there have been complaints by legitimate users that Twitter suspended their accounts in its attempt to delete spamming accounts[16–19]. These tools used by Twitterare based upon ad-hoc methods, which

depend on a person's ability to identify spam manually. This reflects the direct need for better spam detection mechanisms, which are not entirely ad-hoc but are based on other factors as well. These spam detection mechanisms should be accurate and more efficient to avoid untimeliness to authentic Twitter users. In this paper, we have tried to overcome the above problems by building a Twitter spam detection system implemented through supervised machine learning algorithms. The study was carried out on the Twitterdataset, and our major contributions are as follow:

- The Twitter dataset consists of the tweets, whichare posted by the users, as well as information regarding the account from which these tweets were posted. This information includes parameters like a number of followers, verified status, and timestamps. We observed that these parameters could prove to be extremely useful for detecting the authenticity of an account. Such parameters can help us to categorize a particular tweet as spam or non-spam.
- We constructed a system in which we considered 1500 tweets by users to train our algorithms.
- Further, we implemented supervised machine learning algorithms such as Naïve Bayes, KNN,and Meta-bagging on the Twitter dataset to categorize a tweet as spam or non-spam.
- We calculated various parameters such as TPR and FPR to evaluate the presence of our proposed system.
- These experiment results revealed that the use ofsupervised machine learning classifier is highly efficient and gives excellent results with the large datasets like Twitter.
- We evaluated results with six different tactics and compared the results of those with our proposed system (NKMFSD), and our experimental studies reveal that our proposed system gave the best result.

The rest of this paper is organized as follows. Section 2 discusses related works for Twitter behavior and various recent performances of Twitter. Section 3 describes our proposed system (NKMFSD) in detail with flowchart and pseudocode. Section 4 discusses various parameters with experimental results withcomparisons, and Section 5 concludes with future.

2. Related Work

In this section, we will discuss the work, which has been carried out in this Twitter domain. The spam problem of Twitter has drawn the attention to a vast number of researchers worldwide.

Many scientists have conducted experimental studies for devising efficient mechanisms for detecting spam on social networking sites. An experiment was performed and evaluated three different aspects of data, feature, and model[1]. A data of 600 million public tweets were produced by using a commercial URL-based security tool. A probabilistic technique was presentedmutually to exploit the three kinds of associations for discovery experts[13]. Different types of online social communities and blogs are available where people can share their thoughts and feelings. A comparativeanalysisof online communities such as Twitter and Weibo was performed, and a software framework was proposed and also concluded that the communication among people inWeibo is substantively weaker than that of Twitter [20].In recent times, enormous amounts of movements that contain lots of spam or promotion accounts have appeared on Twitter. The movements supportively post useless information, and thus they can pollute more normal users than specific spam or advertising accounts[19]. Organizing or participating in movements has become the foremost procedure to spread spam or promotion information on Twitter. In thisarea of spam detection in Twitter, a study was conducted in which researchers composed and noticedan enormous number of spamtweets[17]. They additionally studied the misleading data in Twitter spam and found that several unreliable contents of spam.Opinion lexicons are used to support automatic sentiment examination of textual passages. A techniquewas offeredthat associated information from habitually annotated tweets and existing hand-made opinion lexicons to increased an opinion lexicon in a supervised fashion[21]. Anotherstudy was also performed to examine influencers on Twitter to determine the features of their tweets over PIAR, data mining research tool established by the University of Salamanca thatassociated graph theory and social influence theory[22]. A modelwas developed for classifying cyberbullying in Twitter.The developed model had a feature-based model that used

features from tweets, such as network, activity, user, and tweet content, and made a machine learning classifier for classifying the tweets as cyberbullying or non-cyberbullying[23]. A study was alsoaccomplished to determine features of spamspreading through Twitter and, explicitly, in which authors studied that how spammers use the certain features of the service to proliferation the efficiency of their campaigns[18].Another study was consideredwhich suggested that how user personality conditions the followee selection process by merging a quantitative examination of personality traits with topology and content[24]. The combined factors were implanted into a recommendation algorithm that calculated the similarity between target users and prospective followers and then graded those latent followees in reducing theorder of significance.

