A Data Science Approach to Social Media Analytics for Discovering Topic-Specific Influentialpatterns from Tweets

Suneetha Dwarapu
Assistant Professor
Department of Computer science and Engineering GITAM Deemed University
Visakhapatnam, India

Shashi Mogalla, PhD
Professor
Department of Computer science and System Engineering
Andhra University
Visakhapatnam, India

ABSTRACT

Social media services like Twitter facilitate communication among people beyond geographical boundaries on the topics / events of their current interest which might differ from time to time. Topic detection from Twitter data would be helpful to find trendy topics / products and current events for target marketing in recommender systems. The information associated with each tweet should be analyzed in three different perspectivesinvolving textual content, social context and the temporal aspects to discover high quality clusters representing the topics and accordingly generate topic-specific influential patterns relating the users. This paper proposes a new framework for discovering topic specific influential patterns and maintaining them as snapshots along the time line. It involves topic identification through tweet clustering based on the textual content, followed by their refinement based on social context that involves informationinherent in the tweeting behavior of users in addition to thetiming of the tweet. Finally the framework proceeds to find the topic-specific influential patterns relating the users interested in a common topic. Dynamically finding patterns of influence enables tracking of information diffusion from influential users to their potential followers.

Keywords

Textual content, social context, Tweeting behaviour, Influential patterns

1. INTRODUCTION

Influential users on Twitter normally have many followers, and they have significant impact on their followers. Their tweets usually contain valuable information and news for others. The followers retweet the influential user's tweets on topics they are interested in because they think that it might be useful for others to know. By tweeting and retweeting people are actually passing information to others at a rapid speed when compared to traditional media services such as newspapers etc., This is one of the ways the product information is spread through the customers nowadays replacing the traditional "word-by-mouth" spread. However, it is also important to take into consideration the large volume of the data produced in social media. In order to transmit the right kind of information to the right kind of users at right time, business executives apply machine learning techniques forsocial media analytics.

The early detection of topics/happenings on social media websites and keeping track of their evolution facilitates business or administrative decision-makers to get additional time for making informed decisions. In order to keep pace with the changing dynamics of a topic, it is essential to identify users who are central and influential by posting regularly on related topics with a good response received from the users of the Social Network. Perhaps following them could be helpful in two ways. In addition to acquiring quality information related to topics or products or events along with the opinions and feedback from the influential users, the business companies may also involve the influential users for rapid information diffusion through their connectivity and following in the network. For example, positive feedback from an influential user on a newly introduced product of a company would be helpful for its sales promotion.

Generally influential users confine to posting tweets ontopics of a domain of their interest and their intensity of postings reflects their authority on the respective topics as wellas their following, which may vary with time. People generally follow multiple influential users for decision making and they would be influenced by different influential persons dependingon the topic or the product to be purchased. Thus the influence flows with respect to the topic as well as the influencing user and such topic-specific influence patterns change with time. In other words, the patterns of influence are determined in terms of a topic, influential person and time. Discovering trendy topics from the tweets made by influential users is an importanttask in social recommender systems in order to use them asthe basis for identifying potential customers for their products and services.

This paper proposes a new framework for discovering topic specific influential patterns and maintaining them as snapshots along the time line. It involves topic identification through tweet clustering based on the textual content, followed by their refinement based on social context that involves information inherent in the tweeting behavior of users in addition tothe timing of the tweet. Finally the framework proceeds tofind the topic-specific influential patterns relating the users interested in a common topic. A new methodology is proposedfor finding the topic-specific influence patterns using bipartite graphs to represent the influential users and their followers as nodes in it. On each of the topics identified, the influence scores are obtained implicitly as the ratio of the number of retweets made by the followers on the tweets related to the trendy topic to the numbertweets made by the influential user on the topic. Influence scores are filtered based on a threshold to identify the potential followers of an influential user on a topic.

The second aspect is based on the natural expectation that the influential users have authority on a selected subset of topics of their current interest and hence, more often than not, are confined to posting tweets on those topics of the domain only. This rationale has driven the work towards adopting entropy which is originally an information theoretic metric forassessing the distinctness of topics represented by different clusters. A new metric named Cumulative Mean Entropy is devised to estimate the randomness associated with the number of tweets made by the influential users on different topics. Comparatively lower values of the cumulative mean entropy of a clustering solution suggests better quality of clusters in representing distinct topics. It is observed that the tweeting behavior of the influential users is more or less confined to a specific subset of topics rather than wondering randomly on all topics. The information theoretic metric for quantification of uncertainty known as Entropy is adapted to define cumulative mean entropy to assess the distinctness of the topics which is used to estimate the quality of clusters generated by the hybrid clustering algorithm.

The framework is useful to identify the potential followersof each influential user on each topic and accordingly to recommend the items / services liked by the influential usersto the potential followers as they share similar interest and are possibly inspired by the influential user.

