IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.8, Issue.5, May 2020 E-ISSN: 2347-2693

Keeping Track of Evolution of Trendy Topics in Social Media

D. Suneetha[1*], M. Shashi[2]

[1]Department of Computer Science and engineering, GITAM deemed to be University, Visakhapatnam, India
[2]Department of Computer Science & Systems Engineering, Andhra University, Visakhapatnam, India

[*]Corresponding Author: dwarapusuneetha@gmail.com, Tel.: +91-8985484764

DOI: https://doi.org/10.26438/ijcse/v8i5.132138 | Available online at: www.ijcseonline.org

Received: 07/May/2020, Accepted: 18/May/2020, Published: 31/May/2020

***Abstract*—** Social media platforms like Twitter facilitate interaction among people on topics of their interest which may vary with time. Hence identification of trendy topics from tweet streams should be a dynamic process. Topic identification with text clustering algorithms that focus on the content of the tweets do not suffice as tweets involves three types of features namely content, social context and temporal features. In this paper the author proposed a frame work that employs incremental clustering involving all the three types of features for clustering stream of tweets to produce set of clusters representing trendy topics at a series of time stamps. The proposed framework provides programmable selection / screening of interesting topics streaming on the Tweeter dynamically through proper parameter setting. Experimentation is done on real world data collected from Twitter on different domains.

Keywords — Social Context, Temporal Features, Incremental Clustering (key words)

I. INTRODUCTION

The pervasiveness of social media made it easier for the people to post anything at anytime from anywhere. Today's online social networking services like Face Book, Twitter, Google+, LinkedIn plays an important role in dissemination of information in a real time basis. Empirical studies show that online social networking service like Twitter is often the first medium to break important news about the events in a matter of seconds. Twitter facilitates real time flow of text messages known as "tweets" coming from different sources covering various kinds of subjects in distinct languages and locations. Millions of users are posting tweets on different topics ranging from technological events to worldwide protests. Extracting valuable information from these tweets is an important task and has become the popular research area.

Real time topic identification about emerging events is valuable if it is discovered timely. There are three types of features associated with tweets namely content-based, social context based and temporal based from which tweets are analyzed to extract knowledge. Clustering algorithms are extensively used for topic identification from text **[1]**. However, simple content based clustering is not appropriate for tweet clustering since, in addition to content based features, tweets have social context as well as temporal features. The authors developed a hybrid clustering algorithm **[1,2]** for tweet clustering based on content as well as social context in the form of re tweets and influential scores of actor who posted the tweets. In addition to social context temporal features of the tweets are also considered for further refining the tweet clusters at successive time stamps. Hence, topic identification from tweets should be dynamic producing set of clusters from tweet collection at successive time stamps along a given timeline and thus calls for incremental clustering.

This paper presents a frame work developed by the authors to automatically discover interesting subtopics dynamically from tweet stream which are used to identify influential users for each of the sub-topics based on the tweets and re tweets posted on the sub topics **[1]**. A new incremental tweet clustering algorithm is proposed to update the set of tweet clusters representing the sub-topics at successive time stamps along the time line.

The rest of the paper is organized as follows; **section 2** presents the related literature, followed by **section 3** that provides the details of the framework and the proposed incremental tweet clustering algorithm. **Section 4** presents the experimentation and analysis of results. **Section 5** concludes the research work with suggestions for future extension.

II. RELATED WORK

A. On summarization and timeline generation for evolutionary tweet streams :

The authors in the paper proposed a novel continuous summarization frame work called SUMBLR to deal with dynamic and fast arriving and large scale tweet streams. It contains three components in the first component they designed an algorithm namely online tweet stream clustering algorithm to maintain distilled statistics and store it in a data structure called as Tweet Cluster Vector

(TCV) and in the second step they designed TCV rank summarization technique for generating summaries. In the last step they effectively designed topic evolution detection method to detect distinct topics of a domain.

B. Topic discovery and future trend forecasting for texts :
The authors in this paper proposed a framework to automatically discover topics from a set of documents and forecast their evolving trend by considering data mining and machine learning as data domains. An association analysis process is applied followed by temporal correlation analysis and ensemble forecasting approach in order to identify the set of topics, discover correlation between the topics and analyze the popularity of the topics in future. Their frame work yields better performance and it is helpful to express large scale text collections in concise form. The frame work is also beneficial for many applications such as modeling the evolution of the direction of research for forecasting future trends of IT industry.

