Advances in Computer Science and Information Technology (ACSIT)
p-ISSN: 2393-9907; e-ISSN: 2393-9915; Volume 3, Issue 4; April-June, 2016, pp. 301-304

http://www.krishisanskriti.org/Publication.html

Improved K-Means for Detection of Human users in Twitter

Isha[1], Sanjeev Dhawan[2] and Kulvinder Singh[3]

[1]M.Tech. (Computer Engineering), Department of Computer Science & Engineering, University Institute of Engineering and Technology (U.I.E.T), Kurukshetra University, Kurukshetra, Haryana
[2]Faculty of Computer Science and Engineering , Department of Computer Science & Engineering, University Institute of Engineering and Technology (U.I.E.T), Kurukshetra University, Kurukshetra, Haryana
[3]Faculty of Computer Science and Engineering , Department of Computer Science & Engineering, University Institute of Engineering and Technology (U.I.E.T), Kurukshetra University, Kurukshetra, Haryana
E-mail: [1]ishasingla587@gmail.com, [2]rsdhawan@rediffmail.com, [3]kshanda@rediffmail.com

Abstract—*Internet has become a significant mean today. The use of internet has increased widely. Internet reduces the manual work and also more time consuming. This is the reason why people today are connected to internet. In recent years Online Social Networks (OSNs) also developed well and plays an equal role. In the digital world of applications, a new application called twitter made a major impact in online social networking and micro blogging. The communication between users is through text based post. Its popularity also attracts many spammers to infiltrate legitimate users account with large amount of spam messages .Online social networking platforms are providing us with a large scale platform to study the human behavior. This paper improves K Means algorithm to separate human from not human users in order to identify normal human activity. K-Means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. Here in this research paper in order to achieve the target results with better accuracy, an efficient approach will be designed by modifying sequential K-Means clustering algorithm to detect spam in Twitter. The data which has been provided for the entire process to be performed is extracted from the social networking website Twitter with the help of R Package as it provides interface with the Twitter web API (Application Programmable Interface). Various calculations has been performed to calculate the accuracy, precision, and recall and based on these results and respective graphs have been obtained.*

Keywords*: Humans, Clustering, K-Means, R package, Social Networks, Spam, Twitter*

1. INTRODUCTION

Twitter is the red hot tool for micro blogging and social networking these days. Started in the late March of 2006 and twitter's off-the-wall the features makes twitter stand tall in this cyber world. As it is era of blogging, micro blogging and people connecting through social sites hence one cannot overlook online blogging and social networking site named Twitter which differs from traditional blogging and has vital add inns. It is a web application which gives users features like Direct Messaging, Following People & Trending Topics, Links, Photos, Videos message, image, or video links to share with their peers/colleagues and with followers such as personal online diaries or news on particular subject also one important aspect to notice is the small message refers to only 140 characters. These short messages are called tweets. With larger user databases in OSNs, twitter is becoming a more interesting target for spammers/malicious users. Spam can take different forms on social web sites and it is not easy to be detected. Spam is defined as the use of electronic messaging system to send unsolicited bulk messages. With the rise of OSNs, it has become a platform for spreading spam. Spammers intend to post advertisements of products to unrelated users. As per twitter policy (http://help.twitter.com) indicators of spam profiles are the metrics such as following a large number of users in a short period of time or if post consists mainly of links or if popular hashtags (#) are used when posting unrelated information or repeatedly posting other user's tweets as your own. There is a provision for users to report spam profiles to Twitter by posting a tweet to @spam. But in Twitter policy there is no clear identification of whether there are automated processes that look for these conditions or whether the administrators rely on user reporting, although it is believed that a combination approach is used. Some spammers post URLs as phishing websites which are used to steal user's sensitive data. To cope up with this problem various techniques have been used which are called as spam detection techniques. Some are supervised and others are unsupervised, the only difference between them is that supervised techniques use the trained data whereas unsupervised techniques use the untrained data. The technique which is discussed in this paper is K-Means technique which is unsupervised one means it uses the untrained data. K-Means clustering algorithm to group the messages based on the similarity of their attributes or features into K disjoint groups using Euclidian distance, to improve the accuracy of spam detection. The data which has been provided as input to this

process is extracted from Twitter with the help of R package. R package is a tool which is very much used for the statistical computing and graphics [1]. On the data which is extracted by R package further steps are performed including Preprocessing, Feature Extraction, and then evaluation. Preprocessing is further performed in two steps i.e., stop word removal and stemming. At the end of this paper the results along the conclusion and future scope are given and a system which is time efficient and accurate is introduced.