2. RELATED WORK

2.1 Clustering and Influential Users

Cha et. al. [1] performed an analysis on the impact of how the Twitter fraternity can stimulate the others. They have used different metrics to analyze the three different dimension of influence such as the retweets, mentions and in degree. Hurtado et. al [2] through their research identified the important topics from articles and also made a trend forecasting on that particular topic. They have used association rule mining and correlation coefficient to mine the frequent sets and identify the important topics.

2.2 Hybrid Clustering

Wang et. al. [4] developed a new method for retrieving the hierarchical relationships of users of social media platforms who exist in common communities which overlap. To develop this algorithm, they have used the multi-fold varied resolution approach which is very granular in nature. When applied on the real world's applications, their algorithm has shown promising results.

2.3 Incremental Clustering

Cai et al [3] using the operations such as the create, retrieve, divide and combine the events created on Twitterand other social media platforms the event monitoring and tracking is done. Based on the threshold values of this MIL events are monitored and tracked. Their proposed methods have shown better performance than the existing methods.

2.4 Other related papers

Wang et. al. [27] developed a new method for retrieving the hierarchical relationships of users of social media platforms who exist in common communities which overlap. To develop this algorithm, they have used the multi-fold varied resolution approach which is very granular in nature. When applied on the real world's applications, their algorithm has shown promising results. Zhao et. al. [28] proposed connecting social media application with e-commerce applications to acquire the user micro blogging information and use it for generation of product recommendations. They have used Chinese micro blogging site SINA WEIBO and e-commerce site JINGDONG for their experimentations and to the effectiveness of the suggested framework. Cai et al [29] using the operations such as the create, retrieve, divide and combine the events created on Twitter and other social media platforms the event monitoring and tracking is done. Based on the threshold values of this MIL events are monitored and tracked. Their proposed methods have shown better performance than the existing methods. Cordeiro et al [30] have carried out a literature review on the existing published research on the different algorithms, frameworks, techniques on the social media platform and Social Network Analysis. They have presented different challenges and the scope for further more research. They indicate that the research must be more focussed on developing better and efficient methods, algorithms and models for social network analysis.

2.5 Disadvantages of the previous works

Content of a tweet conveys the intention of the agent in textual form and is generally used for topic identification.However, the length of the tweet being very short, and also due to conversational nature of the tweets, textual content by itself may not identify the correct topic.

2.6 Problem statement based on previous works

The information associated with a tweet is consideredin three perspectives: content, social context and the timeliness. Content of a tweet conveys the intention ofthe agent in textual form and is generally used for topic identification. However, the length of the tweet being very short, and also due to conversational nature of the tweets, textual content by itself may not identify the correct topic.This research is motivated to use the other two types of information related to tweets also for topic identification. Social context refers to the agent who posted the tweet, andhis general interests as well as the impact it had created among the followers. Accordingly tweets clustered based on the content have to be refined based on social context of the tweets for identifying distinct topics of a domain. In orderto capture the changing dynamics of the social media, tweet stream should be processed to make incremental updates to the topics identified with a provision for adding / discovering new topics online.

3. METHODOLOGY

Popular users with high in-degree estimated in terms of the huge number of followers may not be really influential on all topics [1]. Based on their domain of interest, such users can hold significant influence over a limited set of related topics. Most of the user's in social media are influenced by different influential persons for different topics. For example, a person may be influenced by a movie star for deciding his dressselection for an occasion, while he may be influenced by a professor to decide on a textbook to be bought and similarlybe influenced by an inspiring colleague for decisions on healthinsurance policy.

The information associated with a tweet is considered in three perspectives: content, social context and the timeliness. Content of a tweet conveys the intention of the agent in textualform and is generally used for topic identification. However, the length of the tweet being very short, and also due toconversational nature of the tweets, textual content by itself may not identify the correct topic. This framework is motivatedto use the other two types of information related to tweets also for topic identification. Social context refers to the agent who posted the tweet, and his general interests as well asthe impact it had created among the followers. Accordingly tweets clustered based on the content have to be refined based on social context of the tweets for identifying distinct topicsof a domain. In order to capture the changing dynamics ofthe social media, tweet stream should be processed to make incremental updates to the topics identified with a provision for adding / discovering new topics online.

The proposed framework aims to identify trendy topics from tweet stream and accordingly discover topic-specific influence patterns relating the influential users and their followers. This framework integrates the results of tweet analysis throughthree different perspectives namely content, social context and the timeliness of the tweets in cascade. The functionality of the proposed framework is divided into five modules:

1. Estimates the weights of active users and identifies influential users based on the retweets received for their tweets. It collects the tweets made by the influential users of a domainand represents the tweets in the form tweet vectors.

2. Applies content based tweet clustering to identify the group of tweets that share a common topic.

3. Describes the refinement of content based tweet clusters based on social context of the tweets. Tweeting behavior of influential users is captured using Spearman Rank Correlation coefficient and used for hybrid clustering to obtain distinct topics of a domain.

4. Describes the formation of Topic–Specific Influence Pat- terns with numeric scores indicating the strength of the relation between the influential users with their followers based on the statistics estimated on the tweets belonging to the cluster specific to the topic.