III. METHODOLOGY

Incremental clustering is used for automatic topic identification as it produces appropriate clusters dynamically from tweet streams. At each time stamp along the time line of a tweet stream the collection of tweets may be clustered differently as the topics covered in the tweets may vary with time. However certain topics may continue to be of interest during a portion of the time line marked by a successive time stamps, while new topics may also evolve during this period. Instead of clustering all the tweets again and again at each time stamp incremental clustering aims to refine the set of clusters representing the topics at successive time stamps along the time line. In other words, at each time stamp in the time line the incremental clustering refines the set of clusters discovered at the previous time stamp, except for the first time stamp in the time line wherein initial cluster formation of tweets already available happens. In the context of tweet clustering, the content alone may not accurately represent the topic, so hybrid clustering **[2]** proposed by the authors is used to discover high quality initial clusters. In addition to the content, the tweeting behaviour of the influential users representing the social context is also used in the hybrid clustering. A frame work is developed for the incremental clustering to cluster the tweets as they arrive based on content, social context and temporal features. The proposed framework of incremental clustering is divided into two modules.

Module 1: Initial cluster formation for the tweets available at the first time stamp using hybrid clustering algorithm.

Module 2: Identification of cluster to each of the new tweet in the timeline based on the proposed incremental clustering algorithm; the set of clusters may change with time with possible additions and deletions to the existing set.

Module 1: Identification of initial clusters and refining the previously formed clusters:
The research considers only the re tweeted tweets of influential users. Corresponding to each re tweeted tweet the influential user who made it along with his weight w_i and the list of followers who re tweeted the tweet is recorded by continuously querying twitter search API. The tweets are pre-processed first for content analysis; content of each tweet is represented as a term vector which specifies the prominence of a word in a tweet in terms of TF-IDF scores, where TF is the term frequency and IDF is inverse document frequency. Thus tweets are converted from text format to numeric format. The tweet term vector $<t_i , w_i , tv_i >$ represents the 'i'th tweet, t_i, posted by user whose weight is w_i and tv_i is the term vector representing the tweet content.

The first module of the framework identifies initial clusters/sub-topics in a specific domain based on the content of the tweets posted by the influential users which will be further modified based on tweeting behaviour of the users. K-means clustering algorithm is applied on the content part of tweet term vectors to create initial clusters which may be interpreted as K-topics. Based on the rationale that influential users are prone to confine themselves to a subset of topics of a domain, topics with common influential users with correlated influence patterns are considered indiscernible and hence, mergeable. The content based clusters are further refined based on tweeting behaviour of users in order to identify distinct topics. Influence scores are calculated as the ratio of number of retweets obtained to the total number of tweets they posted on the topic for each influential user within the cluster. Spearman correlation of the ranked list of common users based on their influence scores is estimated for every pair of clusters having considerable number of common users. Highly correlated pairs of clusters are merged and the process is continued to cover all original clusters. The above process of hybrid clustering [2] of tweets generates high quality clusters for identification of distinct topics of a domain as it considers both the content and the tweeting behaviour of the influential users.

