Spam Detection using KNN and Decision Tree Mechanism in Social Network

[1]Saumya Goyal, [2]Prof. R. K. Chauhan , [3]Shabnam Parveen

[1]Assistant Professor, Department of Computer Science & Engineering,U.I.E.T, Kurukshetra University, Kurukshetra, Haryana, INDIA

[1]goyalsaumya4@gmail.com

[2]Professor, Department of Computer Science & Application, Kurukshetra University,Kurukshetra, Haryana, INDIA

[2]rkchauhandcsakuk@gmail.com

[3] Phd Scholar, Department of Computer Science & Application, Kurukshetra University,Kurukshetra, Haryana, INDIA

[3]er.shabnam786@gmail.com

***Abstract*-- Social network (SN) is an online platform extensively used as communication tool by millions of society in order to built social relation with others for career purposes, knowledge point of view, politics and many more. In today's time everyone is online which make it today's most vast network of information. Different SN applications are available like Twitter, Facebook and MySpace through which peoples can communicate with other and send text, audio and video messages. During communication it is possible that a user can performs unwanted activities and send spam messages to disturb communication process. It is difficult to detect these kinds of spam messages. In this paper spam detection mechanism based on decision tree and KNN algorithm has been proposed. In proposed mechanism we apply these algorithms on real datasets of twitter to detect spam messages. To analyse proposed mechanism Weka tool is used. The performance metrics like TP Rate, FP Rate, Precision, Recall, F-Measure and Class are used to measure the execution of proposed mechanism.**

***Keywords*—Social Network, Spam, KNN, decision tree and Weka.**

I INTRODUCTION

Web has turned into an imperative instrument for correspondence, as a result of its quick speed and ease. Regularly, web indexes are the beginning stage for skimming on Internet. So the consequences of raking for a given question are exceedingly critical for business sites. Web spamming debases the nature of the outcomes recovered by the web index. [1] Because of the created innovation, an online interpersonal organization transforms into critical system to trade and impart data to each other. The quantity of online informal organizations have been dug in and utilized by a few distinct clients. Twitter is a standout amongst the most prevalent online networking as it licenses client to mail and read endless presents related on content. Today thousands and a large number of clients [2] share their data with others and there are roughly 400 million tweets each day [3]. It is typical trademark in informal organization that few discrete characters [4] have especially same impact on each other. Tremendous number of accessible clients and amount of traded information on Twitter interpersonal organizations heads to parcel of cyber crime exercises by spammers whose reason for existing is to expand spam messages through URLs of related sites. [6] The inspiration driving spam is to convey data to the beneficiary that contains a draft like publicizing for a (presumable useless, non-existent, or illicit) item, advancement of cause, lure for an extortion plan, or PC malware intended to seize the beneficiary's PC. As it is so modest to send data, just a little portion of focused beneficiaries — perhaps one in ten thousand or less need to react and get to the draft for spam to be productive to its senders [5].

The fame of Twitter is potentially because of its confinement. Client presents are restricted on characters length of 140, and the security model is profoundly constrained: an entire record is either private (just imparting presents on companions) or open, and most clients select "open" sharing the greater part of their substance unreservedly to world. Besides taking after a client is not as a matter of course equal: since all tweets are open, after a client just subscribes an adherent to their open tweets and along these lines clients are urged to take after people, they don't know by and by. This prompted numerous big names utilizing Twitter as medium of associating with their followers, since they can redesign their large number of followers with a solitary 140 character tweet.

Twitter has turned into a prevalent perspective in person to person communication spams [7] because of its weakness and helplessness to assaults. This Twitter's spam abuses the clients with pulling in notifications, for example, "Just saw this photograph of you‟ which is trailed by connection that, takes client to an unapproved site that transfers malware onto the client's PC [8]. In specific situations, by taking welfare of the spear-phishing systems [9]-[10]; the messages may appear to originate among one of the typical companions. Assailants or interlopers utilizes Twitter to intercede coded redesign messages to clients officially contaminated by rebellious code to handle botnets [11] which are gatherings of destroyed PCs that can be coordinated to tell different clients who send false messages or causes an assault over sites with dirtied movement.