2. LITERATURE REVIEW

Identification of anomalous user types in Twitter data is an important precursor to detailed analysis of Twitter behaviours as they could incorrectly skew the results obtained in terms of topics prevalent in the population. Identification of specific types of users as different from the rest of the population is, in essence, a form of creating a profile of the user's interaction with the platform. Existing techniques in spammer detection typically use a pre-classified data set and a combination of behavioural (content, user information, network and topic) to create a classifier that can accurately differentiate spammers from legitimate users with accuracies obtained of around 90%. The main differences in the majority of these approaches are in the features used for classification. Significant work has been done by Alex Hai Wang [2] in the year 2010 which used user based as well as content based features for detection of spam profiles. A spam detection prototype system has been proposed to identify suspicious users in Twitter. Classic evaluation metrics have been used to compare the performance of various traditional classification methods like Decision Tree, Support vector Machine (SVM), Naïve Bayesian, and Neural Networks and amongst all Bayesian classifier has been judged the best in terms of performance. Over the crawled dataset of 2,000 users and test dataset of 500 users, system achieved an accuracy of 93.5% and 89% precision. Limitation of this approach is that is has been tested on very less dataset of 500 users by considering their 20 recent tweets. In year 2010, Lee et al. [3] deployed social honeypots consisting of genuine profiles that detected suspicious users and its bot collected evidence of the spam by crawling the profile of the user sending the unwanted friend requests and hyperlinks in MySpace and Twitter. Features of profiles like their posting behaviour, content and friend information to develop a machine learning classifier have been used for identifying spammers. After analysis profiles of users who sent unsolicited friend requests to these social honeypots in MySpace and Twitter have been collected. LIBSVM classifier has been used for identification of spammers. One good point in the approach is that it has been validated on two different combinations of dataset – once with 10% spammers+90% non-spammers and again with 10% non-spammers+90% spammers. Limitation of the approach is that less dataset has been used for validation. Similarly Benevenuto et al. [4] detected spammers on the basis of tweet content and user based features. Tweet content attributes. . Dataset of 54 million users on Twitter has been crawled with 1065 users manually labeled as spammers and non-spammers. A supervised machine learning scheme i.e. SVM classifier have been used to distinguish between spammers and non spammers. Detection accuracy of the system is 87.6% with only 3.6% non-spammers misclassified. Twitter facilitates its users to report spam users to them by sending a message to "@spam". %. A forward step in the same field was taken by McCord et al. (2011) [5] using user based features like number of friends, number of followers and content based features like number of URLs, replies/mentions, retweets, hashtags of collected database. Classifiers namely Random Forest, Support Vector machine (SVM), Naïve Bayesian and K-Nearest Neighbour have been used to identify spam profiles in Twitter. Method has been validated on 1000 users with 95.7% precision and 97.5% accuracy using the Random Forest classifier and this classifier gives the best results followed by SMO, Naïve Bayesian and K-NN classifiers. Limitation of this approach was that for considered dataset reputation feature had been showing wrong results i.e. it is not able to differentiate spammers and non-spammers, unbalanced dataset has been used so Random Forest is giving best results as this classifier is generally used in case of unbalanced dataset, and finally the approach has been validated on less dataset, and finally the approach has been validated on less dataset. Then onwards in 2013, Lin et al. [6] detected long-surviving accounts in Twitter on the basis of two different features that are URL rate and interaction rate. URL rate is the number of tweets with URL/ total number of tweets and interaction rate is the number of tweets interacting/ total no of tweets. 26,758 accounts have been crawled using Twitter API and 816 long surviving accounts have been analyzed J48 classifier with 86% precision. Limitation of the approach is that only two features have been used for spam profile detection and if spammers keep low URL rate and low interaction rate then this technique will not work as intended. Similarly in 2011, Chakraborty *et al.* [7] have proposed a system to detect abusive users who post abusive contents, including harmful URLs, porn URLs, and phishing links and divert away regular users and harm the privacy of social networks Miller et al. (2014) [8] attempt to treat the identification of spammers as an anomaly detection and not classification problem where outliers are flagged as spammers. They utilize a combination of user metrics and one gram text features. They then test two algorithms: DBSCAN which uses a density based similarity metric and K-Means which uses an Euclidean distance based metric. These approaches achieved an 82% and 71% F1 score respectively with high accuracy but low precision. M.A Fernandes et al. [9] compared classification and clustering approaches to separate human from not human users in Twitter. An initial feature set of 70 variables was reduced to the most relevant for classification, thereby decreasing complexity and improving generalization performance.

