



Automated Detection of Spammers' Profiles using Improved K-Means in Twitter

¹Sanjeev Dhawan, ²Kulvinder Singh, ³Isha

^{1,2}Faculty of Computer Science & Engineering, ³Student of M.Tech (Computer Engineering),

^{1,2,3}Department of Computer Science & Engineering,

University Institute of Engineering and Technology (U.I.E.T),

Kurukshetra University, Kurukshetra (K.U.K)-136119, Haryana, INDIA

Abstract: 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

I. Introduction

Twitter 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. These tweets are public by default and visible to all those who are following the twitter. Hash tags are those which starts with special characters # and which is meant to group similar micro blog topics such as #economics and #amazing. With larger user databases in OSNs, twitter is becoming a more interesting target for spammers/malicious users [1]. Spam (www.spamhaus.org) 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. Spammers are the malicious users who contaminate the information presented by legitimate users and in turn pose a risk to the security and privacy of social networks. Spammers belong to one of the following categories:

- (a) Phishers: are the users who behave like a normal user to acquire personal data of other genuine users.
- (b) Fake users: are the users who impersonate the profiles of genuine users to spend spam content to the friends of that user or other users in the network.
- (c) Promoters: are the ones who send malicious links of advertisements or other promotional links to others so as to obtain their personal information.

II. Related Work

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. 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 it has been tested on very less dataset of 500 users by considering their 20 recent tweets. Similarly Benevenuto *et al.* [3] 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) [4] 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. Miller *et al.* (2014) [5] 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.* [6] 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

III 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.

A. Extracting the data

This is the very first step of performing the detection of user identity. As, it has already been said that for the detection of spammers there is a requirement of some input data. The data from the Twitter can be downloaded with the help of a tool which is called as R package tool. R based Twitter client provides an interface to the Twitter web API. R is a free software environment for statistical computing and graphics. It runs on windows, Linux and Mac OS.

B. Pre-processing

The pre-processing of the data is performed and this is done with the help of the Stop Word Removal Technique. Stop word is referred to as the removal of the common and unnecessary words before or after processing of the natural language data (text). Most commonly used stop words are *the, is, at, which* and *on*. The stop word removal process is followed by Steaming. These both processes are included in the pre-processing section.

C. Feature Extraction

This is referred to as the process of extracting out the features which are redundant in the provided input to the algorithm. When the input given to the algorithm is too large for the processing then some of the features get reduced. This process is called Feature Extraction. In this work, the set maintained is combination of four features and that is, total number of words in a line (word count), no. of URL (url), no. of URL present per word count (URL/word count), Retweets (rt).

For example: Input: Nine-Zero raid: #Rangers to detain 26 #MQM suspects for 90 days #Karachi <http://t.co/yaN0MA8nN0> <http://t.co/nV7EXfWoan>

D. Improve K-Means Clustering

K-Means is a partitioning based clustering algorithm having the data objects with similarities in one cluster based on the Euclidian distance. It is the simplest and most widely used algorithm for clustering. It is an unsupervised learning method that follows an easy procedure for cluster data set. This algorithm always forms *k* number of clusters and in every cluster there must be at least one item. In the initial step, K-means chooses fixed centroid points and number of clusters. It then finds the distance of each data point from every centroid. The points are assigned to the centroid at minimum distance from the point. The centroids are recalculated by taking the arithmetic mean of the data points assigned to that cluster. The performance of K-means increases as the number of cluster increases; hence it is advantageous to apply on large datasets.

Algorithm**Input:***K* is the random number of clusters*D* is the data set having *n* objects**Output:***'k'* clusters are formed.**Method:**

1. Select *k* fixed data objects as the cluster centroids.
2. Assign cluster to each data object to its closest centroid.
3. Recalculate the new centroids for each cluster, by calculating the arithmetic mean of data objects in that cluster.
4. If atleast one data object changes its centroid, then move to the step no 2, else move to next step.
5. Output the final clusters.