978-1-5090-3669-1/16/$31.00 ©2016 IEEE

In Twitter long range informal communication spam [6], spammers make fake records with a specific end goal to take the private information or to circulate business promotions in interpersonal organization for individual's advantage which influence the general security and execution of person to person communication. The principle test is to recognize Twitter social spamming accounts made by double-dealers as their conduct would have numerous assortments with much bigger component space which make it hard to distinguish. The vital issue in identifying tweet social cheaters is their dynamic nature, which makes it troublesome. In customary framework, the execution is consistent by appealing direct routine frameworks, as the fraud dealers get grow new, more slippery strategies to keep away from their detection. [13] Spam recognition in informal organizations is moderately late territory of exploration. A large portion of the looks into around there take after the same general strategy for location: 1) operate experimental study to choose some basic or printed components to inspect; 2) operate characterization and machine learning strategies with corresponding elements to discover designs crosswise over clients and messages; 3) assess whether models in view of examples are powerful in identifying undesirable conduct. Numerous analysts have presented much complex structure or half and half formation to evacuate this issue to accomplish the precision in spam location. [1][12] has presented a cross breed model and obtain numerous of elements to distinguish the action of spammers. Both grouping and order methods are utilized to identify junk and to upgrade execution which makes it more unpredictable and tedious framework, alongside this have not accomplished that level of execution.

Keeping in mind the end goal to tackle these issues, this paper has proposed a rearranged spam identification framework. At first, information is gathered and content and substance components are extricated by pre-preparing. At that point applying classifiers an execution network has advanced and reason better execution of distinguishing spam and non-spam tweets.

II RELATED WORK

In [14] authors explored different features of the spammed videos. For this first the videos from you-tube was collected as the dataset and then labelled them manually into groups of spam and useful videos. On the basis of number of features extracted from database they detected spam videos. Microsoft SQL Server Data mining Tool (SSDT) was applied to categorize random videos as legitimate or spam.

In [15] authors proposed several novel features which were capable of distinguishing genuine accounts from the spam accounts. The features analysed the content entropy and behavioural, profile vectors, bait techniques for recognizing spammers, which were then insert into supervised learning algorithms to generate models for their tool, CATS.

In [16] for detecting trending events from twitter which was discovered by (LSH) locality sensitive hashing technique, authors proposed a novel method. In this paper they have included following challenges: (1) applying LSH to search truly interesting events (2) discover the behaviour of events based on cluster size, geo-location and time (3) building a dictionary using (TF-IDF) in high dimensional data to form tweet feature vector (4) increases-up cluster discovery process during conserving the quality of cluster.

In [17] authors have studied the problems of spam in Chinas famous micro-blog, weibo by creating dataset of 375,430 posts and examined 2,370 users from weibo. Their main goal was the deep analysis of spammed posts and spammers. They contributed their work as follows: one, they worked on undemonstrated assumptions of spammers as botnet users, because it was found that spammers act just like regular users. Two, they investigated the burst properties of useful posts and spammed posts.

In [18] authors have focused on systematically analysing and labelling models that detect review spams. They found that studies could be labelled into three groups which focused on methods to detect individual spammers, spam reviews and group spams.

In [19] authors have introduced multi-scale entropy for identifying and analysing user behaviour on twitter and separately categorised tweeting activities on twitter: promotion and advising activities, individual activities, robotic/automatic activities, newsworthy information dispersion activities and other activities. Though experiments they have achieved great separation above five categories of activities based on Multi-scale Entropy of enjoyers posting time sequence.