5. Describes the incremental clustering of tweets received from the tweet stream in a specific time period to identifynew trendy topics and to refresh the existing topics. Thismodule represents the clusters as cluster vectors so that they are suitable for incremental clustering.

A. *Identification of Influential Users and representation oftheir tweets*

The first step in the methodology is to recognize influential users of a given domain, since they play a significant rolein spreading the data on Twitter by tweeting their opinionsthat influence other users so that the posted tweets are often retweeted by others [6]. The authors consider only those tweets posted by the influential user which acquire retweets and dispose the tweets with no retweets since the impact of such kind of tweets is not evident. Twitter API is used to obtain the retweeted tweets of influential users. Influential users are identified by two level filtering.

In the first level, influential users are identified disregardingthe topic. If a user gets at least two retweets, he is considered as an active user and such active users are assigned a weightage proportionate to the number of retweets produced for his tweets on a logarithmic scale. The weight of a user,j, is denoted by $w_j = log_2(\#RT_j)$ where $\#RT_j$ denotes the number of retweets produced for the tweets made by user, j. However, this weight evaluates the impact of an active user on his followers irrespective of the topic. These weight estimates are also used to represent the tweets made by the users in vector format as described in the next section.

B. *Represent the Tweets in the form Tweet Vectors and applies Content-based Tweet Clustering to identify the group of Tweets that share a Common Topic*

The author proposes to adopt document clustering approaches used in text mining for clustering the tweets that share common topics. Each tweet in the text format is converted into a tweet vector format of fixed length asdetailed below.

C. *Vector Representation of a tweet*

The content of the i^{th} tweet is represented as a two-tuple ***vector format***< tv_i, w_i > where tv_i is the fixed length vector of TF-IDF scores of vocabulary words contained in the tweet, and w_i isthe weight of the user who posted the tweet.

The architectural diagram of the proposed framework is as shown below:

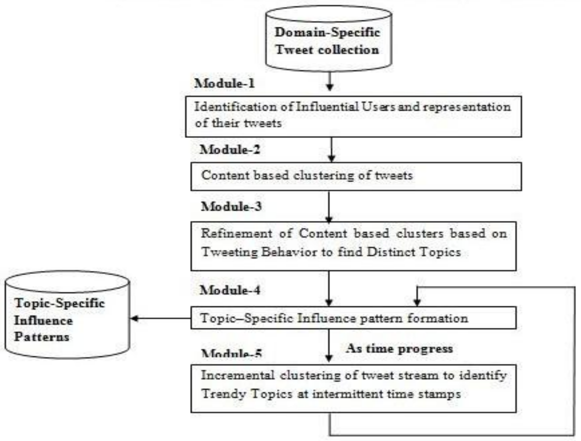
Domain-Specific Tweet collection
Module-1
Identification of Influential Users and representation of their tweets
Module-2
Content based clustering of tweets
Module-3
Refinement of Content based clusters based on Tweeting Behavior to find Distinct Topics
Module-4
Topic-Specific Influence Patterns
Topic–Specific Influence pattern formation
Module-5
As time progress
Incremental clustering of tweet stream to identify Trendy Topics at intermittent time stamps

Figure 1: Comprehensive framework for Topic-specific Influence patterns in Social media

The tweets posted by the influential users of a domain are transformed into tweet vector format, and K-means clustering is applied on the tweet vectors to produce K-clusters representing the topics on which influential users expressed their opinions during the given time period. The optimal K value is found after experimenting with different values of K in the range of 2 to 24 to minimize the SSE

D. *Refinement of Content based Tweet Clusters based on Social Context*

Clustering of tweets is extensively used for automatic topic identification [10]. Majority of the methods for clustering of tweets depends on content of the tweets. But the contentof the tweet may not exactly represent the topic due to the limitations on the length of the tweets and also due to the usage of non standard short forms of words and clauses to represent context sensitive stereotypical responses where the meaning is indicative. In order to accurately identify topics, some more information like tweeting behavior of users is also used with the content of the tweets for improving the qualityof clustering.

In a broad social network such as Twitter, an enormous number of tweets belonging to different domains are posted by a large number of users often dominated by a small number ofinfluential users. However, a specific influential user is inclined to limit to tweet on a subset of distinct topics of a domain. Despite influential users post tweets on various topics, their level of influence on different topics varies unless the topics are somewhat similar. Tweet clusters associated with clearly distinct topics share comparatively less number of common influential users that too with uncorrelated ranking among them for different topics. This consideration is referred to as“***tweeting behavior of influential users*** **[7]”**.

This module presents a new hybrid clustering methodology for distinct topic identification by leveraging information inherent in the tweeting behavior of users in addition to the content of the tweet. Spearmen Rank Correlation metric is adopted[7] for identifying the mergeable clusters based on the tweeting behavior of influential user.

