ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

A Frame Work for Communities Detection in Social Media Networks

K.Gowthami

M.Tech, Dept. of C.S.E., Adikavi Nannaya University College of Engg, Rajahmundry, India

ABSTRACT: Enormous participation of users creates new opportunities to study human social behavior along with the capability to analyze large amount of data streams. Community detection in social media helps to reshape business models. Social sciences frames the concept of community, the problem of community detection in the context of social media provides a compact classification of existing algorithms based on their methodological principles. Positive or negative attitudes can be inferred from the discussions it require a formal interpretation of social media logs and unit of information that can spread from person to person through the social network. Once the social media data such as user messages are parsed and network relationships are identified, datamining techniques can be applied to group different types of communities. The Proposed framework introduces the novel task of detecting communities by clustering messages from large streams of social data. This frame work uses clustering algorithm and groups the user communities based on their activities. This approach is optimized and scalable for real-time clustering of social media data.

KEYWORDS: Social media, Communities, Social Networks, Datamining

I. INTRODUCTION

Social media have grown quickly in popularity in a relatively short time. Social media serves as a ubiquitous public platform which remains accessible to users as a multiple group of internet applications. Within the applications, the user creates individual and unique expression for data exchange [2]. What remains valuable and fascinating is the level of social media data influence as the platform remains an intense portal of human interactions and behaviors. What remains dynamic about social media is the level of opportunity that influences individuals, groups and society. The study of this data by industry specialists seeking new and inventive methods to collect data for analysis remains important to the future of social media [3]. There are so many social media network sites, Twitter and Facebook remain the most well-known but other forms being used are blogs, wikis and platforms with unfiltered text and information which remains of true focus for users.

Industry researchers remain focused on social media application business, bioscience and social science. What has been found is that the social media is extremely valuable to statistical study of information technology, social behavior [4] within quantitative attributes for e-learning process and simulation design for further data mining [5].The rate of digital interaction and exchange of data amongst users increases. These platforms like Facebook and Twitter and some more sites, user expression of opinions and the place many are getting his or her news [6]. These are platforms of free speech and protests [7], organization and sharing of common interests and ways to keep family and friends in the loop of everyday life. Still challenge to industry specialists that collect data from these platforms and social communities. Such data mining actions require the ability to interpret the massive amount of content continuously produced on online social media and manual labelling is infeasible on a large scale. Textual content equals a unit of information and this can also be coded to represent a particular trait of the individual posting the content. Each piece of content also represents a user’s score point and this makes the process of assessing information from the standpoint of learning methods as these acts as individual means of identifying group behaviours.

In this paper, combine text mining, network analytics and data mining together to provide a better description of each user in forum in terms of leadership and sentiment. For each individual user, the authority score, hub score, and rating are calculated by using the network analytics. High authority scores differentiate between users with a high level of leadership where as high hub score represents users with high follower behaviour. The attitude of each individual user in the forum is measured by rating. So by using the attitude measure to identify the positive, negative and neutral.

II. RELATED WORK

The amount of information shared on online social media has been growing during recent years [1]. Much can be learned about the retail and finance behaviors of users by studying social media analysis. It is nothing new that retail companies market via social networks to discover what consumers think about branding, customer relationship management, and other strategies including risk prevention. A good example is the found correlation of data on Twitter with industry market behavior and sentiment posted by users. Wolfram [9] used Twitter data to develop machine learning model using Support Vector Regression and predicted prices of individual stocks and found significant advantages of using social media data for forecasting future prices.

Social media analysis data can be used to track health issues like smoking and obesity for bio-scientific study like Penn State University found innovative systems and techniques to track the spread of infectious diseases because the data social media reflects about users within these groups [10]. Social science applications are concerned, it includes monitoring public responses to announcements, speeches and events with emphasis on political comments and initiatives. It also gives insights about community behavior, social media

Polling within groups and early detection of emerging events like, For example by using the computational linguistics, the automatic prediction impact of news on the public perception of political candidates was implemented. Yessenov and Misailovic [11] use reviewing comments of movie. Karabulut found that Facebook also exhibits and captures major public events in its data.