III PROPOSED WORK

In this section we have presented our proposed mechanism.

```
Algorithm I
    1. Start
    2. Fed real dataset
    3. Extract feature using different feature
       extraction methods.
    4. Apply algorithm 2 for decision tree
       classifier and algorithm 3 for KNN
       classifier.
    5. Distinguish Normal and Spam messages.
    6. END
```

```
    Algorithm II
 1. Input real dataset of Twitter and
    labelled it manually.
 2. Extract features by applying different
    feature abstraction algorithms.
 3. Now apply Decision Tree and K-
    nearest neighbour classifiers to
    classify normal tweets and spam
    tweets.
 4. Then check their performance with the
    help of performance matrix like
    Accuracy, F-measures, TPR and FPR.
```

A. K-Nearest Neighbour classifier

The K-Nearest Neighbour classifier is appraised as an example-based classifier which courses that the teaching records are used for collation rather than an unreserved category representation such as the category of profiles accessed by other classifiers. There is no actual teaching phase. When new document needed to be designated, the k most similar records (neighbours) are detected and if an enough large section of them have been allotted to a certain designation then new records is also allotted to this designation, otherwise not. Additionally, detecting the nearest neighbours can be fastened by exploring traditional indexing methods. To check whether a message is spam or legitimate data, we look at the classes of the notification that are nearest to it. The comparison among the vectors is real time process. This is the overview of k nearest neighbour algorithm.

Above explain algorithm is represented in the form of flowchart in the following diagram.

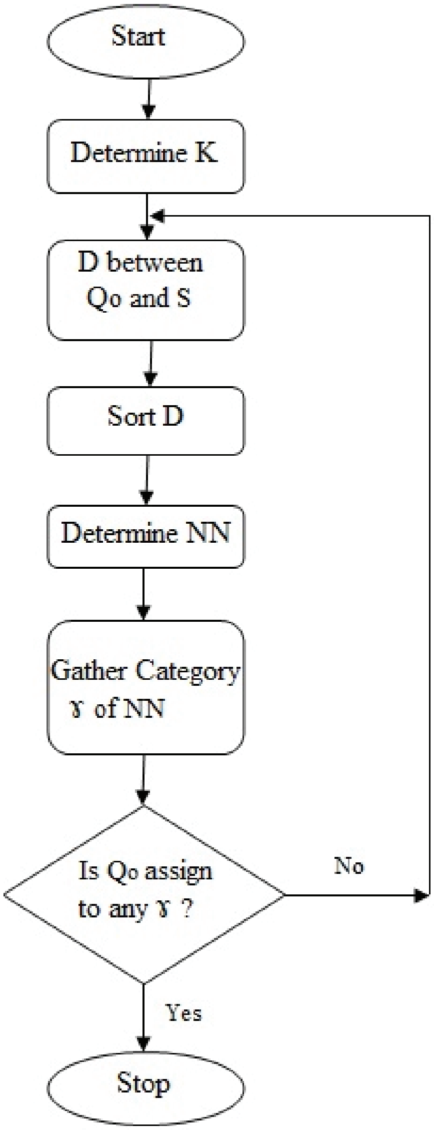


Fig.1.KNN Mechanism

Explanation of above flowchart is as follows:

1. K is the parameter in which number of nearest neighbours is determined.
2. D denotes the distance between target instance (Q_0) and training instance (S).
3. NN is the nearest neighbour on the basis of kth minimum distance.
4. ɤ is the category of the classes.
5. If Q_0 is not assigned to any class then again distance is calculated and if it is assigned then it will do whole process for next target instance.

B. Decision Tree Classifier

Decision tree classifier is one of the method repeatedly used in data mining. Based on several input variables it creates the decision tree model and train it. In this model every leaf presents value of target variables by representing the path from root to the leaf.

Based on predefined attribute value tree can be “trained” by dividing the source set into numerous subsets. This process is iterated on every acquired subset in recursive manner which is wellknown as recursive partitioning. The recursion is acquired when the subset at node has similar value as of the target variable.