E. Identification of distinct topics based on Tweeting Behavior.

In order to extract the topic-specific tweeting behavior of influential users, it was observed that the clusters representing the similar topics often receive tweets from common influential users. According to the natural intuition an influential user, being an authority on a specific set of topics, confines himself /herself to posting regularly on those topics and his topic-specific level of influence will be stable relative to the other users. Since similar topics have common influence users, the correlated topic-specific influence ranking of common users is indicative of the similarity of topics represented by the content-based clusters which, may possibly, represent non-distinct topics. Based on this rationale, this paper suggests that the tweeting behavior of influential user is discovered in terms of correlation intopic-specific influence score ranking of common users contributing to the topics. Specifically, two clusters are said to be highly similar when the relative ranking of the common influential users of the two clusters positively correlate with each other based on their topic-specific influence scores inthe respective clusters. Hence such clusters are mergeable in the process of identifying the distinct topics of a domain and the result of merging should be propagated. In order to estimate similarity between a pair of clusters, the authors proposed [7] a new metric for similarity calculation called as Similarity Score(SS) based on the tweeting behavior of the influential users. Similarity Score is defined as the product of Jaccard Similarity Coefficient and Spearmen Rank Correlation between the pair of clusters.

Similarity score SS(C_i, C_j)=JC(C$_i$, C_j)*$SR(C_i, C_j)$ ...(1)

Where JC(Ci,Cj) is the “Jaccard Similarity Coefficient”, that measures the overlap between two clusters/topics Ci and Cj.

F. Identification of Topic–Specific Influence Patterns

Once the distinct topics of a domain are identified, this module processes each cluster (topic) to generate topic- specific influence patterns among the users of a domain. This module estimates numeric scores indicating the strength of the relation between the influential users with their followers based on the statistics estimated on the tweets belonging the cluster specific to the topic. For each cluster of the domain, it maintains the list of influential users of the cluster; the number of tweets posted by each influential user, number of retweets made by their followers on the topic related to the cluster in the form a Bi-partite graph[6].

The influence score for each follower is estimated based onthe number of tweets posted by influential user IU, and the number of retweets made by the follower F on the specific topic. The potential followers are those followers havinghigher influence score than a pre-specified threshold value. The influence score of an influential user, IU, on each of his/ her follower, F, can be obtained using the formula given below:

$$Influence\ Scoret(IU, F\ ) = \frac{NumberoftimesF \quad retweeted}{\text{NumberoftweetspostedbyIU}} \quad (2)$$

Thus for a given topic, t, associated with a cluster, the influence scores are estimated, and the followers whose scores are higher than the threshold are declared as potential followers of the influential user for the specified topic[6].

G. Incremental Tweet Stream Clustering to Discover New Topics and Refresh Existing Topics

Timeliness is important for events such as cricket matches, COVID-19, US elections, etc., as they arouse temporary interests among users, and those topics are alive only for a limited time period. It is also important to monitor the upcoming topics by continuously analyzing the tweets made by the influential users dynamically in order to know and understand the changing trends about the topics. Such information about emerging events can be immensely valuable if it is discovered and made available timely in order to deliver advertisements to those users who are interested in the product/ service.

As time elapses, the tweet stream extends with new tweets received in tune with the changing dynamics of the trendy topics and new events. Hence, the framework for topic identification should also be capable of incremental update of the topics recognized in the previous time windows.

During incremental clustering of tweets it is essential to dynamically update the clusters as they get new tweets inserted based on their compatibility to the incoming tweets.In order to support efficient matching for compatibility check and insertion of appropriate tweets from the tweet stream, it is mandatory to maintain the essential statistics of each cluster identified at successive time windows. This module represents the clusters as cluster vectors so that they maintain essential information about a cluster for its incremental update. For a cluster C containing the tweets $t_1, t_2 ... t_n$, its tweet cluster vector TCV is defined as detailed below:

TCV(Each cluster C) = (Sum, wsum,w, n, m,ftset,cv).......(3)

Incremental clustering algorithm[8] will process the tweet stream one tweet after the other by matching the tweet vector with TCVs of existing clusters to identify the compatible cluster and inserts the tweet into the most compatible cluster,if found. If no one of the existing cluster TCVs are found to be compatible enough with the incoming tweet then a new cluster will be created to hold it. The goal of the fifth module of the comprehensive framework is to identify the current trending topics on Twitter and removal of outdated topicsafter updating TCV’s in accordance with the new chunk of tweets arrived during the current time window along the time line. The next section presents the details of the incremental clustering methodology for clustering stream of tweets ina given time window along the timeline. It consists of two phases namely cluster allotment phase and cluster labeling phase based on the recent activity.

Cluster Allotment phase: The aim of the incremental clustering is to decide whether the newly posted tweet t belongs to one of the existing clusters/topics or a newly created cluster/topic representing a new event just started.The query is to find out whether or not the new tweet t represented by its tweet vector tv in n-dimensional space will be inserted into an existing cluster or upgrade as new cluster. The clusters whose centroid is closest to t is observed primarily based on cosine similarity of tweet tv to various centroids, cv, and mark the cluster with largest similarity,Maxsim(t), as Cp. Cosine similarity between the new tweet and the centroid is calculated using the formula:

$$Cosine\ similarity(tv, cv) = \frac{\sum_{i=1}^{n} tv_i * cv_i}{\sqrt{\sum_{i=1}^{n} tv_i^2} * \sqrt{\sum_{i=1}^{n} cv_i^2}} \ldots.(4)$$

The cluster with the largest similarity value is marked as Cp. The principle of Minimum Bounding Similarity (MBS) isused to decide whether new tweet t is close enough to one of the existing clusters Cp or not. The Minimum Bounding Similarity (MBS) is defined for a cluster, Cp, as the productof $\gamma \quad sim(C_p)$, *where* γ is the bounding factor between 0 and 1,

$0 < \gamma < 1$ and $sim(C_p)$ is the average cosine similaritybetween the centroid of Cp and the tweets included in the cluster Cp.