3. IMPLEMENTATION

The work is performed in five steps namely; extracting the data, Pre-processing, Feature Extraction, Clustering and results. Extraction involves the use of R package named tool which helps to get the data from the Social Networking Website Twitter, Pre-processing is the method which is to be performed on the data provided so that the complexity can be reduced. After this Feature Extraction is performed and for these certain parameter has been set and according to those parameters whole data gets checked and then evaluation is performed.

3.1 Extracting the data

R package is the tool which is used to extract the data from Twitter. This tool is basically used for statistical Computing and graphical display. It is free environment software. It runs on windows, Linux and Mac OS. R can be easily extended with 6,600+ packages available on CRAN.R package in this research work has been used with the purpose of extracting the tweets from the Twitter Website. R package is an interface which is used with the Twitter Web API.

3.2 Pre-processing

This is performed to reduce the complexity of the data provided by making it simple and easy to read.

To do this two techniques have been used and these are:

3.2.1 Stop Word Removal- This is the process of removing the unwanted and unnecessary words from the text. For this, a separate file is maintained with the name stop words, so that whenever those words comes in the input they automatically gets removed [10].

For example: if we have maintained a file and in that file the words written are „a‟ an‟, the‟ etc. So, whenever these words will appear in the input they automatically get removed and hence, complexity is reduced of the document

3.2.2 Stemming- This is the technique to find out the root or the stem of the word or in other way it is defined as the process of finding out the origin of the word [11].

For example: Relation, Relationships, Relative, Related are the words which has some origin and that is Relate

3.3 Feature Extraction

This is the process in which number of features set is maintained on the basis of which input is examined. In this work, the set maintained is combination of five features and that is no. of spam words per input (scount), total number of" this is considered to be as spam. Words in a line (wordcount), no. of URL (URL), no. of URL present per wordcount (URL/wordcount), Retweets (rt). For example: Input: RT @BBCSport: Stuart McCall appointed #Rangers manager until the end of the season http://t.co/oJRIISNwYI http://t.co/J1AXlQGfvP.

3.4 Improve K-Means Clustering

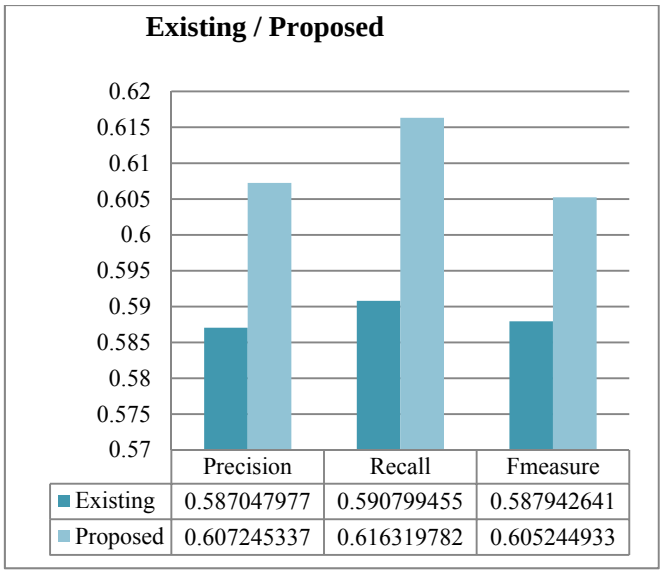
The concept of clustering has emerged for a long time. In database management, clustering data is the process of dividing data element (input data) into groups so that items in the same group are as similar as possible and items in different groups are as dissimilar as possible. Clustering is an unsupervised learning and one of the most useful methods in data mining for detection of natural groups in a dataset-means clustering algorithm and groups data based on their feature values into K clusters. In classification the objects are assigned to predefined classes, whereas in clustering the classes are formed. There are general categories of cluster analysis methods such as tree clustering, block clustering, EM clusters and K-Means clustering. .K-means clustering algorithm, is numerical and one of the hard clustering methods, this means that a data point can belong to only one cluster. This study utilized the K-means clustering algorithm to group the messages based on the similarity of their attributes or features into K disjoint groups using Euclidean distance, to improve the accuracy of spam detection. K is a positive number initialized early, before the algorithm start, to refer to the number of required clusters. Basically, K-means clustering €objects within each cluster are similar to each other and distinct from objects in other clusters. K-means clustering is an iterative algorithm, it starts by defining an initial set of sorted clusters and the clusters are repeatedly updated until no more improvement is possible (or the number of iterations exceeds a specified limit).