The input of algorithm consists of the teaching records S and attribute sets T. The algorithm works on recursively selecting the best attribute to divide the data (step j) and expanding the leaf nodes of the tree (steps n and o) until stopping criteria meet (step c).

```
BuiltTree (S,T)
    a)  START.
    b)  Initialise real data set and extract features from it.
    c)  if Halt_cond(S,T) = true
    d)   then,
    e)  leaf = Generate_node().
    f)  leaf_Mark = classify(S).
    g)  return leaf.
    h)  else
    i)  root = Generate_node().
    j)  root.test.cond = find.best.split(S,T).
    k)    Let V = {v|v is a possible outcome of
        root.test.cond}
    l)    for each v ε V do
    m)  {$S_v$= {e|root.test.cond(e) =v and e ε S}
    n)       child = BuiltTree($S_v$,T)
    o)       add child as descendent of root and mark the
        edge (root→child) as v}
    p)  end for
    q)  end if
    r)  return root.
    s)  END.
```

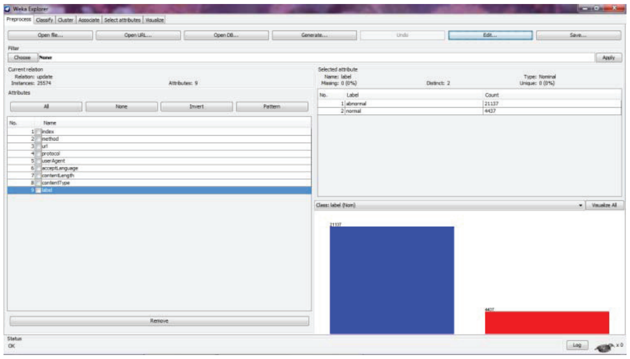
Description of the above algorithm is as follows:

1. Generate_node() function expands the length of decision tree by generating new node. A node in decision tree is either class label represented as node.label or a test condition represented as root.test.cond().
2. The find.best.split() function defines the attributes which should be selected as a splitting state for the teaching records. The choice of test state is depend on entropy, Gini index X^2 statistics to examine the goodness of split.
3. The classify() function determines that which leaf node is assigned to which class label. In most of the cases, the class which has maximum number of teaching records is assigned by the lead node.
4. The Halt_cond() function is to stop the process of growing tree by determining whether all records lie in same attribute values or same class label and another way is to check whether number of records have less value then threshold value.

III RESULTS AND ANALYSIS

A. Tool Used: We used WEKA to analyse real dataset using KNN and Decision tree algorithm. WEKA is open source software for data analysis or data mining[20].

B. Data set used



Filter: None

Current relation
Relation: update
Instances: 25574
Attributes: 9

Attributes

No.	Name
1	index
2	method
3	url
4	protocol
5	userAgent
6	acceptLanguage
7	contentLength
8	contentType
9	label

Selected attribute
Name: label
Type: Nominal
Missing: 0 (0%)
Distinct: 2
Unique: 0 (0%)

No.	Label	Count
1	abnormal	21137
2	normal	4437

Class: label (Nom)

Fig.2. elements used in dataset

Fig shows number of parameters used in dataset like web links of spam, label of spam like normal or abnormal, and content length and so on. Here in this fig Blue line depicts spams and red line shows normal messages or links.

Performance metrics

TP Rate: True Positive Rate is presented as the amount of the actual spam users who are correctly categorized as spam users.

TPR is formulated as:

$$\text{TPR} = \frac{Tp}{Tp+Fn}$$

FP Rate: False Positive Rate is presented as the amount of probability of wrongly classified results for spam and normal detection.

FPR is formulated as:

$$\text{FPR} = \frac{Fp}{Fp+Tn}$$

F-measure: F-measure uses the result of both recall R and precision P

F-measure is formulated as:

$$\text{F-measure} = 2 * \frac{PR}{P+R}$$

Precision and Recall: Precision is the fraction of collection of instances which are relevant where as Recall is fraction of relevant instances that are collected.

Table 1: depicts KNN values with different performance metrics

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
1	0.508	0.904	1	0.949	Spam
0.492	0	1	0.492	0.659	Normal
0.912	0.42	0.92	0.912	0.899	Weighted Avg.

