INTERNATIONAL JOURNAL OF SOFT COMPUTING

MEDWELL PUBLICATIONS

Social Media Analytics: Data Utilization of Social Media for Research

Edi Surya Negara and Ria Andryani
Data Science Interdisciplinary Research Center, Universitas Bina Darma Jl. A. Yani No. 3, 30624 Palembang, Indonesia

Key words: Production, opportunities, various parties, perception, information

Corresponding Author:
Edi Surya Negara
Data Science Interdisciplinary Research Center, Universitas Bina Darma Jl. A. Yani No. 3, 30624 Palembang, Indonesia

Page No.: 111-118
Volume: 15, Issue 5, 2020
ISSN: 1816-9503
International Journal of Soft Computing

Abstract: The amount of production data generated by social media opportunities that can be exploited by various parties, both government and private sectors to produce the information. Social media data can be used to know the behavior and public perception of the phenomenon or a particular event. To obtain and analyze social media data needed depth knowledge of internet technology, social media, data bases, data structures, information theory, data mining, machine learning, until the data and information visualization techniques. In this research, social media analysis on a particular topic and the development of prototype devices software used as a tool of social media data retrieval or retrieval of data applications. Social Media Analytics (SMA) aims to make the process of analysis and synthesis of social media data to produce information can be used by those in need. SMA process is done in three stages, namely: Capture, Understand and Present. This research is exploratory focused on understanding the technology that became the basis social media using various techniques exist and are already used in the study of social media analytic previously.

INTRODUCTION

Development of technology and information which very fast are allowed society to have communication directly, face to face and it is also come into online communication, such as the using of online media. By using the online, the society are able to communicate even they have been separated by mile or more than it. Actually, this phenomena tends to be called as the using of social media in which it is known as online media that connect the users entire the world with development of population too.

The users of social media increase for every year. This can be evidence based on the data which promote by communication and information department at kominfo.go.id and it showed that the user of internet in Indonesia attained 63 million with 95% of them takes the internet to access social media. Based on the survey of Global Web Index in January 2014, the user of internet in Indonesia filled 72.700.000 person from 251.160.124 person in Indonesia. The survey also showed the active users of social media attained 79.7% from the user of internet in Indonesia[1].

Looking to the condition above, it can be said that the user of internet tended to use social media. The uses of social media data are to produce information which can be reviewed or it tends to call Social Media Analytics (SMA) and it has been done in current days. Actually, there are

many fields which function SMA such as: economy, business[2], healthy and epidemiology[3-6], development of social until fields of dynamics city[7, 8] etc. Social media is often used to express phenomena which happen in certain location. Therefore, the data which have been produced from social media are known widely and bigger information. If the data tend to wider form, it will have opportunity to produce accurate information.

Process of SMA has been done by retrieving, storing, processing and visualizing data. Actually, the process is taken as method to make social media to be one of research media which is possible to analyze data and it is used as source of the data in conducting a research.

The challenge which is taken in using data of social media are: each site of social media uses different platform volume, complexity from information and the data unstructured[9]. SMA gets the challenge by providing tolls and framework to collect, evaluates, analyze, conclude and visualize data of social media[10].

To use data of social media in supporting research activities which relate to perception society and social behavior are deal with tools and framework that developed in specific ways to collect, evaluate, analyze, conclude and visualize the data. The problem which needed to response through this research is the tools and framework, such as: what are things which needed to support research activities on social phenomena which have data of social media. Considering the complexity and kinds of platform which are used by each media social, thus this research will discuss SMA at micro-blogger site. Twitter or it is often called with Twitter Data Analytics. This research takes a phenomenon, namely crash of Air Asia with flying code QZ8501 at 28th January 2014 in strait of Karimata and it looks as object to look perception and society behavior toward the phenomenon. Then it will be processed by using SMA, especially for the data on Twitter which discuss the problem.

Literature review

Social media: Internet and Web 2.0 provides a platform which is used to improve services with functionally to: creating and sharing ideas and story (Blogger and Twitter); sharing information and links (Delicious, Digg and Twine); sharing multimedia (You Tube and Flickr), making and sharing knowledge (Wikipedia, Yahoo! Answer and SlideShare) and sharing partnership (Facebook, MySpace and Linkedln) by big groups. These service which are known as social media.

Social media is a platform which gives service through two ways, namely making and sharing are taken as tools of new communication in digital era that can be used to reform networking into community and it gives possibility to have communication through online communication in which it tends to make, manage, edit, comment, tagging, discussion, grouping, connecting and sharing any information which are included on it. Currently, Twitter is one of familiar social media. Actually, Twitter is a micro-blogging which can be used to send message until 140 characters by fasting through platforms. In this case, there are 90% interactions of the Twitter but it comes from other website, namely mobile message, fast message or desktop application.