3. Proposed system

In this section, we will emphasize on theprocedure and framework that we have conceptualized and subsequently implemented to detect spam on Twitter. We have made use of supervised machine learning algorithms to carry out spam detection. The implementedmethod for the Twitter spam problemprimaryemphases on the discovery of tweets containing spam instead of detecting spam accounts. We compared the efficiency of our proposed system with the six traditional classifiers, namely Random Forest, Support Vector Machine, Bayes network, Naïve Bayesian, C4.5 and K-Nearest Neighbor classifiers.Our proposed system is comprised of the Naïve Bayes, KNN, and Meta-bagging.Our proposed system is classifiedinto four major steps such as data Collection, feature extraction, training of classifier and classification. We have used Twitter Streaming API methods provided by Twitter to extract available public data on theTwitterwebsiteto download Twitter tweets and storing these tweets inthefile. The size of this file was 1GB (approx.). In this dataset, each tweet downloaded contains tweet, URL, and many other features that Twitter offers when someone tweets.

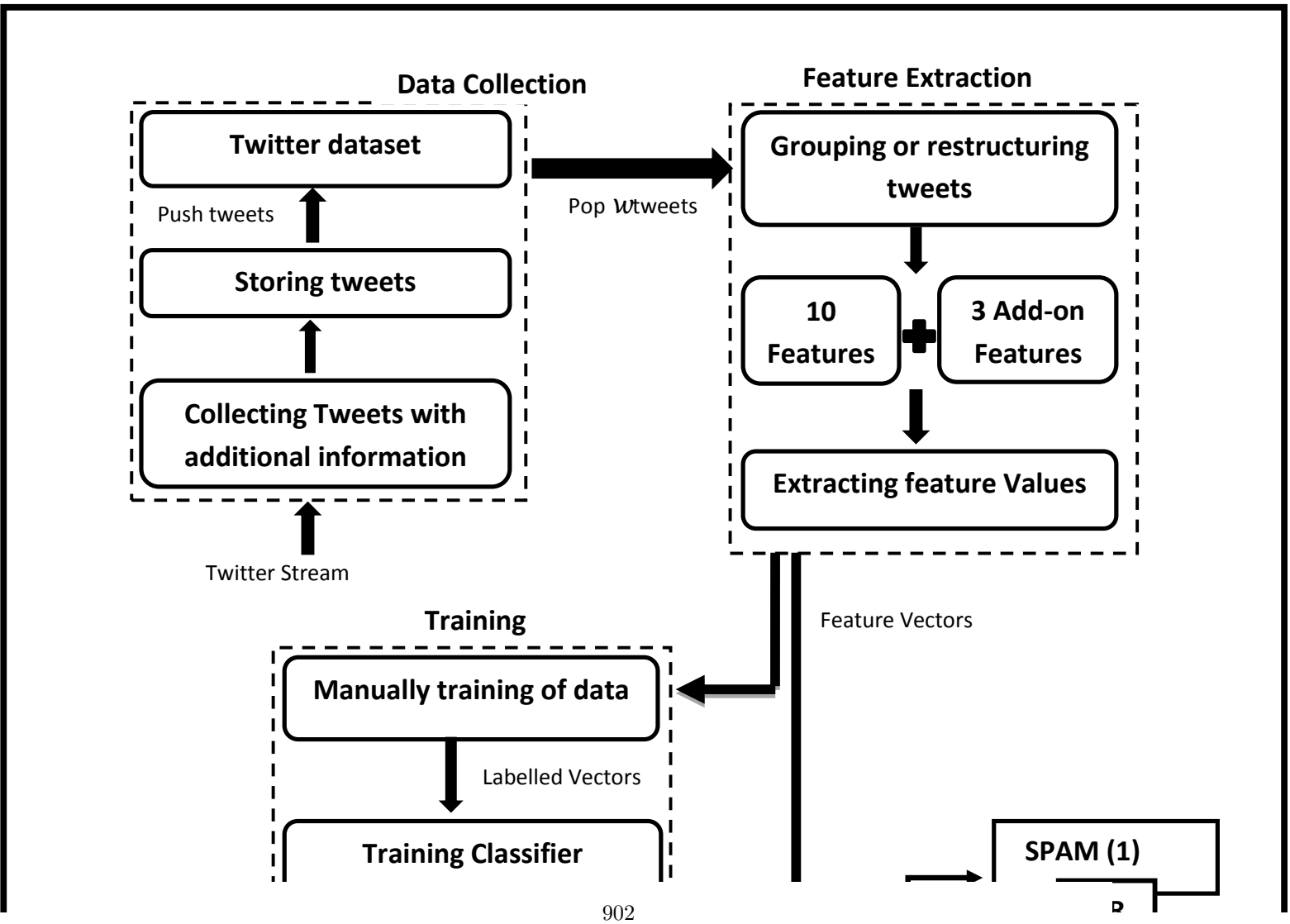
Table 1-Twitter features with description[1]

