International Journal of Advanced Technology in Engineering and Science
Vol. No.9, Issue No. 05, May 2021
www.ijates.com
ijates
ISSN 2348 - 7550

SOCIAL MEDIA DATA ANALYTICS BY PREDICTIVE AND CLUSTERING APPROACH

Mr. Shailesh Kumar Vyas (1)

Department of Computer Science & Engineering,

G.H. Raisoni University, M.P., India

Shailesh.vyas@ghru.edu.in

Mrs. Sonia Bajaj (2)

Department of Computer science & Engineering,

G.H. Raisoni University,M.P., India

Sonia.bajaj@ghru.edu.in

Abstract

Throughout the last years, the net has however seen a wider scope through the event of social media. Supported communication techniques and accessible to any or all, the media promote social interaction through the net. Several social networks exist and there are a unit over 900 social media sites on the market on the net. Variant of fifty eight million tweets per day. Massive information is that the border of the flexibility of an enterprise in term of storing, process and accessing all the information it desires for the effective functioning, and to create choices cut back risks, and conjointly to serve the various customers among on additionally cheap time. The main objective of this study is to spot completely different techniques of analyzing the large social media information. Distinctive these techniques can facilitate in revealing the competition's promoting strategy together with their content, audience, and messages.

Keywords: *Visual Clustering Approach, Twitter classifier models. Big data analytics, social media data, social media analysis, Hadoop, predictive analytics, text mining*

1. INTRODUCTION

The one factor that social media corporations mastery lies in its information. And this they need lots of it, because of their latency to induce users to share data regarding every waking minute. The massive body of knowledge at the disposal of social media corporations mirrors however individuals act with one another, and at the center of those interactions lies priceless data regarding what people and societies hold vital. This volume of knowledge, alongside the quick rate of knowledge flow that social media is renowned for, represent the essence of big data.

By applying analytics to social media information, massive information applications in numerous industries go on the far side the mechanics of interaction to seeing however the content contained within the interactions can have an effect on business performance and people's read of a complete. Content analytics allows corporations to

zero in on unjust data from the messages that users post. As an example analytics tools will be programmed to trace negative or positive sentiment a couple of complete as this might threaten name and revenue.

Facebook insight into user reactions and therefore the ability to roll out or modify its giving. [7]

What's additional, correlating content to demographics of user's age, gender, legal status, geographic location, financial gain levels, instructional accomplishment, and inclination to buy bound product permits an organization to grasp additional regarding the individuals it's managing. Such analysis conjointly reveals however adverts do among completely different client segments. This {can be} terribly helpful as advertisers can react in close to real time and modify campaigns to create additional revenue. Analysis of social media information collected by a distributor may as an example reveal that widowed females between twenty five and thirty five area unit appropriate candidates for a reduction supply on gymnasium instrumentality. Supported this data, the distributor may arrange to target these candidates with discount offers through Twitter, Facebook and different media. If analytics show the uptake and comments area unit dangerous, the supply will be refined to enhance performance.

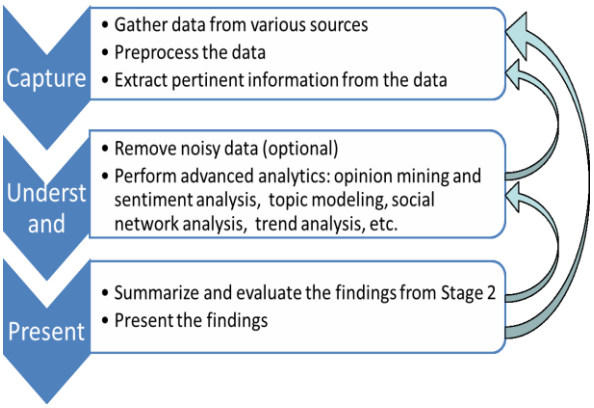
A lot of corporations appreciate the powerful nature of social media for personal-level interaction with their customers. The product customization existed since before social media, the extent and roughness to that it's done by businesses that collect social media information is astounding. Through social media analytics tools, these corporations will build data-driven choices by the minute.