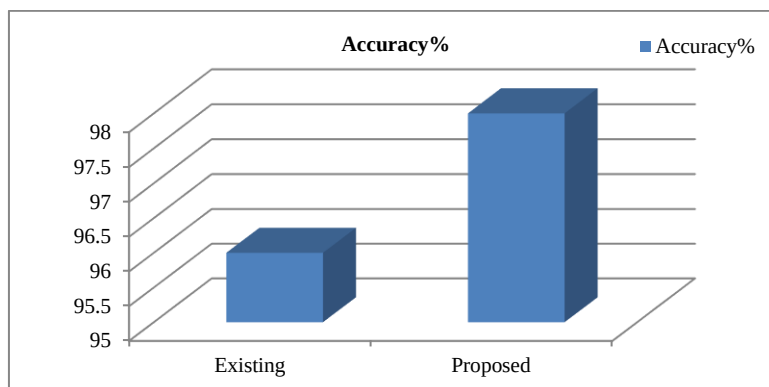
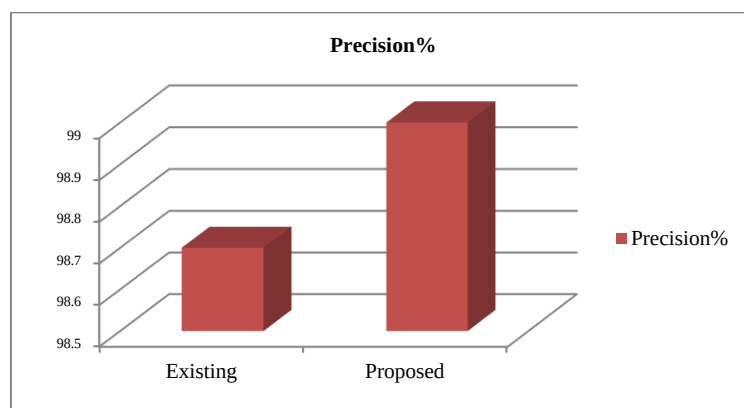
E. Evaluation

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. Improve K-Means is implemented in MATLAB for improving time efficiency and various other parameters [7]. 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 analyzed.

	Precision	Recall	Fmeasure
Existing	0.587048	0.590799	0.587943
Proposed	0.607245	0.61632	0.605245

Table 1: Results on the basis of Untrained Data

After performing the clustering process on the untrained data, the output can be obtained easily in terms of Precision, Sensitivity/Recall/TPR and Fmeasure.

a) Accuracy:**Figure 1: Accuracy****b) Precision:****Figure 2: Precision**

c) Recall:

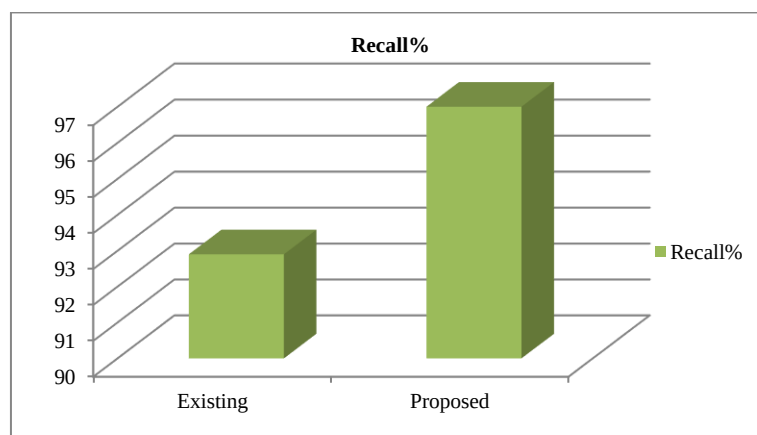


Figure 3: Recall

d) Fmeasure:

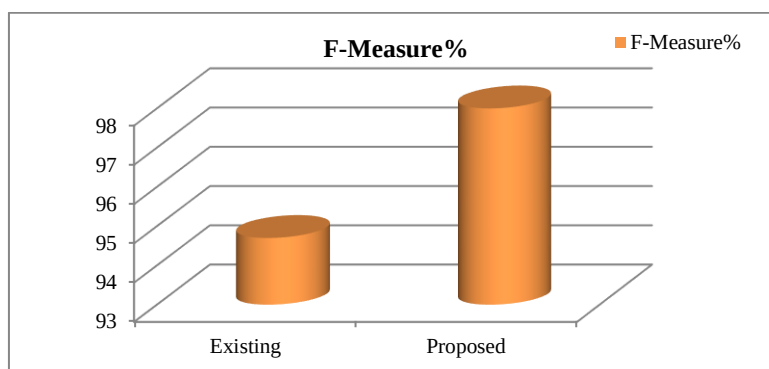


Figure 4: Fmeasure

IV 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] Sanjeev Dhawan, Kulvinder Singh and Isha, "Automated Detection of Spammers in Twitter", Vivechana: National Conference on Advances in Computer Science and Engineering (ACSE-2016). (Accepted and to be published in proceedings).
- [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] 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.
- [4] M. McCord, M. Chuah, Spam Detection on Twitter Using Traditional Classifiers, ATC'11, Banff, Canada, Sept 2-4, 2011, IEEE.
- [5] 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.
- [6] M.A Fernandes, P.Patel, and T. Marwala, Automated Detection of Human users in Twitter. Department of Electrical and Electronic Engineering University of Johannesburg, South Africa, Volume 53, 2015, pages 224-231.
- [7] Isha, Sanjeev Dhawan, Kulvinder Singh, "Improved K-Means for Detection of Human Users in Twitter", Advances in Computer Science and Information Technology (ACSIT), ACSIT 2016, pp. 301-304.(Published)