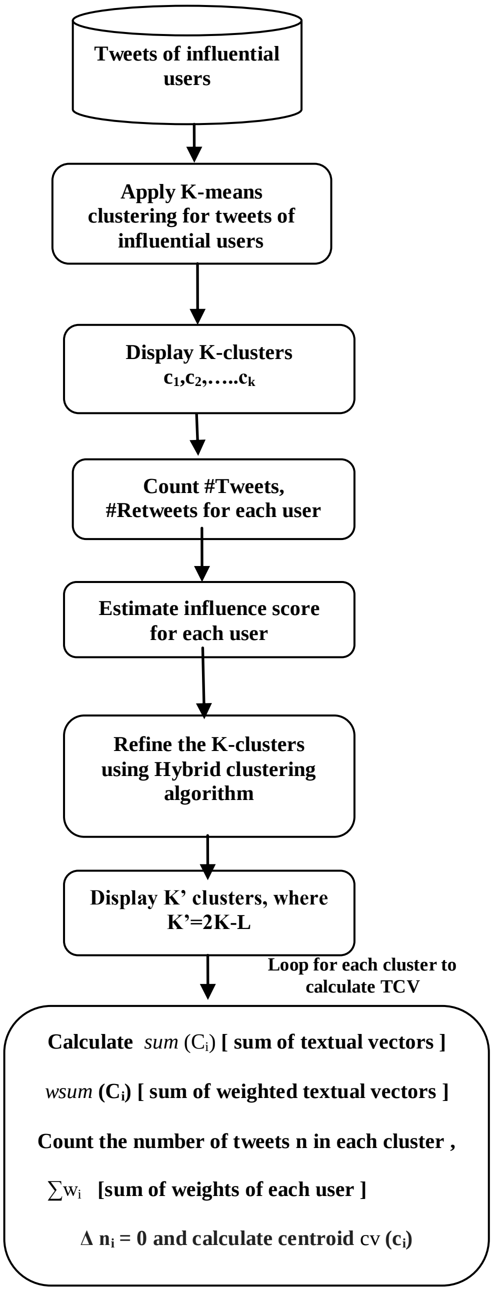
Tweets of influential users
Apply K-means clustering for tweets of influential users
Display K-clusters c1,c2,…..ck
Count #Tweets, #Retweets for each user
Estimate influence score for each user
Refine the K-clusters using Hybrid clustering algorithm
Display K' clusters, where K'=2K-L
Loop for each cluster to calculate TCV
Calculate sum (Ci) [sum of textual vectors]
wsum (Ci) [sum of weighted textual vectors]
Count the number of tweets n in each cluster ,
∑wi [sum of weights of each user]
Δ ni = 0 and calculate centroid cv (ci)

Figure 1: Trendy topic extraction based on the tweets of influential users

Incremental clustering of tweet stream requires to maintain statistics of each cluster identified at successive timestamps. A new data structure named as Tweet Cluster Vector (TCV) **[3]** is used in this paper to store the information about each cluster in each timestamp. For a cluster C containing the tweets $t_1, t_2, \ldots\ldots.t_n$ its tweet cluster vector TCV is defined as the following tuple:

TCV(C) = (*sum*, *wsum*, ∑w, n, cv, ftset, m) where

$sum = \sum \frac{tvi}{|tvi|}$ is the sum of normalized textual vectors

$wsum = \sum_{i=1}^{n} wi * tvi$ Represents sum of weighted textual vectors

∑w is the sum of weights of each user in the corresponding cluster

n is the number of tweets in the corresponding cluster

ftset is the focus set of size m, consisting of m tweets closest to the cluster centroid

cv is the centroid of each cluster which is calculated as wsum/n.

Module 2: Incremental clustering of tweet stream:

The second module of the framework deals with dynamically identifying the current trending subtopics on twitter and removes the outdated subtopics and updates the TCV's of clusters as new tweets arrive along the time line into various clusters. The second module implements incremental clustering methodology proposed by the authors for clustering the stream of tweets as they arrive.

It consists of the following two steps:

a) Allot clusters to each new tweet in time line:

For most events such as football matches, cricket matches, etc., in tweet streams timeliness is important because those events do not last for a long time. These events are modelled as subtopics of the domain 'sports' in this methodology. By considering a stream of domain specific tweets posted by the active users, our goal is to extract trendy topics dynamically from a tweet stream. Suppose a tweet t arrives in time t_s, and there are K active clusters identified at the previous time stamp, our goal is to check whether the newly posted tweet belongs to one of the existing K clusters which were identified at the previous time stamp or about a new event which is just started. The question is to decide whether to absorb t in one of the current clusters or to upgrade t as a new cluster. Each cluster is represented by its centroid estimated from its TCV. The cluster whose centroid is closest to t is found based on the cosine similarity of t to various centroids and mark the cluster with the largest similarity, Maxsim (t), as Cp. Even though Cp is closest to t it does not mean that new tweet t naturally belongs to Cp. The reason is that t may still be very distant from Cp. In such a case, new cluster should be created. The decision of whether to create a new cluster can be made using the principle of Minimum Bounding Similarity (MBS)**[3]**. The average closeness between the centroid to the tweets in the cluster Cp is denoted by Avg_Sim Cp. MBS is defined as β * Avg_Sim Cp where β is a bounding factor between 0 and 1. MBS is used to decide whether t is close enough to Cp: if Maxsim(t) is smaller than it, then t is upgraded to a new cluster; otherwise t is inserted into Cp. This process is repeated whenever a new tweet arrives. Thus after applying incremental clustering on the next time stamp along the timeline the clusters thus obtained are K+P clusters, where K is the number of clusters at the previous time stamp and P is the newly formed clusters during incremental clustering.