Social network analysis has a well-defined relation and background in sociology [12]. With the rapid growth of the web forums and blogs, the user's participation on content creation led to a huge amount of dataset. Hence the advancement of data mining techniques is required. An overall discussion of one news forum called Slashdot, can be found in[13] Social networks, it focus work like facepager. It is used to access data from social media like facebook by using this data to develop a clustering framework using optimized K-Means algorithm that is more accurate than existing methods. Clustering is used as an exploratory analysis tool that aims at categorizing objects into categories, so the association degree between the objects is maximal when belonging to the same categories. Clustering structures the data into a collection of objects that are similar or dissimilar and is considered an unsupervised learning. The application of our method is mainly on finding user groups based on activities and attitude features as suggested in the authority model.

III. METHODS

In this section describe K-Means clustering algorithm and Genetic algorithm in detail.

A. K-Means Clustering

K-Means is an unsupervised clustering algorithm which is used to find groups within the data. Given a set of observations$(x_1, x_2, \ldots, x_n)$, where each observation is a d-dimensional vector, k-means clustering aims to partition the n observations into a set of k clusters (≤ n) such as A= $A_1, A_2, \ldots, A_n$ so as to minimize the within-cluster sum of squares (WSS) which is defined as sum of distance functions of each point in the
cluster to the k centers. The objective function of K-Means is to find

$$\mathrm{A}^{\arg\min} \sum_{i=1}^{k} \sum_{x \in A_i} ||x_i - c_i||^2$$

Where c_i is the centroid of points in A_i

B. Genetic Algorithm

Use of genetic algorithms fall under a larger standard of algorithms called evolutionary algorithms or EA. These EAs work to promote problem solving toward optimizing techniques which promote significant variance in genetic features like inheritance, mutation, selection and crossover of attributes. Much of what takes place happens as random with the population of user/individuals and this can be labelled as a generational group. To evaluate, each group generation is assessed for fitness within the optimization for serving to promote problem solving. The more up to the challenge the individual proves to be within the population, also signifies the rate of individual genom mutation which leads to creating the next generation of genetic characteristics. Thus, the algorithm continues to provide solutions within the generational testing. With the maximum amount of generations produced during the process, the algorithm finalizes and the fitness level standard is reached.

C. Proposed Algorithm

Proposed an optimized K-Means clustering algorithm with Genetic algorithm to find best possible centroids for K-Means algorithm. Genetic algorithm starts by generating initial population based on k provided as a parameter for number of clusters. Then, it computes the fitness of initial population Mean Square Error (MSE). The MSE calculates the distance between the cluster centers and remaining data points.MSE=$\frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2$

Where c is population center Fitness=$\frac{1}{(1+k_n)}$

Fitness value seems not to be changed in next generation. Once Genetic algorithm outputs the best centroids, we run K-Means algorithm on those centroids and get the clustering results. Which further process the K-Means clusters and try to minimize the within cluster variance and maximize distances between clusters. This is done by first calculating a pair wise distance between all clusters. Then the algorithm picks all pairs of clusters which have distances smaller than overall average of cluster distances. Once the pair of clusters are selected, the algorithm re-clusters them using the same K-Means and GA algorithms. Since at this point we are clustering small subset of data.