Feature Name	Description
account_age (time difference)	Age of an account since its formation until the time of sending the most recent tweet
no_follower	Number of followers of this twitter user
no_following	Number of followings/ friends of this twitter user
no_userfavourites	Number of favorites this Twitter user received
no_lists	Number of lists this Twitter user added
no_tweets	Number of tweets this Twitter user sent
no_retweets	Number of retweets this tweet
no_hashtag	Number of hashtags comprisedof this tweet
no_usermention	Number of user declarations included in this tweet
no_urls	Number of URLs included in this tweet
no_char	Number of characters in this tweet
no_digits	Number of digits in this tweet
account verified	Number of accounts verified
no_status	Number of statuses

Table 1 describes 13 lightweight features, which were extracted from our Twitter dataset and used for classification.Account age is a highly crucial factor which can tell us whether the account is reliable or not. Spamming accounts tend to have less account age. Numbers of followers can also help us identify a spamming account. Spammers tend to have very few or no followers. Similarly, spammers also rarely have any user favorites and lists. However, this parameter needs to be monitored carefully. Moreover, the "verified" status of the Twitter account also tells us whether that account is a spammer or not. If the account is not verified, it has a high probability of being a spamming account. For training oftheclassifier, we made a separate file of 1500 tweets from the dataset and manually classify them as spam or non-spam. With the use of this file and 13 features, which were extracted earlier, are used to train our classifier to detect which tweet is spam or non-spam.We employedthethree algorithms namely Naïve Bayes, K-NN, and meta-bagging. Supervised machine learning algorithms like the ones stated make use if a certain set carries out spam detection.These features are divided into different categories based on the magnitude of their values.The probability of tweets being spam on the occurrence of features in various categories is calculated which formed the basis for Naïve Bayes algorithm implementation for classifying tweets as spam or benign.For every tweet in the test dataset,

selected normalized features calculated k nearest neighbors from the training dataset. Each Tweet was subsequently classified in the class of which majority of the neighbors of that tweet lie. The two algorithms were combined using meta bagging algorithm by assigning weights to individual algorithms to produce a combined output for each tweet in the test dataset.This hasbeen illustrated in Figure 1.

The attributes are extracted from the dataset downloaded usingTwitter's online streaming API. A total of 13 features is extractedwhichare used for classification of tweets. These features are divided into different categories based on the magnitude of their values. The probability of tweets being spam on theoccurrence of features invarious categories is calculated which formed the basis for naïve Bayes algorithm implementation for classifying tweets as spam or benign.For every tweet in the test selected normalized features calculated dataset k, nearest neighbors from the training dataset. Each tweet was classified in the class of which majority ofthe neighbors of that tweet lie. The two algorithms were combined using Meta bagging algorithm byassigning weights to individual algorithms to produce a combinedoutput for each tweet in the test dataset.


Data Collection
Twitter dataset
Push tweets
Storing tweets
Collecting Tweets with additional information
Twitter Stream
Pop $w$tweets
Feature Extraction
Grouping or restructuring tweets
10 Features
3 Add-on Features
Extracting feature Values
Feature Vectors
Training
Manually training of data
Labelled Vectors
Training Classifier
SPAM (1)

```
1. Extract attributes of each tweet from the database.
2. Divide the dataset in training and test dataset.
3. Divide the range of each attribute in different categories.
4. Calculate the probability of spam and nonspam for each
category of each attribute based on the training dataset.
5.for each element in test dataset
   calculate ppos and pneg where
   ppos=probability of tweet being spam
   pneg=probability of tweet not being spam
6.if ppos>pneg
   classify tweet as spam
  else
   the tweet is benign
7.if ppos=pneg
8.apply k nearest neighbour algorithm to the test dataset
9.pscore=score in favour of tweet being spam
  nscore=score in favour of tweet being benign
10.fpscore=w1*(ppos*normalization_factor)+w2*(pscore);
11.fnscore=w1*(pnos*normalization_factor)+w2*(nscore);
12.if fpscore>fnscore
   classify tweet as spam
  else
   classify tweet as benign
13.Combine naive Bayes and knn by assigning weight to each algo
  and produce a resultant effect.
14. Calculate the measured TPR and FPR.
```

Figure 2. Pseudo code of our proposed system

4. Results

4.1 Dataset

The dataset has been downloaded from Twitter Streaming API.We have used Twitter Streaming API methods provided by Twitter to findavailable publically data from aTwitter website and stored these tweets in the file. The size of this file was 1GB (approx.). In this dataset, each tweet downloaded contains tweet, URL, and many other features that Twitter offers when someone tweets (Figure 3).