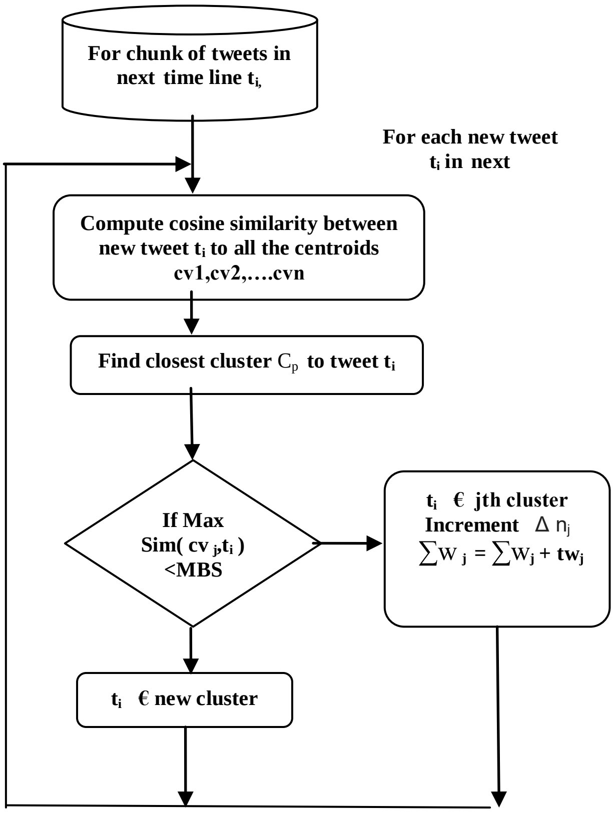
For chunk of tweets in next time line t_i,
For each new tweet t_i in next
Compute cosine similarity between new tweet t_i to all the centroids cv1,cv2,….cvn
Find closest cluster C_p to tweet t_i
If Max Sim(cv $_j$,t_i) <MBS
t_i € jth cluster
Increment Δn_j
$\sum w_j = \sum w_j + tw_j$
t_i € new cluster

Figure 2: Cluster allotment to each new tweet in next timeline

b) Identifying the clusters associated with active and new topics:

Events related to scientific break-through, football match, etc arouse temporary interest among people and those sub topics are active only for a limited time period. Therefore it is safe to delete the clusters representing the sub-topics when they are rarely discussed and it is better to maintain top K active clusters as growing clusters. Once the cluster allotment to each new tweet in the chunk of tweets belonging to the next time stamp is done, the newness of the clusters is estimated to classify the clusters as growing and outdated clusters which represent the active sub topics and outdated sub topics respectively. Mark the newness of newly formed cluster as 1.0.

In order to estimate the newness of an existing cluster, it is necessary to find the number of tweets added into existing cluster during the incremental clustering process denoted by Δn for each existing cluster. It is safe to delete the clusters representing the sub-topic whose Δn is equal to 0 and name such type of clusters as outdated clusters. Then calculate newness for each of the remaining clusters using the formula $\Delta n/n$ where n is the number of tweets originally existed in the cluster at the previous time stamp.

Every newly formed cluster becomes an new cluster for the next time stamp. Among the new clusters, based on the recent activity level some of them may transform into growing clusters and the others may get outdated due to lack of new tweets during the next time stamp based on two parameters α and β respectively. If the number of tweets newly arrived into an new cluster is less than β-minimum activity threshold defined as β times the proportionate share of a cluster from the new tweets arrived during this time period (calculated as the total number of tweets added in the latest time interval divided by the number of growing clusters, K) then the new cluster is considered as outdated and hence removed. Among the new clusters with more than β-minimum activity threshold if the number of tweets newly arrived into an new cluster is less than α-minimum activity threshold defined as α times the proportionate share of a cluster then it continues to be an new cluster in the next time stamp. If the number of tweets added to the cluster is more than the α-minimum activity threshold it is transformed into a growing cluster. It may be noted that $\beta < \alpha$.

By arranging the growing clusters based on decreasing value of its newness top K active clusters are recognized and maintained at the specific time stamp. Every newly formed cluster becomes new cluster for the next time stamp.

Outdated clusters are to be maintained for at least one more time stamps after which they are deleted unless they become α-minimum active again. Each of the Growing cluster represents a trending topic while a New cluster represents an evolving topic. An Outdated cluster corresponds to a topic which is no longer interesting.