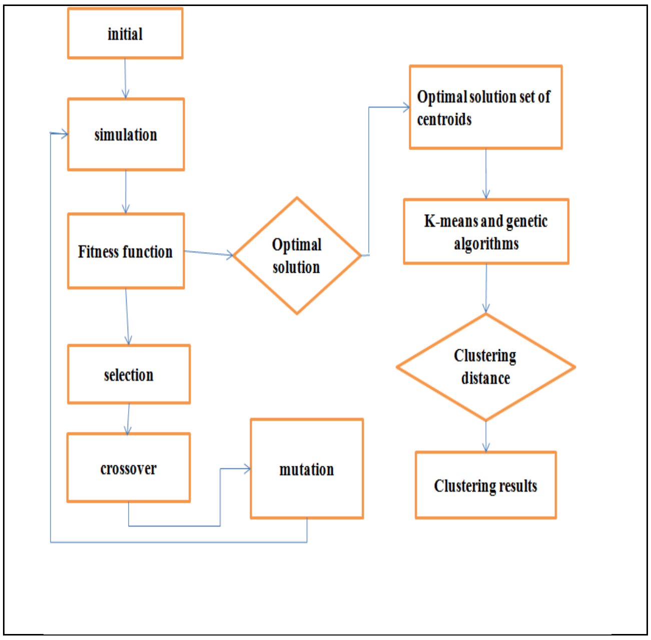


Fig: 1Flow diagram of Genetic Algorithm

Algorithm 1 Optimized K-means-Genetic algorithm

```
INPUT: Number of clusters C, Input dataset S, number of iterations T, population P
OUTPUT: Outputs are dataset with C clusters
Procedure
  Let i=1
   While $i \leq T$ do
     Fitness = ga (P(i), S)
     If fitness (i) > fitness (i+1)
       then
       centers = ga-centers(S, C)
     end if
   end while
   clusters = k-means(S, centers)
   dist = pairwisedistance(centers)
   for each dxy $\epsilon$ dist $\leq$ average(dist)
     do while
       max iterations do
       centers = ga-centers(S,2)
       newcluster = k-means(S, centers)
     end while
   end for
end procedure
```

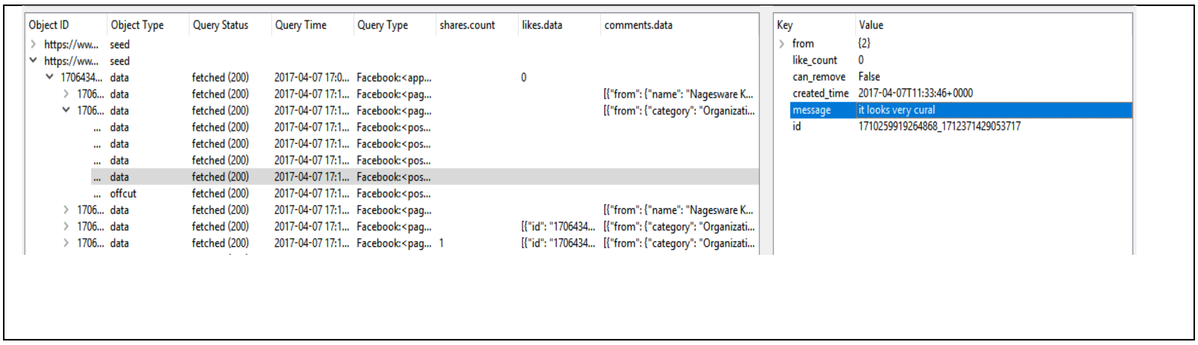
IV. DATASET DESCRIPTION

This section explains the dataset and experimental setup used in this project. Facepager is used it access the users posts, post comments and its data which is posted by the user are some other facebook users

There are many user communities that can serve as participant pools and it was found that an active community. Most of the users are registered and leave comments by their nickname, although some participate anonymously. by using attributes of facepager which is available in facepager tool to fetch data according to the selected attribute for example select post comments, then it will access the page and its continents like comments .In order to make facebook data useful for K-Means algorithm, processed this data using facepager tool .For each post comments are retrieve and finds its positivity and negativity and calculated its result.. The positive and negative attitude of each posts called rating. This rating is also calculated by extracting bag of words across all of the published posts using sentiment analysis. The description of each individual community user is defined by means of an authority score (leadership), a hub score (follower), and the level for the attitude of positivity or negativity. The opinion of the leaders posts spread quickly over a large number of users, hence they represent an important category. What has been found is that followers do not create influence over other users as predicted. In other words, their attitude does not influence other users. For example comment A. it is good for use B. Harm to health. Where 'A' is positive comment and 'B' is negative comment.