3.5 Evaluation

On the basis of the process performed above, Precision, Recall, F Measure are evaluated and then the corresponding graphs get generated.

4. RESULT AND ANALYSIS

Entire work is performed in MATLAB which is abbreviated as Matrix Laboratory, it is used to perform the data visualization, data analysis and it is a very high-level interactive programming language [12]. Improve K-Means is implemented in MATLAB for improving time efficiency and various other parameters. For performing this work, a dataset of 150 records has been maintained and those records are basically the tweets, which are downloaded from Twitter using the R package tool and that data is saved as tdata.csv , here extension .csv means "comma separated values" A Comma Separated Values files stores tabular data in plain text. Plain text means that the file is interpreted a sequence of characters, so that it is human readable with a standard text editor. Each line of the file is a data record. On various Percentage of Untrained data the Precision, Recall and F-measure has been analysed.



	Precision	Recall	Fmeasure
Existing	0.587047977	0.590799455	0.587942641
Proposed	0.607245337	0.616319782	0.605244933

Fig. 1: Performance comparison with existing approach

5. CONCLUSION AND FUTURE SCOPE

This study proposed a improve K-Means clustering for detecting human or not human users. Results have been obtained on different percentages of the untrained data and the respective graph has also been obtained. Using this approach our suggested features can achieve 99% Precision, 97% recall, and 98 % F-Measure but the features which are identified here are just related to the spam data like no. of URL in the tweet or number of spam words etc. In future there is need to increase the testing dataset and researcher can do the work of human or not human user detection by including the features like user's friend, frequent links etc or can introduce an entire new methodology by using any other clustering technique.

REFERENCES

[1] Data Mining with R, http://www.rdatamining. Com/docs/introduction-to-data-mining-with-r Accessed on 20-May-2016

[2] Alex Hai Wang, Security and Cryptography (SECRYPT), Don't Follow Me: Spam Detection in Twitter, Proceedings of the 2010 International Conference, Pages 1-10, 26-28 July 2010, IEEE.

[3] Kyumin Lee, James Caverlee, Steve Webb, Uncovering Social Spammers: Social Honeypots + Machine Learning, Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval, 2010, Pages 435–442, ACM, New York (2010).

[4] Fabricio Benevenuto, Gabriel Magno, Tiago Rodrigues, and Virgilio Almeida, Detecting Spammers on Twitter, CEAS 2010 Seventh annual Collaboration, Electronic messaging, Anti Abuse and Spam Conference, July 2010, Washington, US.

[5] M. McCord, M. Chuah, Spam Detection on Twitter Using Traditional Classifiers, ATC'11, Banff, Canada, Sept 2-4, 2011, IEEE.

[6] Po-Ching Lin, Po-Min Huang, A Study of Effective Features for Detecting Long-surviving Twitter Spam Accounts, Advanced Communication Technology (ICACT), 15th International Conference on 27-30 Jan. 2013, IEEE.

[7] Ayon Chakraborty, Jyotirmoy Sundi, Som Satapathy, SPAM: A Framework for Social Profile Abuse Monitoring.

[8] Miller, B. Dickinson, W. Deitrick, W. Hu, and A. H. Wang. Twitter spammer detection using data stream clustering Technical report, Department of Computer Science, Houghton College, Houghton, NY, USA, 2014.

[9] M.A Fernandes, P.Patel, and T. Marwala, Automated Detection of Human users in Twitter. Department of Electrical and Electronic Engineering University of Johannnesburg, South Africa, Volume 53, 2015, pages 224-231.

[10] StopWords,https://en.wikipedia.org/wiki Stop - words. Accessed on 20-May-2016

[11] Stemming https://en.wikipedia.org/wiki/ Stemming. Accessed on 20-May- 2016

[12] MATLAB, http://in.mathworks.com/products /matlab/ Accessed on 20-May-2016