IV. DATA SET AND EXPERIMENTATION

A data set of one lakh retweeted tweets on a specific domain namely **sports** were collected by continuously querying the Twitter search API. From this collection of domain specific tweets those tweets which were responded in the form of retweets were considered as the task relevant data for our experimentation. Eighty thousand tweets were thus obtained along a time line. One-fifth of the tweets posted in the first stage of the time line are pre processed and clustered based on content and social context of the tweets using hybrid clustering algorithm proposed by the authors [2]. The remaining part of timeline is divided into 4 intervals delimited by time stamps. Incremental clustering algorithm is applied for each time interval.

While processing the tweets in the first stage of the timeline, 11 distinct topics are identified based on content and social context of the tweets using hybrid clustering algorithm. Remaining part of the tweets in the timeline is divided into 4 intervals. Incremental clustering algorithm is then applied for each time interval (each interval contains 350 tweets). The tweets in each interval are processed in order to identify the topics which are growing (topics which are trending), which are outdated (the topics which may be closed) and newly evolving topics (the topics which are recently started) at each time stamp. The tweets in the first time interval are processed identification of clusters to each new tweet in the first interval is done to check whether the tweet is related to an existing topic or a new topic by using the MaxSim formula. Based on the number of tweets newly added to the existing clusters (Δn value), they may be labelled as growing clusters or outdated clusters. Clusters whose Δn value is equal to zero are marked as outdated, otherwise they are considered as growing clusters. In addition to that while processing the tweets in a time interval some new clusters are also formed.

Similarly the tweets in the subsequent time intervals are processed, cluster allotment to each new tweet is done and Δn value is also checked for each stage. New clusters which are formed during processing the tweets in a next time interval are considered as new clusters. Some new clusters which are formed in the previous time interval will be tested in this part and classified as growing, or outdated, or remains as new by using α-minimum activity and β-minimum activity levels as shown in the **Figure 3.**

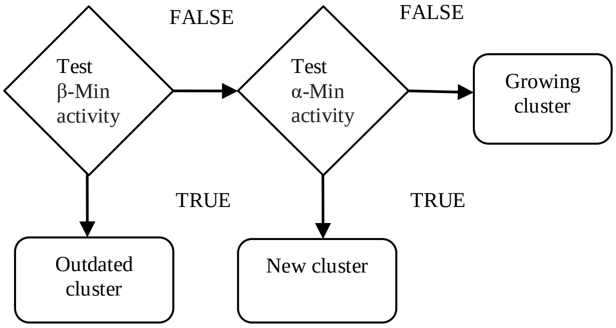

Figure 3: Procedure for every new cluster in previous time

For every new cluster, if the number of tweets newly arrived into the new cluster is less than β-minimum activity threshold then it is considered as outdated. Otherwise the new cluster is tested with α-Min activity level to check whether the new cluster is transformed into a growing cluster or not.

Results Analysis :

The number of Growing, New and Outdated clusters formed during successive time periods are displayed in the following Tables 1, 2, and 3 for different threshold values for α and β. The observed number of growing clusters, new clusters and outdated clusters at each time stamp are displayed in the below graph as shown in the Figures 4, 5, and 6. Relative distribution of growing, outdated and new clusters depends on the topic evolution during the study period (the time of collecting the tweets).

It may be observed from the results that higher values of β supports elimination of more number of new clusters as outdated clusters unless they receive sufficient number of tweets during the next time interval and hence focuses on lesser number of evolving topics. Similarly higher values of α screen newly formed clusters from becoming Growing clusters more stringently so that identification of trendy topics is more selective. Thus the proposed methodology provides programmable selection / screening of interesting topics streaming on the Tweeter dynamically through proper parameter setting.

1. The observed number of growing clusters, new clusters and outdated clusters at each time stamp with **β =0.1 and α=0.5**

Table 1: #growing topics, # outdated topics, #new topics at each time stamp by taking β = 0.1and α=0.5

β=0.1and α=0.5	#clusters	#growing	#outdated	#new
Initial	13			
1st time stamp		11	2	28
2nd time stamp		27	6	26
3rd time stamp		43	10	17
4th time stamp		50	12	20

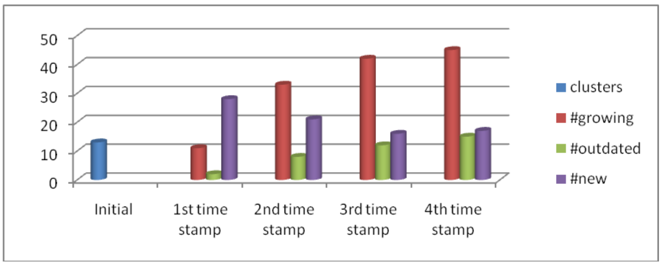

Figure 4: # growing topics, # outdated topics, #new topics at each time stamp with β = 0.1and α=0.5.