Furthermore, social media analytics tools mean that companies will look on far side the chatter contained in unstructured information to seek out purposeful data which will guide choices and action. Through analysis of applied mathematics information like impressions per post, audience distribution, interactions on mobile versus desktop, responses, click-through rates for URLs embeds, and transactional history, an organization will live the effectiveness of its social media strategy for promoting complete recognition and loyalty.

Big information conjointly makes it potential to achieve insight into the roles individuals play among social media teams. Users with an oversized range of followers as an example, will be thought of to be influencers. By singling out such individuals, an organization will observe the flow in discussion forum and even participate in such forums.

He described a topic modeling framework for discovering health topics in Twitter, a social media website. That is an exploratory approach with the goal of understanding what health topics are commonly discussed in social media. The statistical model has been designed and demonstrated in detail for this purpose, the Ailment Topic Aspect Model (ATAM), along with our system for filtering conventional data available over social media platform based on health keywords and supervised classification. He showed how ATAM and other topic models can automatically infer health topics in 144 million Twitter messages from 2011 to 2013. ATAM has found many relevant clusters of data available on Twitter, some of which correlate with seasonal influenza ($r = 0.689$) and allergies ($r = 0.810$) temporal surveillance data, as well as exercise ($r = .534$) and obesity ($r = 2.631$) related geographic survey data in the United States. These outcomes describe that it is easy to find the topics based on statistic correlation along with the reliable information despite using very less manual observations and no historical information that is used to train the model, an extension to the previous work. Additionally, as a result of this work it has been observed that many different health issues can be recognized in social media using a single general purpose statistical model.

The data analytics in social media is accomplished into the following steps: *capture*, *understand*, and *present*. (See Figure 1). The *capture* stage involves obtaining relevant social media data by monitoring or "listening" to various social media sources, archiving relevant data and extracting pertinent information. This process can either be done by a company itself or through a third-party vendor. Not all data that are captured will be useful. The *understand* stage selects relevant data for modeling, removes noisy, low quality data, and employs various advanced data analytic methods to analyze the data retained and gain insights from it. These stages are conducted in an ongoing, iterative matter rather than strictly linearly.



2. BACKGROUND AND LITERATURE RE- VIEW

Sweeny *et al.* [11] proposed k anonymity model, a dataset is said to be k anonymous (k 1) if the published data have k anonymity preservation if the information for each per- son contained in the published data cannot be categorized from at least (k 1) persons whose information also appears in the published data. Machanavajjhala *et al.* [12] showed that k anonymous dataset may leak user's privacy. If background knowledge is available to an attacker then k anonymity may not protect the user's privacy in the k anonymous dataset. In this paper author present a diversity model, it is a group based anonymized t used to provide privacy for data sets by reducing the granularity of a data representation. L diversity requires that each equivalence class has at least a well-represented values for each sensitive attribute. Li *et al.* [13] Proposed t closeness model, i.e., an equivalence class is said to have t closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. The t closeness model extends the l diversity model by treating the values of an attribute distinctly by taking into account the distribution of data values for that attribute.

Liu *et al.* [14] demonstrate the degree based attack in social network data. In this paper author discussed that even after social networks data is anonymized does not guarantee user's privacy by simply removing the user's identity. A user may be identified by node degree in the social network graph. The author proposed k-degree anonymous solution, a graph called $k-$anonymous if for every node v, there exist

Zhou *et al.* have considered d neighborhood node attack in social network data. Here d represents the distance of neighbors from the targeted node and d *1*. If an attacker has the information of neighbors and relation between them of targeted user's node then the attacker may re-identify the targeted user's node from the social network data. In this paper, the author shows that the neighborhood attack may be real in practice and proposed an algorithm that can handle the A neighborhood attack problem. The proposed algorithm can use minimum DFS code to get an isomorphic check. But as d increase the number of possible DFS tree for every component increase exponentially.