Current days, there are many kinds of social media, such as: Social networks, Blogs, Wikis, Podcast, Forums, Content communities, Micro-blogging, etc that can be used to get certain aims[11]. By implementing related theories on social presence, media richness and social process Kaplan and Haenlein has claimed that social media can be divided into six types, namely: Collaborative projects, Blogs and micro-blogs, content communities, social networking sites, virtual game worlds and virtual communities. But if it looked from the category, social media has been divided into four categories, namely: Social networking, social collaboration, social publishing and social feedbacks[12].

Social media analytics: Social Media Analytics (SMA) tends to activity in developing and evaluating tools of information and framework to collect, evaluate, analyze, conclude and visualize data of social media[10, 13].

Gartner Research defines SMA is a process to look, analyze, measure and predict digital interaction, relationships, topic, ideas or contents at social media[19]. SMA aims to have analysis process and synthesis data of social media until it gives information which functionally for the stakeholders. Process of SMA takes three steps, namely Capture, Understand and Present[10]. The steps of SMA can be seen at Fig. 1.

The capture on the process of SMA tends to process in collecting data of social media which relevant toward needs by using crawler tools that have been connected to the Application Programming Interface (API) of social media, such as: Facebook, Twitter, LinkedIn, YouTube, Pinterest, Google+, Tumblr, Foursquare, Internet forums, blogs and micro-blogs, Wikis, news sites, picture sharing sites, podcasts and social bookmarking sites, etc. The data which has been produced from Capture process will be saved into database and it is provided to further process, namely Understand process. In this step, the data is processed to give its information which relevant with needs and it includes creating a model the data form[13].

After finishing the Capture process, the next is Understand process. Actually, the Understand process at SMA is a process to choose the data which relevant to

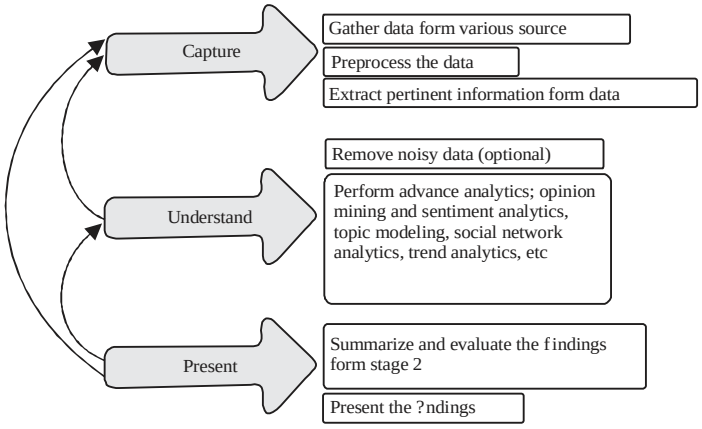

Fig. 1: Social media analytics process[13]

apply data modeling, break of noise which includes at the data, selecting the data with high quality and making process to analyze the data in which it takes to have the best information[13]. In this step, process of analysis data get statistic method, text mining, data mining, Natural Language Processing (NLP), machine translation, machine learning and network analysis[13]. Many techniques analysis data of social media which are possible to use in producing information: Opinion mining (or sentiment analysis), Topic modeling, Social network analysis, Trend analysis and Visual Analytics[14]. The last step on process of SMA is Present. Initially, Present is a process to show or visually information which get from step of Understand[13]. In this case, there are many kinds of visualization that can be used to show the information from the analysis process.

MATERIALS AND METHODS

Micro-blogger twitter terminology: Twitter is one of media social which popular and it have 8th rank at Alexa rank. Initially, Twitter comes from idea of Jack Dorsey in 2006 and he looks habitual of society which wants to share their activities when they have quality time with other people. In the development of Twitter, Jack Dorsey combined pola of communication from one to be more and it took as basic of pola for communication on Twitter. It gives possibility to the users of Twitter to share information with other people.

Nepplenbroek etc. describe architecture of Twitter development take model ”4+1” in which it had been developed by Kruchtens[22]. This model is used to describe architecture of software by focusing logical, process, physical and development and scenario views. By using model of Kruchtens[22], describe architecture of Twitter with Logical view, Process view, Physical view, Development view and Scenario view.