2. The observed number of growing clusters, new clusters and outdated clusters at each time stamp with β =0.2 and α=0.5

Table 2:# growing topics, # outdated topics, #new topics at each time stamp by taking β =0.2 and α=0.5

β=0.2and α=0.5	#clusters	#growing	#Outdated	#new
Initial	13			
1st time stamp		11	2	28
2nd time stamp		30	8	21
3rd time stamp		44	10	16
4th time stamp		51	12	19

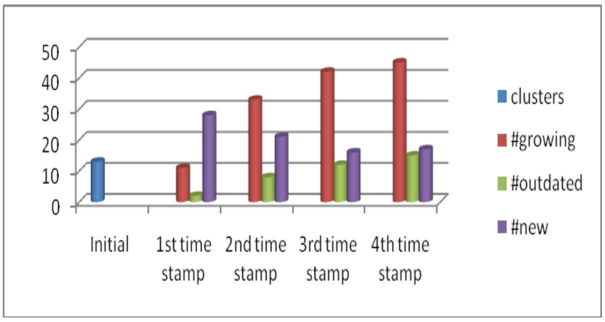

Figure 5: # growing topics, # outdated topics, #new topics at each time stamp with β = 0.2and α=0.5.

3. The observed number of growing clusters, new clusters and outdated clusters at each time stamp with β =0.3 and α=0.75

Table3:# growing topics, # outdated topics, #new topics at each time stamp by taking β =0.3 and α=0.75

β =0.3and α=0.75	#clusters	#growing	#Outdated	#new
Initial	13			
1st time stamp		11	2	28
2nd time stamp		33	8	21
3rd time stamp		42	12	16
4th time stamp		45	15	17

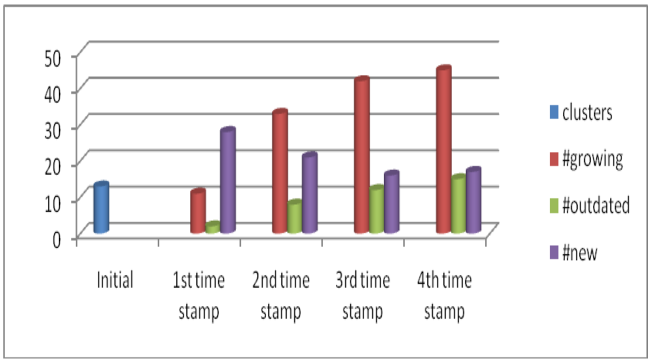

Figure 6: # growing topics, # outdated topics, #new topics at each time stamp with β = 0.3 and α=0.75.

V. CONCLUSION AND FUTURE SCOPE

An incremental framework for Automatic Topic identification as they evolve in a social media is developed. Tweets are clustered by considering the tweet features related to three aspects namely content, social context, as well as timeliness. Trendy topics in a specific time period are identified and it is useful for target marketing, recommender systems, and other e-commerce applications. In this paper the authors proposed an incremental clustering approach to identify the latest trendy topics by dynamically updating the clusters as they receive new tweets along the timeline. Top K trendy topics are identified based on the proposed newness measure.

REFERENCES

[1] D.Suneetha ., M. Shashi , "*Discovering trendy topics and their influence patterns relating users of social media*", Journal of advanced research in dynamical & control systems JARDCS, Vol **11**, No.**2 2019**, ISSN: **1943-023X.**

[2] D.Suneetha ., M. Shashi, "*Hybrid clustering for identification of distinct topics of a domain using user influence pattern*", International journal of innovative technology and exploring engineering IJITEE, ISSN: **2278-3075**, Vol **8** Issue:**2S2**, December **2018**.

