International Journal of Innovative Research in Computerand Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.194 |

||Volume 8, Issue 4, April 2020||

Detecting Spammers in Social Media Using Data Mining Techniques

Dr. R. Vijayalakshmi[1], Dr. R. Deepalakshmi[2,] S.Gowtham[3], M.Pranav Subiksun[4], P. Nithish Kumar[5]

Associate Professor, Dept. of CSE, Velammal College of Engineering and Technology, Madurai, Tamil Nadu, India[1]

Professor, Dept. of CSE, Velammal College of Engineering and Technology, Madurai, Tamil Nadu, India[2]

UG Students, Dept. of CSE, Velammal College of Engineering and Technology, Madurai, Tamil Nadu, India[3,4,5]

ABSTRACT: Now a day's human relations and information are maintained by social media networks.To communicate with others .maintain records, sharing knowledge and exchanging media between social media networking sites are used. The social media such as Facebook ,Twitter , LinkedIn ,etc are used for this purpose. Twitter allows media users to share their interests, opinions and knowledge to others in the form of tweets and messages. At the same time some of the users misguide the genuine users. These users who misguide other users are referred to as spammers. The non-spammers may retweet these spam tweet to others and the cycle continues. To avoid this spam messages the proposed architecture uses a methodology of using machine learning algorithm to classify the spammers.

KEYWORDS: Social media, Spammer , Non-Spammer, tweets, Machine learning ,classification algorithm

I. INTRODUCTION

The proposed system architecture has been implemented using Linear Kernel Support Vector Machine-based classifier to classify the data and for predicting the efficiency of the proposed work. The goal is to identify the conditions under which coordination is more likely to arise from networks that are constantly pulsating with information. The proposed system aims to relax these assumptions and allow actors to repeatedly activate as a function of the dynamics unfolding in the rest of the network. This modification aligns the propose model of contagion more closely with what is observed in many empirical networks.

The system has theoretical and empirical reasons to allow repeated activation to be the driving force of contagion dynamics. The main reason is that most instances of diffusion does not mainly involve a single activation but many activations building up momentum in time. Before a hashtag(#) becomes a trending topic, a period of buzz is first required prior duely to the protest day. The Actors decide whether they want to engage in an online conversation or to take part in a protest. This is what threshold models are able to capture. What threshold models cannot capture is the period of information exchange. The proposed Model aims to capture this temporal dimension.

The proposed System assumes that exposure to information is the driving force underlying contagion. What makes this model different from previous models is that failure to trigger a chain reaction depends not only on the distribution of threshold or the impact of network structure on activation dynamics; it also depends on whether the network facilitates coordination, that is, an alignment of actions in time − which is an important organizational goal for social movements that want to gain public visibility in social media or use online networks to manage mobilization By focusing on coordination dynamics, the proposed system is in a better position to explain why, more often than not, large-scale contagion fails to take off. If the network is not conducive to coordination contagion ends up trapped in local activity clusters and, therefore, fails to synchronize the actions of the majority.

The main assumption is that actors reach their activation zone at different speeds. The speed of activation is a function of two parameters: ω, which determines how quickly the actor reaches the threshold zone and ε, or the strength of the signal received from other actors − which, in our case, is restricted to actors one step removed in the network.

II. METHODOLOGY

The dataset is a list of tweets that contains the fields of user id and their tweets.The tweet is a piece of message that is posted by the user.The selected dataset is going to be used for tracking and clustering user interest. The data is pre-processed to remove reviews from anonymous users, since we would like to associate each review with a unique user.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.194 |

||Volume 8, Issue 4, April 2020||

The Data pre-processing is the process of detecting, correcting or removing, corrupt or inaccurate records from the dataset. The records which provide the incorrect clustering results are detected and removed from the dataset

The tweets are broken up as a sequence of strings into pieces such as words, keywords, phrases, symbols and other elements called tokens. Next the stop words such as “the, a, and,..” are detected from the reviews and it’s removed. The porter stemming algorithm is used for stemming the words. After the stemming process, the term frequency are constructed from the stemming words.

Data splitting is the act of partitioning available data into two portions, usually for cross-validatory purposes. One. portion of the data is used to develop a predictive model. and the other to evaluate the model's performance. Separating data into training and testing sets is an important part of evaluating data mining models. Mostly, when you separate a data set into a training set and testing set, most of the data is used for training, and a smaller portion of the data is used for testing.

Classification is used to classify each item in a set of data into one of predefined set of classes or groups. The goal of classification is to accurately predict the target class for each case in the dataClassification algorithm is used for classifying the data based on their interest.Classification result displays the Spam and Non-spam details.The goal of classification is to accurately predict the target class for each case in the data.The overall clustering report is generated based on their interest in twitter.The report which contain dynamically distributed topics in the tweeter and their level. The level that describe how many times Spam and non-spam detailsandalso it will show the result of particular things.