V. RESULTS AND DISCUSSIONS

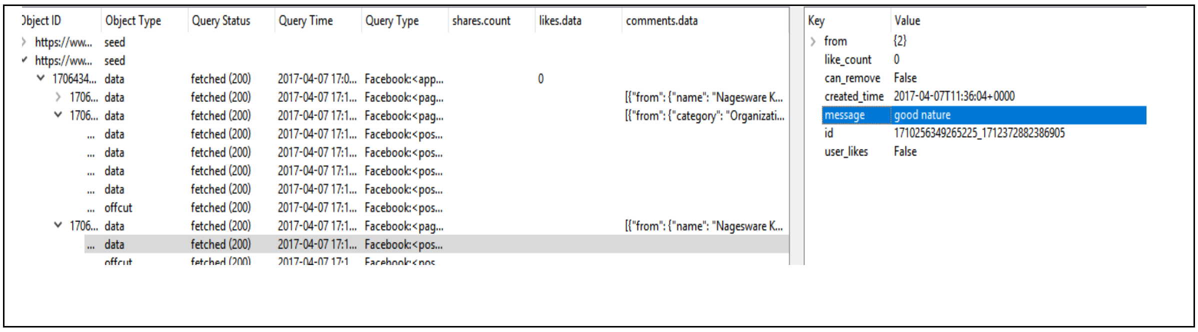
A set post page comments will be fetched from facebook by using fetching methods like in figure 2(a) and 2(b) shows that post comment data which is fetch from facebook .In fig 2(a) contain a negative comment data and in 2(b) positive comment data of selected post .Figure 3 shows that shared post, number of times it will be shared from one person to another. It shows that which post will be share more times compare to other posts in a page. Data will be accessed from facebook data. Figure 5 shows those communities. which is define according to the user activities .In this pie chart define communities by using the data which is fetching from social media like facebook, In facebook users create different kind of post and discussion on it like facebook comments. Fetch those data from facebook and find positive and negative comments in that data. Communities are framed according to the users activities in social media. In figure 6 shows that positive and negative comments, where 90.70 % are positive community and 9.29% are belongs to negative community it will be find based on post comments of users. this pie chart shows some rang of post comment values of selected page of facebook .page may be user created page , user like page are friends or followers page in facebook .



Object ID	Object Type	Query Status	Query Time	Query Type	shares.count	likes.data	comments.data
› https://ww...	seed						
⌄ https://ww...	seed						
⌄ 1706434...	data	fetched (200)	2017-04-07 17:0...	Facebook:<app...		0	
› 1706...	data	fetched (200)	2017-04-07 17:1...	Facebook:<pag...			[{"from": {"name": "Nagesware K...
⌄ 1706...	data	fetched (200)	2017-04-07 17:1...	Facebook:<pag...			[{"from": {"category": "Organizati...
...	data	fetched (200)	2017-04-07 17:1...	Facebook:<pos...			
...	data	fetched (200)	2017-04-07 17:1...	Facebook:<pos...			
...	data	fetched (200)	2017-04-07 17:1...	Facebook:<pos...			
...	data	fetched (200)	2017-04-07 17:1...	Facebook:<pos...			
...	offcut	fetched (200)	2017-04-07 17:1...	Facebook:<pos...			
› 1706...	data	fetched (200)	2017-04-07 17:1...	Facebook:<pag...			[{"from": {"name": "Nagesware K...
› 1706...	data	fetched (200)	2017-04-07 17:1...	Facebook:<pag...		[{"id": "1706434...	[{"from": {"category": "Organizati...
› 1706...	data	fetched (200)	2017-04-07 17:1...	Facebook:<pag...	1	[{"id": "1706434...	[{"from": {"category": "Organizati...

Key	Value
› from	{2}
like_count	0
can_remove	False
created_time	2017-04-07T11:33:46+0000
message	it looks very cural
id	1710259919264868_1712371429053717

Fig: 2(a) shows sample data of social media which contain negative comment



Object ID	Object Type	Query Status	Query Time	Query Type	shares.count	likes.data	comments.data
› https://ww...	seed						
✓ https://ww...	seed						
⌄ 1706434...	data	fetched (200)	2017-04-07 17:0...	Facebook:<app...		0	
› 1706...	data	fetched (200)	2017-04-07 17:1...	Facebook:<pag...			[{"from": {"name": "Nagesware K...
⌄ 1706...	data	fetched (200)	2017-04-07 17:1...	Facebook:<pag...			[{"from": {"category": "Organizati...
...	data	fetched (200)	2017-04-07 17:1...	Facebook:<pos...			
...	data	fetched (200)	2017-04-07 17:1...	Facebook:<pos...			
...	data	fetched (200)	2017-04-07 17:1...	Facebook:<pos...			
...	data	fetched (200)	2017-04-07 17:1...	Facebook:<pos...			
...	offcut	fetched (200)	2017-04-07 17:1...	Facebook:<pos...			
⌄ 1706...	data	fetched (200)	2017-04-07 17:1...	Facebook:<pag...			[{"from": {"name": "Nagesware K...
...	data	fetched (200)	2017-04-07 17:1...	Facebook:<pos...			
	offcut	fetched (200)	2017-04-07 17:1	Facebook:<pos			