[3] Zhenhua Wang., Lidan Shou, Ke Chen., Gang Chen, and Sharad Mehrotra, "*On summarization and time line generation for evolutionary tweet streams*", IEEE Transactions on knowledge and data engineering,Vol **27** No.**5**, My **2015**

[4] Lei Tang ., and Huan Liu ," *Leveraging social media networks for classification*",Springer, Data mining and knowledge discovery , **DOI 10.1007/s 10618-010 210-X**

[5] Xufei Wang ., Lei Tang ., Huan Liu., and Lei Wang , "*Learning with multi resolution overlapping communities*"Springer Knowledge information Systems , DOI 10.1007/s 10115-012-05550.

[6] Xia Hu., Lei Tang., Jilang Tang., and Huan Liu, "*Exploiting social relations for sentiment analysis in micro blogging*", ACM,**2013**,acm **978-1-4503-1869-3/13/02**.

[7] Yi-Chen.Lo., Jho-Yin-Li ., Mi-Yenyeh ., shou-de Lin ., and Jian Pei, "*What distinguishes one from its peers in social networks ? Data mining and knowledge discovery*" , **2013**, vol**(27)**, **396-420**, **DOI 10.1007/s 10618-013-0330.I.**

[8] Macro Pennacchiotti ., and Ana-Maria popascu , "*A machine Learning approach to twitter user classification*", Proceedings of the Fifth international conference on weblogs and social media ,**2013**.

[9] Jiliang Tang ., and Huan . Liu , "*Unsupervised feature selection for linked social media data*" , Knowledge discovery in data bases KDD, **2012**, ACM **978-1-4503-1462-6/12/08**.

[10] Jilang Tang ., Hujji Gao ., and Huan Liu , mtrust : "*Discerning multifaceted trust in a connected world*", WSDM,**2012**, ACM978 **1 -4503-0747-5/12/02**.

[11] Volkova.S , Twitter data collection : "*Crawling users ,neighbours and their communication for personal attribute prediction in social media*",**2014**.

[12] Volkova.S ., Coppersmith.G ., and Van Dume . B , "*Inferring user political preferences from streaming communication*"*s* , in Proceedings of the association for computational linguistics (ACL).**2014**

[13] Zamal ,F.A., Liu.W ., and Ruths .D, Homophily and "*latent attribute inference inferring latent attributes of twitter users from neighbours*" , In proceedings of international AAAI Conference on weblogs and social media,**387-390**.

[14] Volkova .S .,Wilson .Theresa ., and David .Y , "*Exploring demographic language variations to improve multi lingual sentiment analysis in social media*" ,Proceedings of the 2013 conference on empirical methods in natural language processing,**2013**,**1815-1827**.

[15] Yusuf Perwej, " The Hadoop Security in Big Data: A Technological View point and analysis", International Journal of Scientific Research in Computer Science and Engineering, Vol **17**, Issue **3**, pp:1-**4**, **(2019)**.

[16]. J.A. Alkrimi, Sh A Toma, R.S.Mohammed, C.E.George, "Using Knowledge Discovery to Enhance Classification Techniques for Detect Malaria-Infected Red Blood Cells, IJSRNSC, Vol-**8**,Issue -**1. (2020)**

Authors Profile

D.Suneetha received M.Tech degree in Computer Science and Technology from Gitam College of Engineering Visakhapatnamin in 2007. She is persuing Ph.D in JNTUK, Kakinada. She has published many technichal research papers in various international journals, presently she is working as an Assistant Professor in department of Computer Science and Engineering at GITAM deemed to be university, Visakhapatnam, Andhra pradesh, India. Her areas of research intrests include Data Mining Deep Learning, Machine Learning, and Pattern Recognition.

Prof. M Shashi received her B.E. in Electrical and Electronics and M.E. in Computer Engineering with distinction from Andhra University. She received a Ph.D. in 1994 from Andhra University and received the best Ph.D. thesis award. She is a professor in the Department of Computer Science and Systems Engineering, Andhra University, Andhra Pradesh, India. Prof. Shashi was a recipient of the AICTE career award as a young teacher in 1996 and also received the Andhra Pradesh State award as the Best Teacher for Engineering stream in 2016. She is the coordinator for the Center for Data Analytics, Andhra University sponsored by ISEA Project phase II, Ministry of Electronics and Information Technology (MeitY), India. She recently completed three consultancy projects on Deep learning in NLP domain for a Japanese Software Company, Exa Wizards, TOKYO, JAPAN. Her research interests include Data Mining, Artificial intelligence, Pattern Recognition, and Machine Learning. She is a member of the Computational Intelligence group of IEEE, a life member of ISTE, CSI, and a fellow member of the Institute of Engineers (India).