2.1 Need and Importance of Problem

While massive information is bothered with all types of sources, it's calculable that the bulk of it comes from unstructured sources. Collectively may think, social media constitutes maybe the most important supply of unstructured sources for giant information. Likes, tweets, views, comments, favorites, and everything else that users will do an act with in any social media platform will be collected and analyzed by interested parties. Within the digital age, social media is very important for any business. Maintaining a presence on platforms like Facebook and Twitter is important as a result of it permits individuals to act with the corporate on apparently personal level that aids businesses across multiple fronts. It's conjointly vital for the typical client. Facebook alone boasts two billion monthly users, regarding twenty sixth of the world's entire population. Its important then to think about that massive information from social media will arrive in an unbelievable range of forms. [8]

3. OBJECTIVES OF THE RESEARCH

1. Identify the problems associated with data extraction in social media networks and possible avenues of research that may help to address these issues.
2. To Access huge amounts of data with less failures in utilizing the data to curb the cost of rising healthcare and by inefficient systems that stifle faster and better healthcare benefits across the board.
3. To describe the framework for data collection and analysis of multiple data streams, filters, and supervised classifiers for identifying of health records.
4. To analyze millions of health related tweets and separate, categorize non-ailment topics and Ailment topics.
5. To represent the outcome results using visual clustering approaches for clear visual view of clusters.

4. Scope of Research

The proposed work is designed in a way of supervised approach, i.e. it uses VCA for defining of health clusters based on symptoms of disease and classier model is built for giving treatment labels or solutions for diseases. Twitter model is evaluated for specific health cases and it is proved that social media networks are more useful for healthcare system. There is large scope to extend healthcare system for addressing of many challenging issues of social media networks.

5.TOOLS USED FOR BIG DATA ANALYTICS

In order to data analysis over over social media, every word should be treated as verb rather than noun. Prominently, it is a strategically collection of information from social platforms to follow the right path towards the result.

This method begins by prioritizing business goals. For example, your focus could also be to double the quantity of recent guests to your web site.

The second step is decisive key performance indicators (KPIs). During this case, your chief social media KPI would doubtless be based mostly on engagement stats. These can be broken down into:

- Likes and shares your post receives
- Replies and comments
- (most importantly) clicks your links and content earn

By collecting this data, you can figure out how social media factors into meeting your business goal. From there, you can keep going in the direction you're headed or adjust your approach.

The following tools can be used for data analytics: [6]

- Keyhole
- Agora Pulse
- Brand watch
- Buffer
- BuzzSumo
- Crowd booster
- Edgar
- Google Analytics and many more

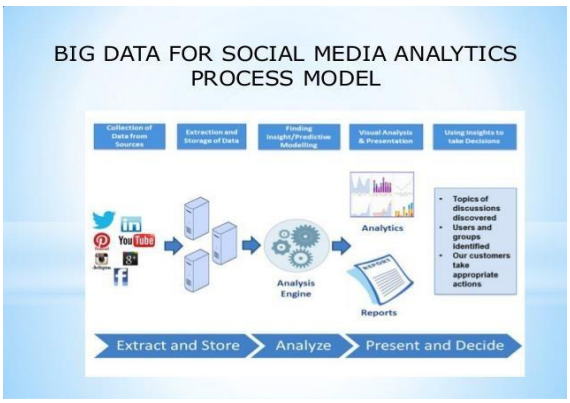


Fig-1: Process model for big social analytics

6.HOW BIG DATA IS ANALYZED

Big data is analyzed using different techniques which gives different outcomes. To use which technique is dependent on the user's need for analyzing the big data.