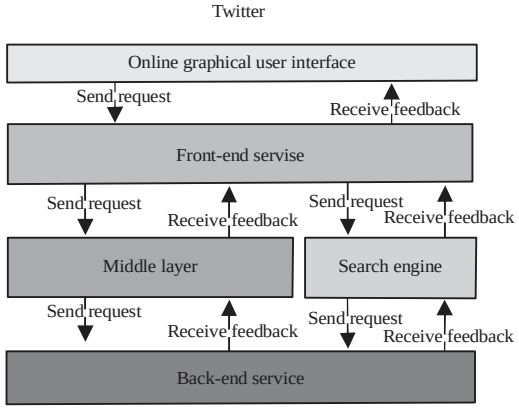

Fig. 2: Architecture of Twitter

Architecture of Twitter development can be seen at Fig. 2. Appendix of Back-end Service from Twitter saved all of Tweets which has been posted by members with MSQL as database of saving data. In the line of Search Engine, Twitter used Apaches Lucene. Search Engine at the Twitter used invert indexing method in which it is used to separate Twess to be words of a sentence. Actually, line of Middle Layer at the architecture of Twitter is taken as requesting system until it cannot burdening Back-end Service. In the first, Line of Middle Layer is implemented by Starling by using program language Rubby on Rails.

RESULTS AND DISCUSSION

Capture data: Based on the research which had been conducted, it can be seen there were prototype application of retrieval data and framework for social media analytics. In this research, the writer used Twitter as object of the research. To analyze the data of Twitter, it needed certain techniques in order the data (text) on the Twitter to be information which took as objects of the research.

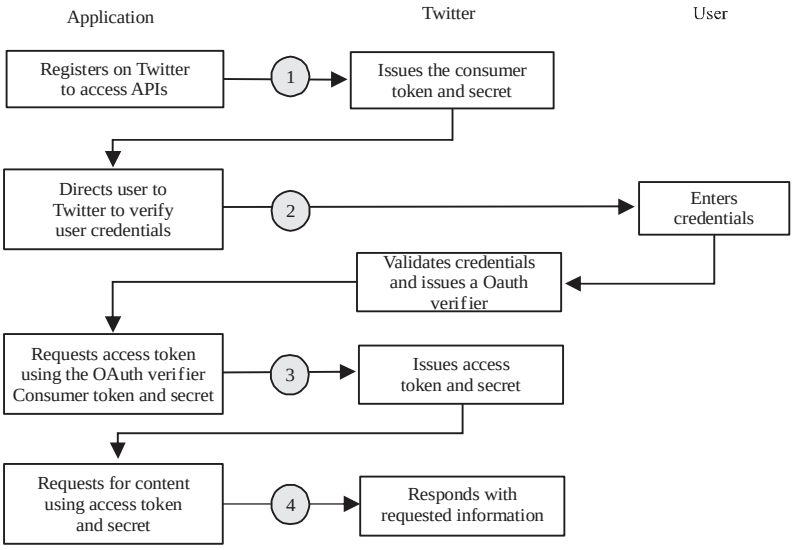

Fig. 3: Oauth proses

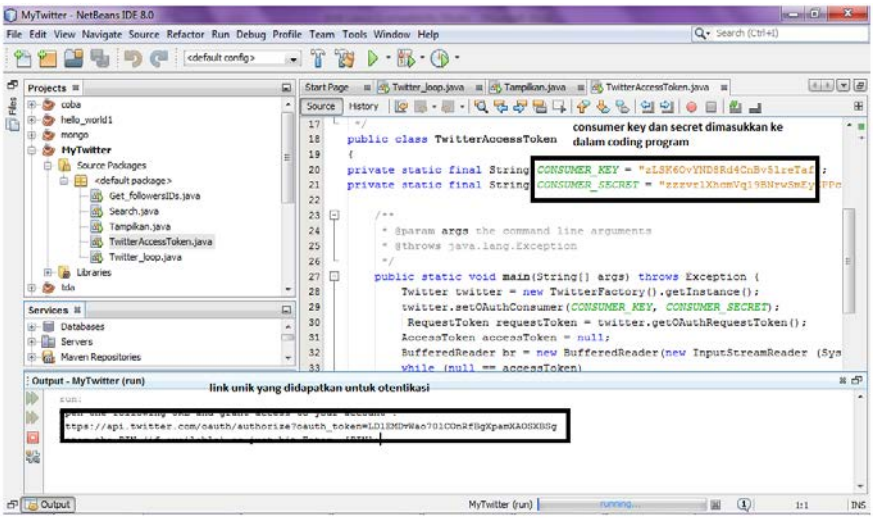

Fig. 4: Code PIN authentication

The first step that should be done in social media analytics is retrieving data of Twitter. To have retrieving, it should register the application which developed toward Twitter and it needs to be done in order there is permission to have user credential in which it will be used into process of retrieving data of Twitter. This process is aim to get authentication from the Twitter toward data access on the Twitter. Process of authentication can be showed into Fig. 3. After the process retrieving of the Twitter data successes, the next step is saving the data to database of MongoDB, like Fig. 4.