Key	Value
› from	{2}
like_count	0
can_remove	False
created_time	2017-04-07T11:36:04+0000
message	good nature
id	1710256349265225_1712372882386905
user_likes	False

FIG: 2(B) SHOWS SAMPLE OF SOCIAL MEDIA WHICH CONTAIN POSITIVE COMMENT

.

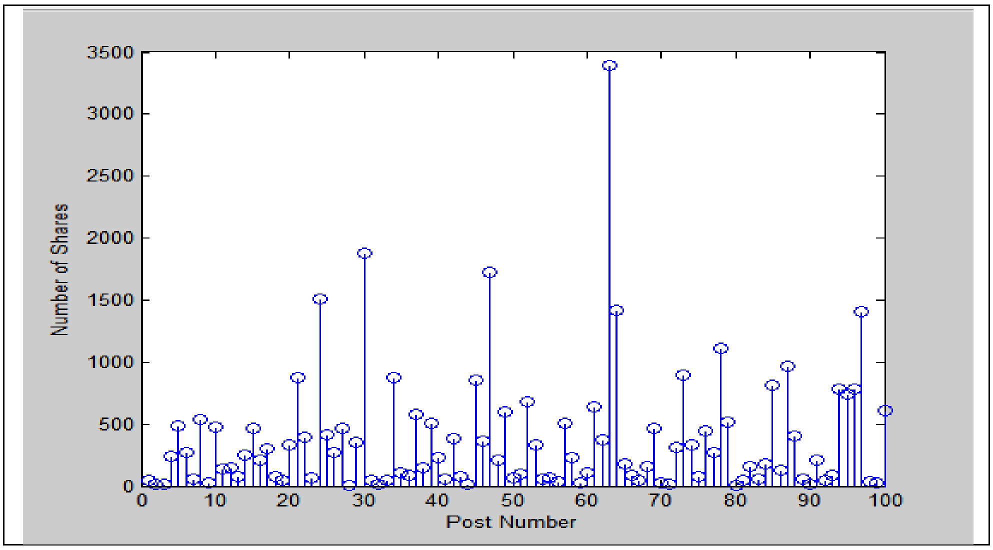
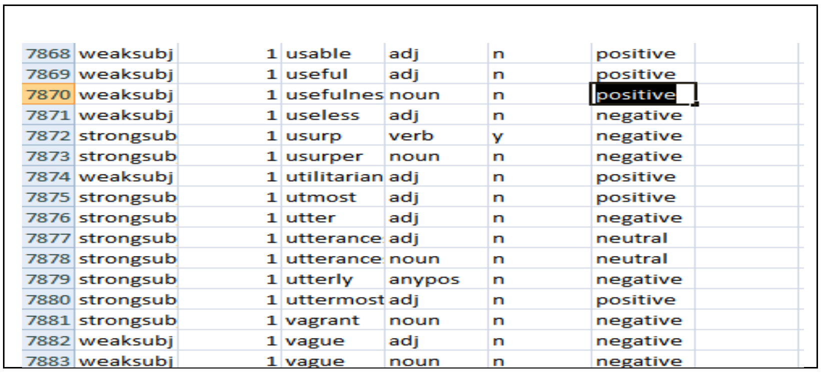
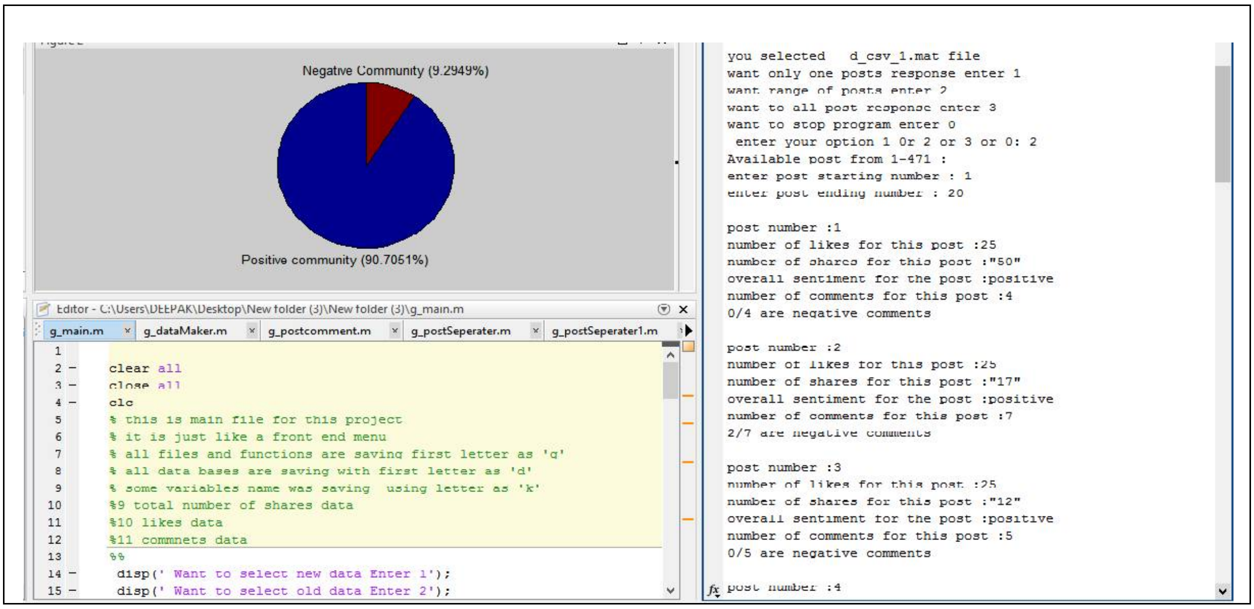