		Predicted	
		Spam	Non-Spam
True	Spam	TP	FN
	Non-Spam	FP	TN

Figure 3. Values of TP, FP, TN& FN[1].

4.2 Performance Metrics

For spam detection tactics, some metrics are widely used by the researchers. Positives and Negatives: A way to estimate the classifier's presentation is to use true negatives (TN), false positives(FP), true positives (TP), and false negatives (FN)[1, 25]. These metrics are defined as follows.

a) TP tweets of class S appropriately classified as belonging to class S.
b) FP tweets not fitting to class S wrongly classified as fitting to class S.
c) TN tweets not fitting to class S appropriately classified as not fitting to class S.
d) FN tweets of class S wrongly classified as not fitting to classes.

Table 2. Comparison of results with existing methodologies

S.No	Methods	Dataset	
		TPR	FPR
1	Random Forest	92.9	5.6
2.	C4.5	92.4	8.4
3.	Bayes Network	75.3	8.7
4.	Naive Bayes	97.3	77.1
5.	Knn	91.9	11.1
6.	SVM	79.1	18.9
7.	**Proposed system (NKMFSD)**	**95.84**	**99.54**

Table 2 shows the values of TPR and FPR of 6 other spam detection strategies[1] in which Naïve Bayes gives the best results regarding FPR and TPR. Each of these strategies is compared with our strategy, and results were calculated.Table two shows the results of our proposed system for

spam detection strategy, whichis implemented with Naïve Bayes,Knn, and meta-bagging machine learning algorithms. Our result increased 2% regarding TPR and 17% regarding FPR. The value of FPR=99.54% and TPR=95.84%.The relations of TP, FP, TN, and FN in social spam detection are shown in Table 2.

$$\text{TPR} = \frac{\text{TP}}{\text{TP+FN}} \quad (1)$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP+FN}} \quad (2)$$

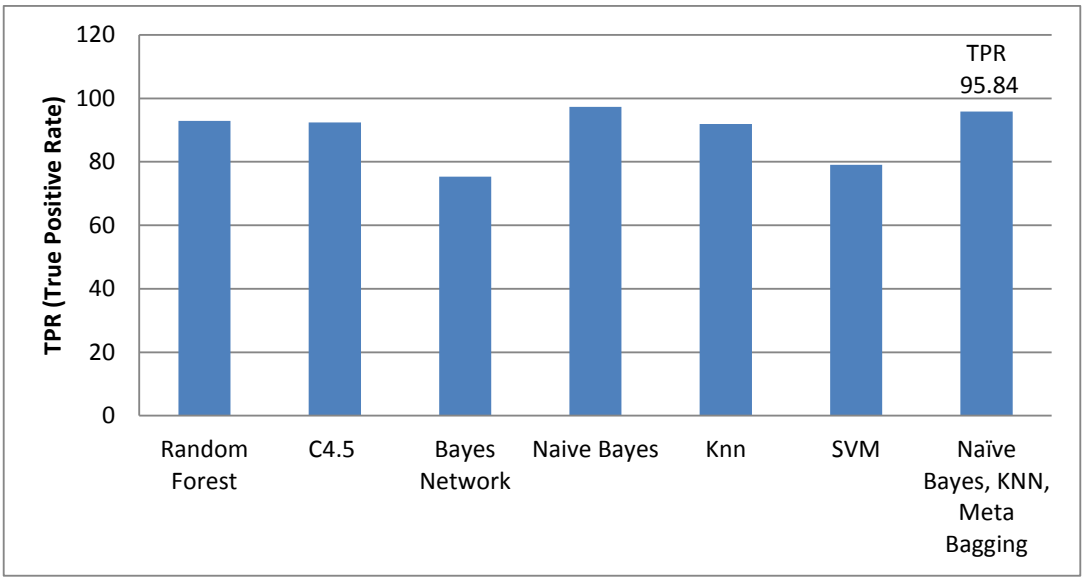


Figure 4. True Positive Rate of each strategy

Figure 4 shows the value of TPR for Table 2 and 2% increase in the value of TPR with other approaches except for Naïve Bayes. Figure 5demonstrates the value of FPR for Table 2 and 17% increase in the value of FPR with other methodologies.