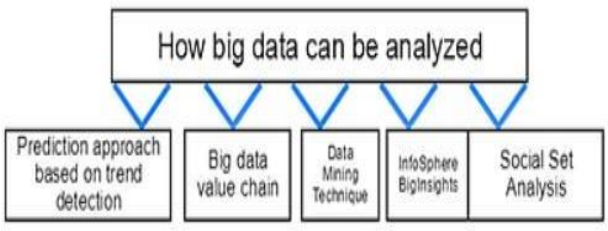


Fig-2: Ways of analyzing social bog data

Strategic approach based on flow detection: It provides the enhanced Strategic approach supported flow detection. First, a variable time deviation distance-based K-medoids rule is applied to cluster programs' quality evolution into four trends. Then, four flow -specific Strategic models are designed one by one using random forests regression. Consistent with the options extracted from an electronic program guide and early

viewing records, freshly printed programs are classified into the four trends by a gradient boosting decision tree. Finally, by combining prediction values from the trend-specific models and the classification probability, the planned approach achieves higher predication results.

Big data value chain: It presents a scientific framework to decompose massive information systems into four successive models, particularly information generation, information acquisition, information storage, and information analytics. These four modules form a big heap of information that is organized as a sequential information flow. Following that, paper provides a deep survey of various methodologies and procedures from analysis and business communities. Additionally, it presents the prevalent Hadoop framework for addressing massive information challenges. Finally, paper outlines many analysis benchmarks and potential analysis directions for giant information systems.

Data mining techniques: The increasing reliance on social networks needs data processing techniques that is likely to facilitate reforming the unstructured information and place them among a scientific pattern. The goal is to investigate the information mining techniques that were utilized by social media between 2003 and 2015.

Info Sphere Big Insights: It presents the usage of Twitter during a range of planned subjects that is that the largest social networking web site wherever Twitter information is in increasing at high speeds that also considered as huge amount of data is generated day by day. Then, describing well the manner within which massive information technology, such as, Info Sphere Big Insights allows process of this information, that are primarily collected from social networks by Apace Flume and keep in Hadoop storage.

Social cluster Analysis: It presents an innovative method to bulk information analytics known as social cluster analysis. Social cluster analysis consists of a progressive modules for the

Philosophies of procedure science, theory of social information, abstract and formal models of social information, and an analytical framework for combining massive social information sets with structure and social information sets.

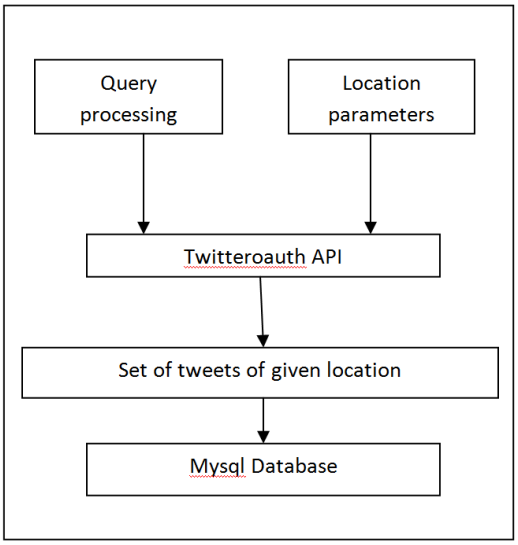
7. Methodology

7.1 Data Extraction

Information available at social network platform like messages, post and audio video were used as a data source. It is possible to extract tweets in a large scale from Twitter using the twitter public API that they provide. In our case we used the “twitteroauth” version of the public API by Williams (2012). This version has been implemented in PHP and can be run directly on the local host or on web servers. The query could contain several parameters. Twitter provides a large set of filtering parameters so that a well-defined set of tweets can be obtained. API is able to run the query so the query generation is the first step of this module then API will run this query and as a result of this execution the concerned twitter information is given as output in the browser. This data was directly inserted into a MySQL database for the use later on. Every available record on social networking sites like user name, tweet id, text etc. But our focus is on the text and id of a particular user so that we can extract the meaningful information out of them. Initially the twitter API allowed tweet locations in the form of latitude and longitude to be available with every tweet were the user has made his/her location public.