Fig: 3 Graph for share posts



7868	weaksubj	1	usable	adj	n	positive
7869	weaksubj	1	useful	adj	n	positive
7870	weaksubj	1	usefulnes	noun	n	positive
7871	weaksubj	1	useless	adj	n	negative
7872	strongsub	1	usurp	verb	y	negative
7873	strongsub	1	usurper	noun	n	negative
7874	weaksubj	1	utilitarian	adj	n	positive
7875	strongsub	1	utmost	adj	n	positive
7876	strongsub	1	utter	adj	n	negative
7877	strongsub	1	utterance	adj	n	neutral
7878	strongsub	1	utterance	noun	n	neutral
7879	strongsub	1	utterly	anypos	n	negative
7880	strongsub	1	uttermost	adj	n	positive
7881	strongsub	1	vagrant	noun	n	negative
7882	weaksubj	1	vague	adj	n	negative
7883	weaksubj	1	vague	noun	n	negative

Fig: 4 shows positive and negative words


Negative Community (9.2949%)
Positive community (90.7051%)
Editor - C:\Users\DEEPAK\Desktop\New folder (3)\New folder (3)\g_main.m
g_main.m
g_dataMaker.m
g_postcomment.m
g_postSeperater.m
g_postSeperater1.m

```
clear all
close all
clc
% this is main file for this project
% it is just like a front end menu
% all files and functions are saving first letter as 'g'
% all data bases are saving with first letter as 'd'
% some variables name was saving  using letter as 'k'
%9 total number of shares data
%10 likes data
%11 commnets data
%%
 disp(' Want to select new data Enter 1');
 disp(' Want to select old data Enter 2');
```

```
you selected   d_csv_1.mat file
want only one posts response enter 1
want range of posts enter 2
want to all post responce enter 3
want to stop program enter 0
 enter your option 1 0r 2 or 3 or 0: 2
Available post from 1-471 :
enter post starting number : 1
enter post ending number : 20

post number :1
number of likes for this post :25
number of shares for this post :"50"
overall sentiment for the post :positive
number of comments for this post :4
0/4 are negative comments

post number :2
number of likes for this post :25
number of shares for this post :"17"
overall sentiment for the post :positive
number of comments for this post :7
2/7 are negative comments

post number :3
number of likes for this post :25
number of shares for this post :"12"
overall sentiment for the post :positive
number of comments for this post :5
0/5 are negative comments

post number :4
```