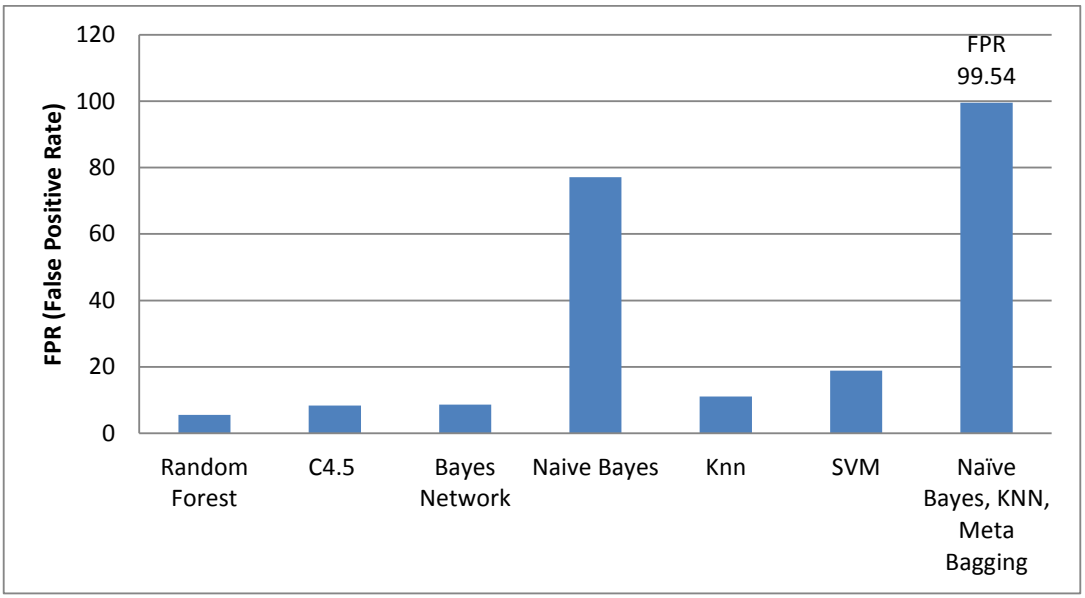


Figure 5. False positive rate of each strategy

5. Conclusion and future work

We developed a classifier to evaluate the detection of spam or non-spam tweets based onsupervised machine learning algorithms such as Naïve Bayes, KNN and Meta baggingon theTwitter dataset to detect spam or non-spam tweets.The results demonstrate that our spam detection system has a 99.54% FPR and 95.84% TPR using Naïve Bayes, Knn, and meta-bagging machine learning algorithms.Overall, there has been a 2% improvement in the true positive rate and a 17% improvement in the false positive rate when we compare our approach with six existing methods.Our results showthat our classifier gives the best performance by comparing FPR and TPR as parameters with other six strategies. The proposed system achieved 99.54% FPR and 95.84% TPR. We present a newsystem to detect spam tweets onTwitter by using the three machine learning algorithm which isNaïve Bayes, Knn, and meta-bagging. We also presentthereal-time detection of tweets as spam or non-spam. Apart from this, we can also consider using other classifier algorithms such as Ada-boost and A priori, which could give us

even better results and analyze the changing efficiency with changing thevalue of a number of clusters in knn.