But due to security issues and user complaints this was stopped in 2012. This means that the geographical location from where the tweet was created is not available with the tweet. Our next purpose is to find the location of a user so this location is used as a filtering parameter then a query is generated using this filtering parameter. And this query is also considered as main query. So in compliance with this restriction we had to extract tweets based on a fixed set of locations. For our research we decided to focus on one nation, USA. We have taken a sample of tweets from various locations of USA and observed some tweets and extracted information. These locations have been chosen randomly and no such constraints have been applied for choosing the locations. We decided to go with data from New York, Los Angeles, Boston, Chicago, Dallas, San Francisco and Philadelphia for the experiments. Although we have used a free map tools by Viklund (2015) to choose the area of a particular location. Three parameters have been used to describe the city like latitude, longitude and center of the city so that the maximum portions of the city can be covered as much as possible. Even if a bit of excess was covered it does not really matter as those areas are generally very lightly populated and will not give results anyways. The latitude, longitude and radius are all values assigned to the 'locations' parameter in the query build. So now we have prepared various types of data sets that has been taken from the various cities.

The keyword we choose here in Ailments and Disease. Even though it is possible to analyze any disease in twitter the major issue is that ailment is to be trending in twitter. Remembering that reasonable amount of data about the ailment is available. But our proposed model will obtain results of any required data given that good amount of data available on twitter. So only tweets which contain the ailment will be obtained.



7.2 Data processing

In order to filter out these useless data we mainly used a tool of Natural Language Processing. This tool is an open source tool and it is developed by The Stanford NLP Group (SNLP Group 2015) .this tool is very efficient and widely used for processing and extracting the information. The strength of this NLP tool is that it can relate the words of a sentences grammatically which is also an output of this process. According to advanced linguistics several such relations are available in the English language. Generally we do not consider all kinds of relations generated by the NLP tool. Only relevant relations are observed that can fulfill our purpose of research.

So the dependent relations are chosen among all the relations generated by the SNLP tool. These dependencies are listed and explained in the Stanford Type Dependencies Manual (SNLP Manual 2015). The reason 50 dependencies are defined in the SNLP is because these are the only word relations which are useful to information analysts, even though linguistics defines several other word relations within a sentence. Out of these 50 dependencies we chose three which will be useful to us.

7.3 Implementation

The implementation process starts with the assessment on all the collected information from the tweets. We also consider a numeric parameter for the assessment of tweets having sentimental information. This process is done with the help of has been done using the SentiWordNet (2015) tool, which is also a supporting tool of Natural language processing tool. We take a word and a part of speech as an input to this SentiWordNet tool .This tool process the group of part of speech and the word itself then SentiWordNet generates a numeric score between −1 and 1 where lower value refers to more negative sentiment and higher value refers to higher sentiment. We will find the summation of scores generated for each word of social media messages by the tool. We cn also observe that the SentiWord is not able to identify sentences because only words and their related parts of speech can be taken as input by the tool. The part of speech and the word will depend completely on the sentence itself. So the SentiWord performs a mapping of each word of the sentence to its related part of speech. There is a mechanism known as Speech tag extraction that is used for extracting the information out of the tweets. SentiWord is also combined with the SNLP and is used to recognize a word associated with the part of speech within a given sentence. Before proceeding with the SentiWord, the analysis on twitter data is performed first using the POS tagger which will separate the tweet into individual words and assign a part of speech to it. This is required because by only assessing the word itself it is not possible to determine any sort of opinion, what part the given word plays within a sentence is always defined by the part of speech it using. A custom program is implemented to perform mapping or normalization of the POS tags allocated by the POS tagger. Knowing that SentiWord only recognizes nouns, adjectives, adverbs and verbs, any parts of speech other than these three had to be mapped to any of these. A mapping can be performed in such a way that if a word is allocated the verb tag, which can identify a verb in present tense, it will be assigned the Verb tag by the mapper. This group of words and normalized values of these words with POS tags are then sent to SentiWord and the output of SentiWord will contain the final value of each word and then final score is generated by adding the every numeric sentiments for the twitter data.