Cluster labeling phase:

Once clusters are allotted to each new tweet in the specific time window along the timeline, the next step is to label the clusters into appropriate stage of their life cycle based onthe extent of tweeting activity in the present time window. Newness measure is devised by the author to quantify the extent of tweeting activity related to a cluster in the present window. Every newly formed cluster becomes new cluster for the next time window.

All new clusters are classified as infant stage indicatingthat they are possibly related to upcoming topics. The clusters with less than the minimal tweeting activity in the presenttime window are classified as outdated and discarded as they are obsolete and no longer discussed by the twitter users. A grace period of at least one time window is given for the clusters labeled as outdated to regain the growing cluster status before discarding them. The clusters with tweeting activity maintained from the previous time windows as well as in this time window are classified as growing and adultclusters and these clusters indicate the current trendy topics.

Among the new clusters based on recent activity level some of the new clusters may change the type from new clusters to growing clusters and others may get labeled as outdated because of insufficient number of tweets added into the new clusters during the next time window. The question is to decide whether the newly created cluster in the previous time window becomes growing cluster or an outdated cluster based on two parameters namely α and β respectively.

If the number of tweets newly arrived into a new cluster is less than β-minimum activity threshold then the new cluster is considered as outdated and hence removed. β-minimumactivity threshold is defined as β times the proportionate share of a growing cluster from the chunk of tweets arrived during this time window.

4.* DATA SETS USED AND EXPERIMENT RESULTSOBTAINED

A data set of one lakh tweets on the “sports” domain based on the related hash tag is gathered by repeatedly questioning the Twitter search application. Tweepy is used for data collection. From this mixture of domain-specific tweets, those tweets acknowledged with retweets are 21360 tweets; those tweets are considered as the task appropriate data for the research. The active users who posted those 21360 tweets are recognized with the help of Twitter API. The distinct active users identified are 11600. The weight of each active useris calculated, based on a threshold (value as 1.5); influential users are filtered in the first stage. Influential users thus obtained are 5400. K-means clustering is applied to identify the trendy topics of a domain, namely sports, based on the content of the tweets made by influential users.

Thus content-based tweet clustering generates 14 clusters/topics, on which influential users expressed their opinions in the specified time duration. A hybrid clustering algorithm [6] is applied to content based tweet clusters [5]based on the tweeting behavior of influential users. The similarity between pairs of topics is estimated, and similar clusters are identified and merged together as a single cluster. During merging process of similar clusters some new clusters may be formed. A variable L is used to monitor/index the clusters being generated by merging pairs of similar clusters.L is initialized with the value of K (existing number of clusters) and it is incremented whenever a new cluster is formed. The author proposes to use entropy as it has the potential to quantify the randomness or uncertainty in data distribution. Specifically, mean entropy is used to measure the quality of the clusters before and after merging. Better quality of clusters is indicated by lower values of entropy. Each of the 14 clusters contains tweets made by influential users on specific topics. The number of tweets in each cluster is shown in Table 1.

Table 1: Tweets in each Cluster by taking K value as 14

Cluster number	1	2	3	4	5	6	7	8	9	10	11	12	13	14
#tweets	997	2155	141	517	1060	526	265	728	262	1488	1105	656	788	989

Similarly tweets related to “politics” and “health” domains are collected and distinct topics are identified applying the methodology discussed above. The hybrid clustering solutions generated for the datasets belonging to the three domains are analyzed to assess the impact of merging the content based clusters using social context for distinct cluster formation.The line graph below depicts the percentage of reduction in cumulative mean entropy for the clustering solution in three different domains: sports, politics and health.

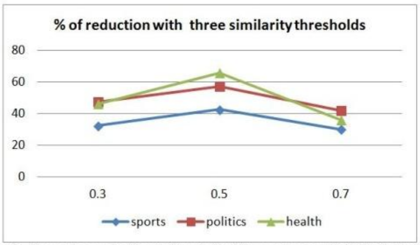

Figure 2: Percentage reduction in Cumulative Mean entropy with Hybrid clustering

It is observed that the best threshold value is 0.5. The proposed hybrid clustering methodology could generate better quality clusters that capture distinct topics of a given domain using the content of the tweet and the tweeting behavior of influential users. The cumulative mean entropy CME of clusters formed before and after merging and the percentage reduction in entropy for distinct clusters is shown in the Table 2 forthe best similarity threshold.