Table 2: Confusion Matrix of KNN

Spam	Normal
21137	0
2255	2182

Table 3: depicts Decision Tree values with different performance metrics

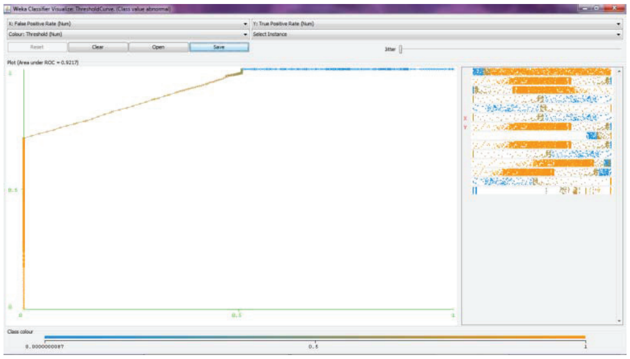
TP Rate	FP Rate	Precision	Recall	F-Measure	Class
1	1	0.827	1	0.905	Spam
0	0	0	0	0	Normal
0.827	0.827	0.683	0.827	0.748	Weighted Avg.

Table 4: Confusion Matrix of Decision Tree

Spam	Normal
21137	0
4437	0

Table 1 and Table 3 values of performance matrix like TP Rate, FP Rate, Precision, Recall, F-measure. These values are computed by Weka Tool by applying KNN and Decision Tree algorithms. Table 1 depicts KNN reading while Table 2 depicts Decision Tree readings.

Table 3 and Table 4 shows confusion matrix of KNN and Decision Tree. A Confusion matrix contains information about classifiers used for the performance analysis. The information contained is both actual information and predicted information by classifier.



Weka Classifier Visualize: ThresholdCurve. (Class value abnormal)

X: False Positive Rate (Num)

Y: True Positive Rate (Num)

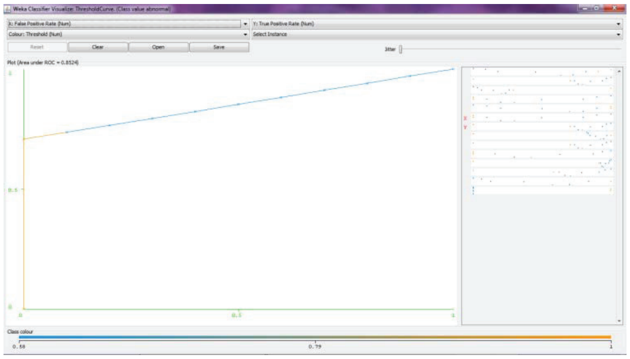
Colour: Threshold (Num)

Select Instance

Plot (Area under ROC = 0.9217)

Fig.3. ROC Analysis of KNN classifier describing spam detection

Fig 3 depicts ROC curve of spam detected by KNN algorithm.



Weka Classifier Visualize: ThresholdCurve. (Class value abnormal)

X: False Positive Rate (Num)

Y: True Positive Rate (Num)

Colour: Threshold (Num)

Select Instance

Plot (Area under ROC = 0.8524)

Fig:4 ROC Analysis of Decision tree classifier describing spam detection

Fig 4 depicts ROC curve of spam detected by decision tree algorithm. In decision tree algorithm detection rate is low as compare to KNN algorithm.

IV CONCLUSION AND FUTURE WORK

Social network is the widest network for exchanging information all over the world. Along with the benefits of the useful information available on social networks spammed messages are also wide spreading with high speed. In this paper Decision tree and KNN classifier algorithms are applied to detect those spam and normal messages and links. Weka tool is used to analyse these algorithms on real dataset. Results show that KNN algorithm is better than Decision tree algorithm. In future we will continue working on it and proposed novel mechanism to contrast spam's with high detection rate and compare with above mentioned algorithms.

REFERENCES

[1] S. J. Soman and S. Murugappan, "Detecting malicious tweets in trending topics using clustering and classification," *2014 International Conference on Recent Trends in Information Technology*, Chennai, 2014, pp. 1-6. doi: 10.1109/ICRTIT.2014.6996188

[2] S. Fiegerman, "Twitter now has more than 200 Million monthly active users," Mashable. [Online]. Available: http://mashable.com/2012/12/18/twitter-200-million-active-users/. Accessed: July. 22, 2016

[3]]H. Tsukayama, "Twitter turns 7: Users send over 400 million tweets per day," in *Washington Post*, Washington Post, 2013. [Online]. Available: http://articles.washingtonpost.com/2013-03-21/business/37889387_1_tweets-jack-dorsey-twitter. Accessed: July. 24, 2016.