III. SYSTEM ARCHITECTURE

The below block diagrams explains the architecture of the proposed system (Fig 3.1 Architecture Diagram and Fig 3.2 Flow Diagram):

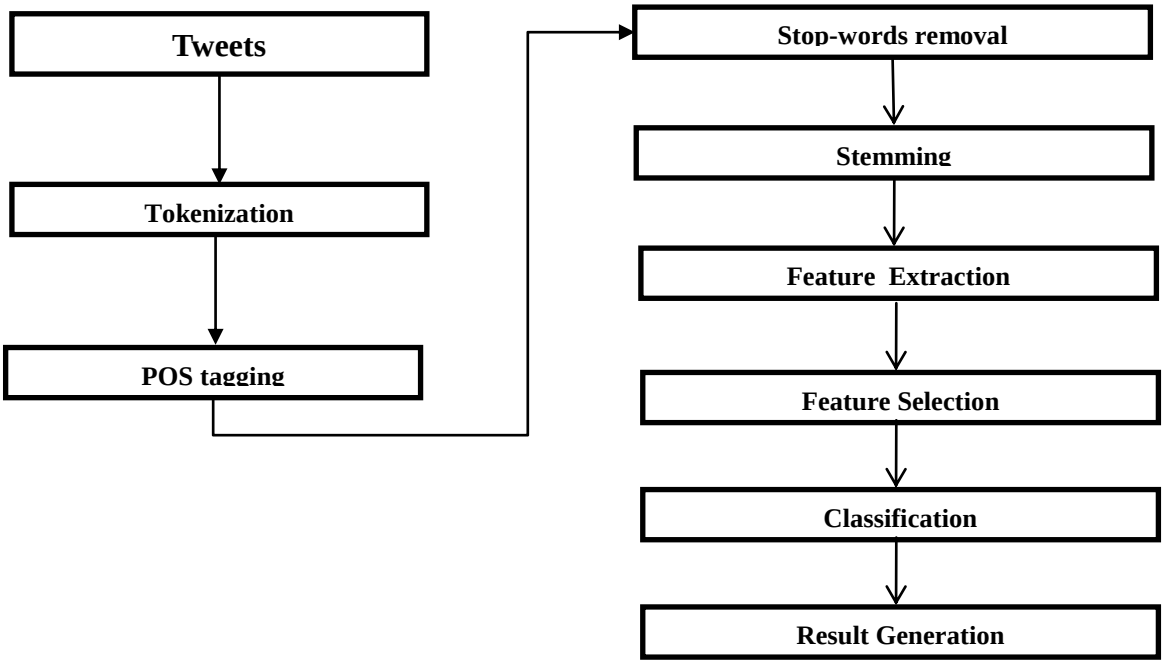

Fig 3.1 Architecture Diagram

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.194 |

||Volume 8, Issue 4, April 2020||

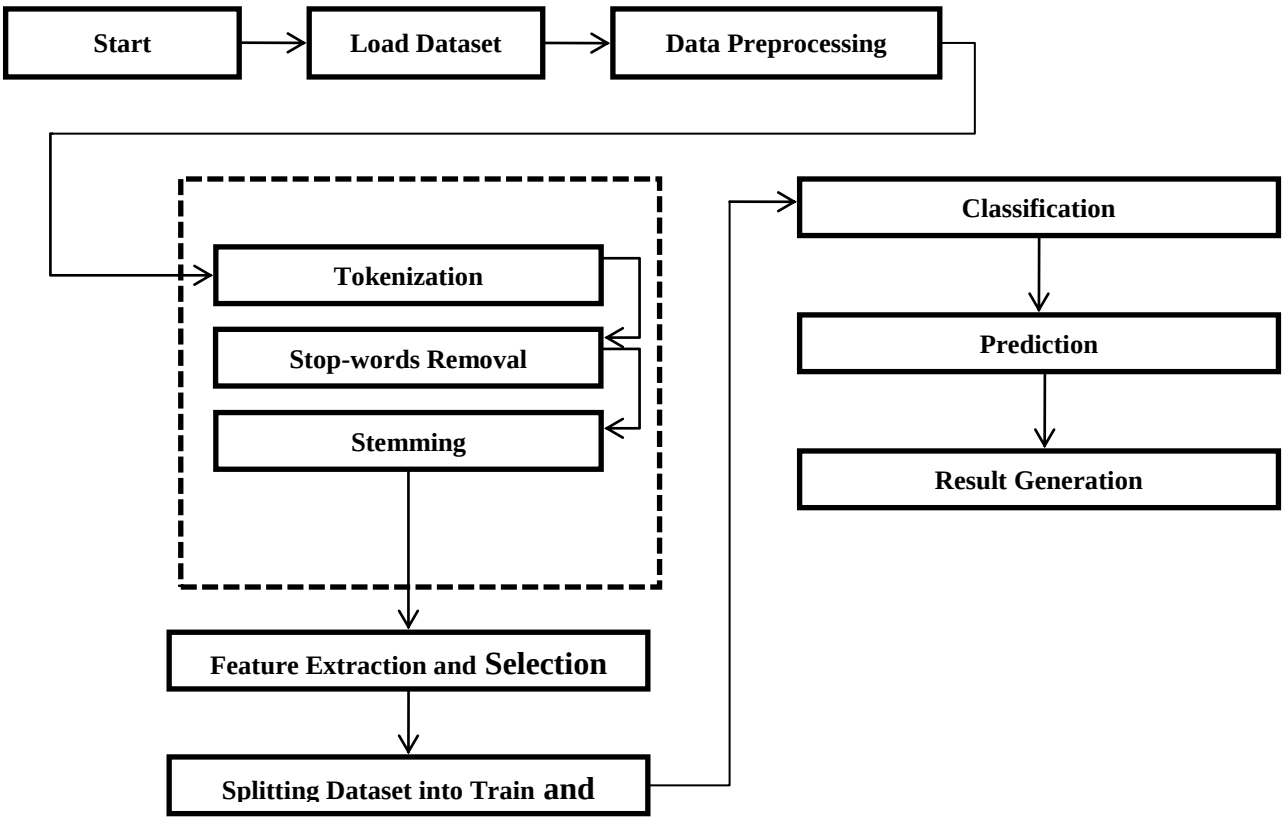

Fig 3.2 Flow Diagram

IV. RESULT

This project enhances the performance of the overall clustering and prediction.It finds the different clusters and it is summary effectively and quickly.It alleviates the sparsity problem and reduces the information loss.The accuracy of the clustering and classification result is highly increased. It effectively tracks and cluster the user by avoiding the sparsity problems and also this project enhances the performance of the overall tracking and clustering.

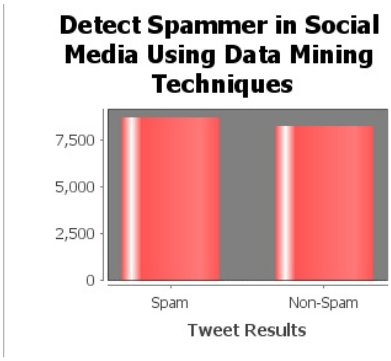

Fig 4.1 spammer vs non-spammer

The tweets are classified as spam and non-spam data and the results are represented graphically in the figure 4.1.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.194 |

||Volume 8, Issue 4, April 2020||

V. CONCLUSION AND FUTURE WORK

This project enhances the performance of the overall clustering and prediction.It finds the different clusters and it is summary effectively and quickly.It alleviates the sparsity problem and reduces the information loss.The accuracy of the clustering and classification result is highly increased