1. The value of cumulative mean entropy of cluster solution for “sports” domain before merging is 877 and after merging is 373, which indicates a clear reduction in mean entropy due to merging of clusters.

2. The value of cumulative mean entropy of cluster solution for “politics” domain before merging is 547 and after merging is 314 and it indicates reduced entropy after merging.

3. The value of cumulative mean entropy of cluster solution for “health” domain before merging is 636 and after merging is 419. In this domain also, the clusters finalized after merging some of the content based clusters as per the tweeting behavior of the influential users are better due to reduced uncertainty reflected in reduced entropy after merging.

Table 2: The of reduction in clustering solution, when the similarity threshold is set to 0.5

Domain	H(U) before	Ha(U) after	% reduction in CME
Sports	877	373	42.5
Politics	547	314	57.4
Health	636	419	65.8

Sample Results of Topic-specific Influence Patterns formation: The distinct clusters of a domain are further explored individually to extract the topic-specific influential patterns relating to the influential users and their followers based on the number of retweets made by them. The algorithm for module 4 is applied on the collection of tweets belonging to each cluster representing a specific topic; influence scores are estimated for every influential user-follower pair. Influence scores for a small set of users related to a topic represented by a cluster are shown in the Table3 given below:

Table 3: Topic-Specific Influence Score of each Follower to each User in one Cluster

Sl. no	User name	# Tweets posted	Follower name	# Retweets	Influence scores
1.	Wsls Scores(A)	23	BubbaTyl (D)	1	0.043478
			Rattilffmark (E)	20	0.869595
2.	11Alivenews(B)	16	Times land (F)	11	0.6875
			BubbaTyl (D)	4	0.25
3.	SEN news(C)	20	Hokiegunns(G)	10	0.5
			Jimrmp(H)	2	0.086957

Higher values of influence score indicate that the influential user has a stronger influence on the follower while making decisions on the issues related to the specific topic. The analysis is done based on the influence scores. Specific threshold value, 0.1 is considered to filter the influence patterns. The followers whose score is greater than the threshold are recognized as potential followers for a given influential user on a specific topic.

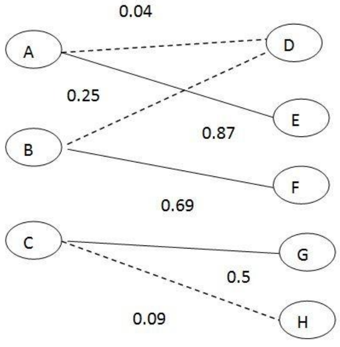

Figure 3 : Bi-partite graph representing Topic-specific Influence Patterns

Results of Incremental Clustering :

The remaining part of the tweet stream is divided into four intervals in the timeline, called as time windows. Each interval contains 350 tweets. An incremental clustering algorithm is then applied to each time window. After fixing α and β thresholds defined by Minimum Bounding Similarity, the tweets in each interval are processed in order to identify the topics which are growing (Trendy topics), the topics that are outdated (topics that may be closed), and newly evolving topics (topics that are recently started). The results of incremental clustering at successive time windows are shown in figure 4 which may vary for different α and β settings.

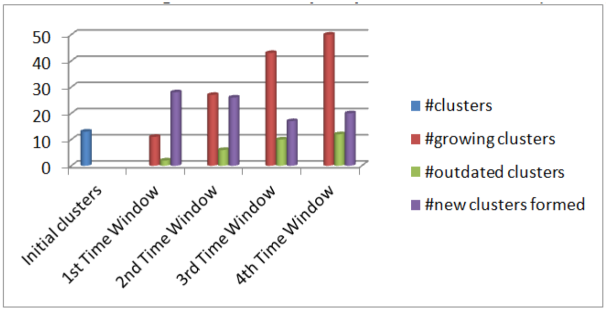

Figure 4: **#Growing topics, #New topics, #outdated topics at each time stamp by taking** β**=0.1, and** α**=0.5.**

Thus the proposed incremental clustering methodology provides programmable selection / screening of interesting topics streaming on the tweeter dynamically through proper parameter setting.

5. CONCLUSION AND FUTURE SCOPE:

A comprehensive framework is developed for dynamically identifying topic-specific influential patterns relating the users of micro blogging social networking platforms. The traditional approach to topic identification with text clustering algorithms that focus on content part of the tweets do not suffice as tweets involve three types of features namely content, social context and temporal patterns. Tweeting behavior of Influential Users reflecting the social context is captured using Spearman Rank Correlation coefficient and used for Hybrid Clustering to obtain Distinct Topics of a Domain. Incremental clustering is used to keep track of the changing dynamics in topic evolution. The framework is successfully developed synergizing various concepts of data science applicable to social media analytics.

The functionality of the framework is described in five modules for pre processing tweets followed by discovering and refining the topics based on textual content, social context and temporal aspects and a separate module for identifying topic-specific influence patterns.