8. SOCIAL MEDIA AS Big DATA

Social media become one of the prominent and relevant data sources for big data. Social media data produced a wide varieties of Internet applications and websites, with most popular being Facebook, Twitter, LinkedIn, YouTube, Flicker and Instagram. Every day trillions and trillions of data are being generated in social media websites.

With growth of these social media sites allow users to be connected and creates an environment for interacting, sharing and collaborating communication. Information that are generated has been spread to many different areas such as everyday life (e-business, e-tourism, hobbies, friendship, etc.), education, health, and daily works. Exponential growth of social media has produced a challenging issue for conventional data analysis algorithm techniques ,such as machine learning, data mining, statistics, and so on, due to their high computational complexity for large datasets.

9. SECURITY ISSUE IN SOCIAL MEDIA

Social media becomes the best way for getting the information about a person because users are sharing lot of their information on social media. Social media users have up- loaded their personal information in user-profile and share lots of personal information. As much information available on social media attracts the attackers to steal the personal information of users. Because of that attackers use a different kind of attacks to access the user-profile data from social media and that problem raises the security issues in social media [10].

10. SOCIAL MEDIA AS BIG DATA

Social media become one of the prominent and relevant data sources for big data. Social media data produced a wide variations of Internet applications and websites, with most popular being Facebook, Twitter, LinkedIn, YouTube, Flicker and Instagram. Every day trillions and trillions of data are being generated in social media websites.

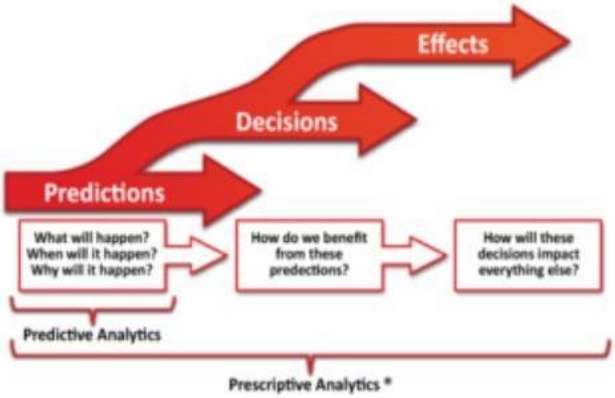
With growth of these social media sites allow users to be connected and creates an environment for interacting, sharing and collaborating communication. Information that are generated has been spread to many different areas such as everyday life (e-business, e-tourism, hobbies, friendship, etc.), education, health, and daily works. Exponential growth of social media has produced a challenging issue for conventional data analysis algorithm techniques ,such as machine learning, data mining, statistics, and so on, due to their high computational complexity for large datasets.

11. LIMITATIONS

In proposed work we are taking the tweets of common people who are posting their day to day routine health problems. We are processing the twitter data which is extracted by “twitteroauth”. The data processing technique which we using is open source natural language processor so it will process 50 predefined relations. In future people may discuss about many ailments in social media. For extraction and processing of those tags may be complicated in future.

12. CHALLENGES

Big information can still accelerate the intrusion of social media corporations into People’s Privacy. As Facebook, Twitter, Instagram and Pinterest still monetize their offerings, it might seem that the advantages that massive information can have for social media within the future can become even additional personalized. A study printed by researchers from Cambridge and Stanford Universities shows that Facebook will use its information to predict people’s temperament with additional accuracy than shut friends and families.