In future the proposed system can store all the data for increasing hadoop storage the processing speed. This process will increase the effectiveness of the data storage, processing and classification results. These data classified as spammers can be stored in the cloud database and the tweets can be classified based on these results of the proposed system. This minimizes the risk of spam data to be retweeted again and again so that the processing efficiency of the tweet is improved.

REFERENCES

[1] Aral, Sinan, Walker, Dylan, 2012. Identifying influential and susceptible membersof social networks. Science 337, 337–341.

[2] Lakshman prabhu S,Joel J.P.C Rodriguez,Victor Hugo,2017 Effective features to classify big data using social internet of things Grant 074-U01

[2] Aral, Sinan, Muchnik, Lev, Sundararajan, Arun, 2009. Distinguishinginfluence-based contagion from homophily-driven diffusion in dynamicnetworks. PNAS 106 (51), 21544–21549.

[3] Aral, Sinan, Muchnik, Lev, Sundararajan, Arun, 2013. Engineering socialcontagions: optimal network seeding in the presence of homophily. Netw. Sci.1 (02), 125–153.

[4] Backstrom, Lars, Boldi, Paolo, Rosa, Marco, Ugander, Johan, Vigna, Sebastiano,2012. Four degrees of separation. In: Proceedings of the 3rd Annual ACM WebScience Conference, Evanston, Illinois: ACM, pp. 33–42.

[5] Barabási, Albert-László., 2009. Scale-free networks: a decade and beyond. Science325 (5939), 412–413.

[6] Barberá, Pablo, Wang, Ning, Bonneau, Richard, Jost, John, Nagler, Jonathan, Tucker,Joshua, González-Bailón, Sandra, 2015. The critical periphery in the growth ofsocial protests. PLoS One 10 (11).

[7] Bond, Robert M., Fariss, Christopher J., Jones, Jason J., Kramer, Adam D.I., Marlow,Cameron A., Settle, Jaime E., Fowler, James H., 2012. A 61-million-personexperiment in social influence and political mobilization. Nature 489, 295–298.