6. FUTURE EXTENSION

Social media data is growing leaps and bound in the recent times and it is desirable to apply the recent developments in Big data analytics for collecting and processing the social media data at regular intervals using highly scalable procedures. An intensive study of regulations for accessing and collecting social media data without violating the privacy restrictions is called for by the research communities. The proposed framework has three different modules for processing the tweets based on textual content, social context and temporal aspects in cascade. Recent developments in deep learning architectures, in general, and specifically the applicability of hybridization of Recurrent and Convolution networks may be explored for comprehensive processing of the tweets.

Future plans to apply this frame work in real worldapplications

Formation of implicit groups of users that share common interests along with the trendy topics of the domain could be discovered by clustering the tweets posted by the influential users only. This observation helps to automatically filter out enormous volume of tweets for topic identification.

Significant research attention has been invested in modeling information diffusion in social networks with emphasis on investigating the roles of different users, topic-specific impact of influential users on their followers, identification of distinct topics of a specific domain and monitoring the evolution of trendy topics and their life cycle.

Contributions of the paper:

1. Formation of implicit groups of users that share common interests along with the trendy topics of the domain could be discovered by clustering the tweets posted by the influential users only. This observation helps to automatically filter out enormous volume of tweets for topic identification.

2. Tweets have three different types of information in the form of text content, social context, the timing and hence the author proposed to process tweets through all these perspectives for discovering high quality clusters representing distinct topics.

3. An integrated framework for processing the tweets posted by the influential users in each of these perspectives in cascade to achieve their integration is developed to discover and maintain trendy topics along the time line.

4. It was observed that the impact of influential users on their followers is topic specific. Users follow multiple influential users and consider them as the authorities on different topics and accordingly follow distinct influential users for making decisions related to distinct topics or events.

5. Spearman Rank correlation coefficient originally designed for identifying redundant metrics is used for a novel purpose of identifying non-distinct topics of a domain based on the social context of the tweets. The content based clusters having highly correlating relative ranking of influential users are considered non-distinct and hence merged to form distinct topics.

6. A hybrid clustering algorithm is developed to discover topics from the tweets based on their textual content and refine them based on the social context of the tweet.

7. It is observed that the tweeting behavior of the influential users is more or less confined to a specific subset of topics rather than wondering randomly on all topics. The information theoretic metric for quantification of uncertainty known as Entropy is adapted to define cumulative mean entropy to assess the distinctness of the topics which is used to estimate the quality of clusters generated by the hybrid clustering algorithm.

8. The framework facilitates discovery of new topics while keeping track of the changing dynamics of the trendy topics previously identified. The life cycle of the topics with differentstages named as infant cluster, growing cluster, outdated cluster is maintained by devising appropriate statistics for categorizing the topics into different stages of their life cycle. Newness measure is devised by the author to quantify the extent of tweeting activity related to a cluster in the present window.

9. Incremental clustering is performed by proposing a new methodology with an appropriate representation for tweet clusters as vectors suitable for incremental update of the clusters based on Minimum Bounding Similarity estimation proposed by the author.

10. A new methodology is proposed for finding the topic-specific influence patterns using bipartite graphs to represent the influential users and their followers as nodes in it. On each of the topics identified, the influence scores are obtained implicitly as the ratio of the number of retweets made bythe followers on the tweets related to the trendy topic to the total number of tweets made by the influential user on the topic. Influence scores are filtered based on a threshold to identify the potential followers of an influential user on a topic.

7. REFERENCES

[1] Cha, Meeyoung, Hamed Haddadi, Fabricio Benevenuto, and Krishna P. Gummadi. ”Measuring user influence in twitter: The million follower fallacy.” In fourth international AAAI conference on weblogs and social media. 2010.

[2] Hurtado, J.L., Agarwal, A. and Zhu, X., 2016. Topic discovery and future trend forecasting for texts. Journal ofBig Data, 3(1), p.7.

[3] Cai, Hongyun, Zi Huang, Divesh Srivastava, and Qing Zhang. ”Indexing evolving events from tweet streams.” IEEE Transactions on Knowledge and Data Engineering 27, no. 11 (2015): 3001-3015.

[4] Wang, Xufei, Lei Tang, Huan Liu, and Lei Wang. ”Learning with multi-resolution overlapping communities.” Knowledge and information systems 36, no. 2 (2013): 517-535.

[5] Wang, Chi, Jie Tang, Jimeng Sun, and Jiawei Han. ”Dynamic social influence analysis through time-dependent factor graphs.” In 2011 International Conference on Advances in Social Networks Analysis and Mining, pp. 239-246. IEEE, 2011.

[6] D.Suneetha, M.Shashi, “Discovering Trendy Topics and their Influence Patterns Relating Users of Social Media”, Journal of Advanced Research and dynamic Communication Systems JARDCS, Vol.11,No.2,ISSN :1943-023X, 2019.

[7] Dwarapu Suneetha, Mogalla Shashi, “Hybrid Clusteringfor Identification of Distinct Topics of a Domain Using User Influence Pattern”, International Journal of Innovative Technology and Exploring Engineering IJITEE, Vol.8,Issue- 2S2, ISSN: 2278-3075,2018.