13. CONCLUSION

Social media operator fetch and store data from the social media user for the purpose to share among a huge varieties of third party consumers. As the fetched information often contain sensitive data, network operator release the complete graph in an anonymized and sanitized versions. But it does not provide full guarantee of the user privacy. The attacker may use structural based attacks on social network data to re-identify the users and get the user's information. Mostly researches focus on social network data attacks and find out the different type of attack patterns (like; degree based attack, neighborhood attack, sub graph attack). Using these attacks, an attacker may re-identify the user in the social network data and acquire the user's information.

This work is based on controlling graph based attack in the social network graph. The graph based attack is based on the targeted neighbor's information and the relationship between them. Most researches happened for the social network anonymization by adding dummy edges and dummy vertices information in the social networks. Adding dummy data in the social networks creates the information loss and it caused for the inappropriate result in a research study. The proposed anonymization process will increase the number of isomorphic neighborhood networks by adding dummy edges in the social network graph. Therefore, a user may not be re-identified in social network graph based on its unique neighborhood network.

REFERENCES

1. Culotta, A.: Towards detecting influenza epidemics by analyzing Twitter messages. KDD Workshop on Social Media Analytics (2010)
2. Culotta, A.: Lightweight methods to estimate influenza rates and alcohol sales volume from Twitter messages. Language Resources and Evaluation, Special Issue on Analysis of Short Texts on the Web (2012)
3. Lampos, V., Cristianini, N.: Tracking the flu pandemic by monitoring the social web, IAPR 2nd Workshop on Cognitive Information Processing (2012)
4. Maskawa, S., Aramaki, E., Morita, and M.: Twitter catches the flu: Detecting influenza epidemics using Twitter. Conference on Empirical Methods in Natural Language Processing (2010)
5. Yoon, S., Elhadad, N., Bakken, S.: A Practical Approach for Content Mining of Tweets, American Journal of Preventive Medicine, 45(1), (2013)
6. Golder, S., Macy, M.W.: Diurnal and Seasonal Mood Varies with Work, Sleep and Day length Across Diverse Cultures. Science 333(6051): 1878–1881, (2011)
7. Moreno, M., Christakis, D.A., Egan, K.G., Brockman, L.N., Becker, T.: Associations between displayed alcohol references on Facebook and problem drinking among college students, Arch PediatrAdolesc Med (2011)

8. Michael, J. Paul., Mark, D.: Discovering health topics in social media using topic models, plos one, 9(8), 1-14, (2014)
9. Michael, J. Paul., Abeed, S., John, S.B., Azadeh, N.,Matthew, S., Karen, L.S., Graciela, G.: Social Media Mining for public health monitoring and surveillance, 468-479, Pacific Symposium on Biocomputing (2016)
10. Blei, D.: Probabilistic topic models, Communications of the ACM 55(4): 77–84,35, (2012)
11. A.Machanavajjhala, D.Kifer, J.Gehrke, and M. Venkitasubramaniam, "*A*—diversity: Privacy beyond *k*—anonymity, *ACM Trans. Knowl. Discov. Data*, vol. 1, no. 1, Mar. 2007. [Online]. Available: http://doi.acm.org/10.1145/1217299.1217302
12. N. Li, T. Li, and S.Venkatasubramanian, "*t*—closeness: Privacy beyond k—anonymity and A—diversity," in 2007 *IEEE 23rd International Conference on Data Engineering*, April 2007, pp. 106–115.[Online].Available: https://doi.org/10.1109/ICDE.2007.367856
13. L. Liu, J. Wang, J. Liu, and J. Zhang, *Privacy Preservation in Social Networks with Sensitive Edge Weights*,pp.954–965.[Online]. Available:https://epubs.siam.org/doi/ abs/10.1137/1.9781611972795.82
14. B. Zhou and J. Pei, "Preserving privacy in social networks against neighborhood attacks," in 2008 *IEEE 24th International Conference on Data Engineering*, April 2008, pp. 506–515. [Online]. Available:https: //doi.org/10.1109/ICDE.2008.4497459