Fig : 5 post comment result

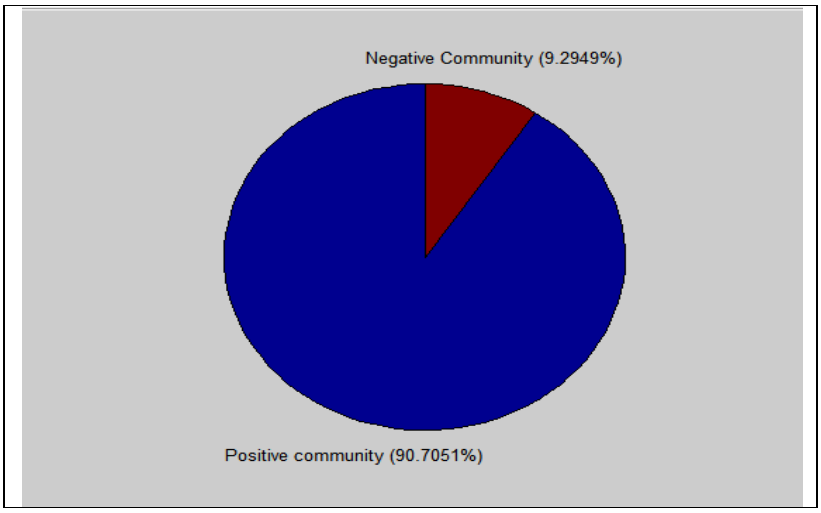


Fig: 6 pie chart for communities

VI. CONCLUSION

This project proposes a novel method to analyze social media data. In this method ,the method used to fetch data is facepager .The fetched data will be export into csv files .These csv files make access data easily to perform clustering algorithm .By separating the post comments according to the activities and find the result and make the communities based on the result of post comments.

REFERENCES

1. Emilio Ferrara, Mohsen JafariAsbagh, Onur Varol,Vahed Qazvinian, Filippo Menczer, and Alessandro Flammini. Clustering memes in social media. In Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pages 548–555. IEEE, 2013.
2. Andreas M Kaplan and Michael Haenlein. Users of the world, unite! the challenges and opportunities of social media. Business horizons, 53(1):59–68, 2010.
3. Bogdan Batrinca and Philip C Treleaven. Social media analytics: a survey of techniques, tools and platforms. AI & SOCIETY, 30(1):89– 116, 2015.
4. David Lazer, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. Life in the network: the coming age of computational social science. Science (New York, NY), 323(5915):721, 2009.
5. Claudio Cioffi-Revilla. Computational social science. Wiley Interdisciplinary Reviews: Computational Statistics, 2(3):259–271, 2010.
6. Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In Proceedings of the 19th international conference on World wide web, pages 591–600. ACM, 2010.
7. Michael D Conover, Clayton Davis, Emilio Ferrara, Karissa McKelvey, Filippo Menczer, and Alessandro Flammini. The geospatial characteristics of a social movemen communication network. PloS one, 8(3):e55957, 2013.
8. Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. Journal of Computational Science, 2(1):1–8, 2011.
9. M Sebastian A Wolfram. Modelling the stock market using twitter. School of Informatics, page 74, 2010.
10. Marcel Salathe, Linus Bengtsson, Todd J Bodnar, Devon D Brewer, John S Brownstein, Caroline Buckee, Ellsworth M Campbell, Ciro Cattuto, Shashank Khandelwal, Patricia L Mabry, et al. Digital epidemiology. PLoS Comput Biol, 8(7):e1002616, 2012.
11. Kuat Yessenov and Saˇsa Misailovic. Sentiment analysis of movie review comments. Methodology, pages 1–17, 2009.
12. John Scott. Social network analysis. SAGE Publications Ltd, 2013.
13. Vicenc¸ G´omez, Andreas Kaltenbrunner, and Vicente L´opez. Statistical analysis of the social network and discussion threads in slashdot. In Proceedings of the 17th international conference on World Wide Web, pages 645–654. ACM, 2008.