[8] D.Suneetha, M.Shashi, “Keeping Track of evolution of Trendy Topics in Social Media”, International journal of computer sciences and engineering IJCSE, Vol.8, Issue.5, ISSN:2347-2693, May 2020.

[9] Culotta, Aron. ”Training a text classifier with a single word using Twitter Lists and domain adaptation.” Social Network Analysis and Mining 6.1 (2016): 8.

[10] Wang, Z., Shou, L., Chen, K., Chen, G. and Mehrotra, S., 2014. On summarization and timeline generation for evolutionary tweet streams. IEEE Transactions on Knowledgeand Data Engineering, 27(5), pp.1301-1315.

[11] Lo, Yi-Chen, et al. ”What distinguish one from its peers in social networks?.” Data mining and knowledge discovery 27.3 (2013): 396-420.

[12] Tang, Lei, Huan Liu, and Jianping Zhang. ”Identifying evolving groups in dynamic multimode networks.” IEEE Transactions on Knowledge and Data Engineering 24.1 (2011): 72-85.

[13] Tang, Jiliang, Huiji Gao, and Huan Liu. ”mTrust: discerning multi-faceted trust in a connected world.” Proceedings of the fifth ACM international conference on Web search and data mining. 2012.

[14] Volkova, Svitlana, Theresa Wilson, and David Yarowsky. ”Exploring demographic language variations to improve multilingual sentiment analysis in social media.” In Proceedings of the 2013 Conference on Empirical Methodsin Natural Language Processing, pp. 1815-1827. 2013.

[15] Tang, Lei, and Huan Liu. ”Leveraging social media networks for classification.” Data Mining and Knowledge Discovery 23, no. 3 (2011): 447-478.

[16] Riquelme, Fabián, and Pablo González-Cantergiani. ”Measuring user influence on Twitter: A survey.” Informationprocessing management 52.5 (2016): 949-975.

[17] Pennacchiotti, Marco, and Ana-Maria Popescu. ”Amachine learning approach to twitter user classification.” Fifth international AAAI conference on weblogs and social media. 2011.

[18] Mart´ın EG, Lavesson N, Doroud M. Hashtags and followers. Social Network Analysis and Mining. 2016 Dec 1;6(1):12.

[19] Zhao WX, Li S, He Y, Chang EY, Wen JR, Li X. Connecting social media to e-commerce: Cold-start product recommendation using microblogging information. IEEE Transactions on Knowledge and Data Engineering. 2015 Dec 17;28(5):1147-59.

[20] Gromov, Vasilii A., and Anton S. Konev. ”Precocious identification of popular topics on Twitter with the employment of predictive clustering.” Neural Computing and Applications 28, no. 11 (2017): 3317-3322.

[21] Liu, H. and Tang, L., 2010. Toward Collective Behavior Prediction via Social Dimension Extraction. IEEE Intelligent Systems, 25(4), pp.19-25.

[22] Wang, Chieh-Jen, Yung-Wei Lin, Ming-Feng Tsai, and Hsin-Hsi Chen. ”Mining subtopics from different aspects for diversifying search results.” Information retrieval 16, no. 4 (2013): 452-483.

[23] Liu, Lu, Jie Tang, Jiawei Han, Meng Jiang, and Shiqiang Yang. ”Mining topic-level influence in heterogeneous networks.” In Proceedings of the 19th ACM international conference on Information and knowledge management, pp. 199-208. 2010.

[24] Cataldi, Mario, Luigi Di Caro, and Claudio Schifanella. ”Emerging topic detection on twitter based on temporal and social terms evaluation.” In Proceedings of the tenth international workshop on multimedia data mining, pp. 1-10. 2010.

[25] Alshahrani, Mohammed, Fuxi Zhu, Lin Zheng, Soufiana Mekouar, and Sheng Huang. ”Selection of top-k influential users based on radius-neighborhood degree, multi-hops distance and selection threshold.” Journal of Big Data 5, no.1 (2018): 1-20.

[26] Khan MA, Bollegala D, Liu G, Sezaki K. Multi-tweet summarization of real-time events. In2013 International Conference on Social Computing 2013 Sep 8 (pp. 128-133). IEEE.

[27] Wang, Xufei, Lei Tang, Huan Liu, and Lei Wang. ”Learning with multi-resolution overlapping communities.” Knowledge and information systems 36, no. 2 (2013): 517-535.

[28] Zhao WX, Li S, He Y, Chang EY, Wen JR, Li X. Connecting social media to e-commerce: Cold-start product recommendation using microblogging information. IEEE Transactions on Knowledge and Data Engineering. 2015 Dec 17;28(5):1147-59.

[29] Cai, Hongyun, Zi Huang, Divesh Srivastava, and Qing Zhang. ”Indexing evolving events from tweet streams.” IEEE Transactions on Knowledge and Data Engineering 27, no. 11 (2015): 3001-3015.

[30] Cordeiro, Mário, Rui P. Sarmento, Pavel Brazdil, and João Gama. ”Evolving networks and social network analysis methods and techniques.” Social Media and Journalism: Trends, Connections, Implications (2018): 101.