References

1. Chen C, Zhang J, Xie Y, et al. (2016) A Performance Evaluation of Machine Learning-Based Streaming Spam Tweets Detection. IEEE Trans Comput Soc Syst 2:1–12. doi: 10.1109/TCSS.2016.2516039
2. Li J, Liu C, Yu JX, et al. (2016) Personalized Influential Topic Search via Social Network Summarization. IEEE Trans Knowl Data Eng 28:1820–1834. doi: 10.1109/TKDE.2016.2542804
3. Jankowski-Lorek M, Jaroszewicz S, Ostrowski Ł, Wierzbicki A (2016) Verifying social network models of Wikipedia knowledge community. Inf Sci (Ny). doi: 10.1016/j.ins.2015.12.015
4. Huang S, Zhang J, Wang L, Hua X-S (2016) Social Friend Recommendation Based on Multiple Network Correlation. Multimedia, IEEE Trans 18:287–299. doi: 10.1109/TMM.2015.2510333
5. Nettleton DF, Salas J (2016) A data driven anonymization system for information rich online social network graphs. Expert Syst Appl 55:87–105. doi: 10.1016/j.eswa.2016.02.004
6. Katarya R, Verma OP (2017) Effectual recommendations using artificial algae algorithm and fuzzy c-mean. Swarm Evol Comput. doi: 10.1016/j.swevo.2017.04.004
7. Katarya R, Verma OP (2016) Recent developments in affective recommender systems. Phys A Stat Mech its Appl 461:182–90. doi: doi:10.1016/j.physa.2016.05.046
8. Katarya R, Verma OP (2016) A collaborative recommender system enhanced with particle swarm optimization technique. Multimedia Tools Appl 75:1–15. doi: 10.1007/s11042-016-3481-4
9. Katarya R, Verma OP (2015) Restaurant Recommender System Based on Psychographic and Demographic Factors in Mobile Environment. In: IEEE Int. Conf. Green Comput. Internet Things 2015. pp 907–912
10. Katarya R, Jain I, Hasija H (2014) An Interactive Interface for Instilling Trust and providing Diverse Recommendations. In: IEEE Int. Conf. Comput. Commun. Technol. ICCCT-2014. pp 17–22
11. Deshmukh JS, Kumar A (2017) Applied Computing and Informatics Entropy based classifier for cross-domain opinion mining. Appl Comput Informatics. doi: 10.1016/j.aci.2017.03.001
12. Cao H, Lin M (2017) Mining smartphone data for app usage prediction and recommendations : A survey. Pervasive Mob Comput 37:1–22. doi: 10.1016/j.pmcj.2017.01.007
13. Wei W, Cong G, Miao C, et al. (2016) Learning to Find Topic Experts in Twitter via Different Relations. IEEE Trans Knowl Data Eng 28:1764–1778. doi: 10.1109/TKDE.2016.2539166
14. Tan S, Li Y, Sun H, et al. (2014) Interpreting the Public Sentiment Variations on Twitter. IEEE Trans Knowl Data Eng 26:1158–1170.
15. Kim Y, Shim K (2014) TWILITE : A recommendation system for Twitter using a probabilistic model based on latent Dirichlet allocation. Inf Syst 42:59–77. doi: 10.1016/j.is.2013.11.003
16. Chen C, Zhang J, Xiang Y, et al. (2016) Spammers Are Becoming "smarter" on Twitter. IT Prof 18:66–70. doi: 10.1109/MITP.2016.36
17. Chen C, Wen S, Zhang J, et al. (2016) Investigating the deceptive information in Twitter spam. Futur Gener Comput Syst -. doi: http://dx.doi.org/10.1016/j.future.2016.05.036
18. Antonakaki D, Polakis I, Athanasopoulos E, et al. (2016) Exploiting abused trending topics to identify spam campaigns in Twitter. Soc Netw Anal Min 6:48. doi: 10.1007/s13278-016-0354-9
19. Zhang X, Li Z, Zhu S, Liang W (2016) Detecting Spam and Promoting Campaigns in Twitter. ACM Trans Web 10:1–28. doi: 10.1145/2846102
20. Han W, Zhu X, Zhu Z, et al. (2016) A Comparative Analysis on Weibo and Twitter. Tsinghua Sci Technol 21:1–16.
21. Bravo-Marquez F, Frank E, Pfahringer B (2016) Building a Twitter Opinion Lexicon from Automatically-annotated Tweets. Knowledge-Based Syst 108:65–78. doi: 10.1016/j.knosys.2016.05.018
22. Lahuerta-Otero E, Cordero-Gutiérrez R (2016) Looking for the perfect tweet. The use of data mining techniques to find influencers on twitter. Comput Human Behav 64:575–583. doi: 10.1016/j.chb.2016.07.035
23. Al-garadi MA, Varathan KD, Ravana SD (2016) Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. Comput Human Behav 63:433–443. doi: 10.1016/j.chb.2016.05.051

24. Tommasel A, Corbellini A, Godoy D, Schiaffino S (2016) Personality-aware followee recommendation algorithms: An empirical analysis. Eng Appl Artif Intell 51:24–36. doi: 10.1016/j.engappai.2016.01.016
25. Sobitha Ahila S, Shunmuganathan KL (2016) Role of Agent Technology in Web Usage Mining: Homomorphic Encryption Based Recommendation for E-commerce Applications. Wirel Pers Commun 87:499–512. doi: 10.1007/s11277-015-3082-y