[4] Il-ChulMoon,Dongwoo Kim, Yohan Jo and Alice O, "Analysis of twitter lists as a potential source for discovering latent characteristics of users," in CHI 2010 Workshop on Microblogging: What and How Can We Learn From It?, 2010.

[5] Neetu Sharma, GaganpreetKaur,et all, "Survey on Text Classification (Spam) Using Machine Learning",(IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (4) , 2014, 5098-5102

[6] Gordon V. Cormack, *David R. Cheriton, "*Email Spam Filtering: A Systematic Review*"*, Foundations and Trends ®in Information Retrieval Vol. 1, No. 4 (2006) 335–455©2008.

[7] C. Yang, R. C. Harkreader, and G. Gu, "Die free or live hard? Empirical evaluation and new design for fighting evolving twitter spammers", in 14th International Symposium, (RAID 2011), CA, USA, Proceedings in LNCS Series, Springer, Vol. 6961, pp. 318–337, 2011.

[8] N. Villeneuve, "Koobface: Inside a crimeware network", Munk School of Global Affairs, (JR04-2010), 2010.

[9] K. J. Nishanth, V. Ravi, N. Ankaiah, and I. Bose, "Soft computing based imputation and hybrid data and text mining: The case of predicting the severity of phishing alerts", in Expert Systems with Applications, Vol.39, Issue 12, pp.10583–10589, 2012.

[10] V. Ramanathan, and H. Wechsler, "phishGILLNET–phishing detection using probabilistic latent semantic analysis", in EURASIP Journal on Information Security, 2012.

[11] J. Nazario, "Twitter-based botnet command channel", [Online].Available:http://asert.arbornetworks.com/2009/08/twitter-based-botnetcommandchannel.Accessed: July,22,2016

[12] S. J. Soman and S. Murugappan, "Bayesian probabilistic tensor factorization for malicious tweets in trending topics," *2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT)*, Kanyakumari, 2014, pp. 895-900

[13] G. Stafford and L. L. Yu, "An Evaluation of the Effect of Spam on Twitter Trending Topics," *2013 International Conference on Social Computing*, Alexandria, VA, 2013, pp. 373-378. doi: 10.1109/SocialCom.2013.58

[14] R. Chowdury, M. N. Monsur Adnan, G. A. N. Mahmud and R. M. Rahman, "A data mining based spam detection system for YouTube," *Digital Information Management (ICDIM), 2013 Eighth International Conference on*, Islamabad, 2013, pp. 373-378.

[15] A. A. Amleshwaram, N. Reddy, S. Yadav, G. Gu and C. Yang, "CATS: Characterizing automation of Twitter spammers," *2013 Fifth International Conference on Communication Systems and Networks (COMSNETS)*, Bangalore, 2013, pp. 1-10.

[16] K. B. Shakira, A. Abdolreza, "Cluster-discovery of Twitter messages for event detection and trending", Journal of Computational Science, Volume 6, January 2015, Pages 47-57, ISSN 1877-7503, http://dx.doi.org/10.1016/j.jocs.2014.11.004.

[17] K. Chen, L. Chen, P. Zhu and Y. Xiong, "Unveil the Spams in Weibo," *Green Computing and Communications (GreenCom), 2013 IEEE and Internet of Things (iThings/CPSCom), IEEE International Conference on and IEEE Cyber, Physical and Social Computing*, Beijing, 2013, pp. 916-922.

[18] S. He, H. Wang and Z. H. Jiang, "Identifying user behavior on Twitter based on multi-scale entropy," *Security, Pattern Analysis, and Cybernetics (SPAC), 2014 International Conference on*, Wuhan, 2014, pp. 381-384.

[19] H. Atefeh, A. T. Mohammad, S. Naomie and H. Zahra, "Detection of review spam: A survey", 2015 Elsevier, pp. 3634-3642.

[20] http://www.cs.waikato.ac.nz/ml/weka/, accessed on 23-08-2016.