In the process of retrieving, the data which got are user name, re-tweet count, tweet followers count, source and tweet mentioned count, tweet ID and tweet text. User name is name of the user or Twitter, re-tweet count describe how many times the status re-tweeted by the other user. Tweet followers count belongs to total of follower from the user of its account, source is media that has been used to up-date the tweets, tweets mentioned count show how many the tweet which followed, tweet ID is ID of the Twitter user and the tweet text is content of the tweet.

In the process of retrieving data, there are many factors which affect it, namely connection of the internet, time to collect data and up-data new news that will be done. The connection of the internet is very important to

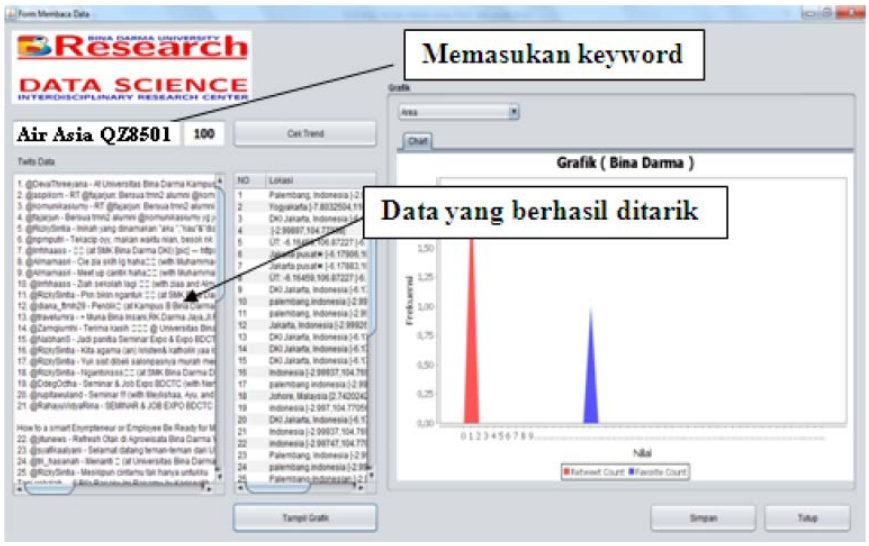

Fig. 5: Application of data retrieval

process retrieving. The connection of the internet which is stable will support the process of retrieving data. Contrary, if the connection un-stable, it will be error or process of retrieving is to be slow and break-down.

The second factor is the longest of process retrieving data. If the retrieving data of Twitter is done into longest time, it will have more data from the process. Then, the third factor is up-data new news. This is happened because retrieving data which has been done by Twitter relate to real time. Therefore, in this process streaming API is needed. In this case, the data which is taken is real-time data. Retrieving data can be done at new news which posted 10 days before. Furthermore, the method which is used to retrieving data is REST API. Thus, the new news which is happening or have been happened will be easy to know the process of development.

Representational State Transfer API (REST-API) is one of architecture model of software to distribute hypermedia system like WWW. The term is promoted firstly by Roy Fielding on his doctoral dissertation in 2000 and he is a major writer of HTTP specification[15]. Specifically, REST tends to collection of principles networking architecture which stress role of definition and keep of sources. This term is often used with widely understanding to describe all of simple face-to-face which convey the data on HTTP specific domain without additional line like SOAP or tracking session which used HTTP cookies. REST-API on the Twitter can be used to access status or timelines of the Twitter user. REST-API can take 3.200 new tweets from the user include re-tweet:

- Main parameter; every page, we can take 200 Tweet from the user
- Rate limit; an application is allowed to have request into 300 requests

Understand data: In this step, the data which took from the Twitter entered to the database. Database which used in this research was MongoDB.

Furthermore, the data which success saved would be analyzed to get clear data without noises. The clear data can be used as data for the research. To get comprehension toward the data which showed as information, it needs to be visualized into bubble graph form or the other graphs based on the needs.

The process saving of the data should be done directly or it tends to call as direct storing. It needs to be done because the data which got from retrieving data are real-time data of Twitter. Thus, it needs to use a database which possible to save the data directly (Fig. 5).

The result of retrieving data which got from the process of retrieving data tended to the data in text form or document. The >100 data texts which took through retrieving data. It will be difficult to create primary data manually because process of retrieving data indicate that the data which got should be manage into database directly.

In this research, MongoDB or Mongo database is taken because it tends to document or text database. The easy of using the database is primary key should not use to insert the data. Because of MongoDB will make ID object or primary key automatically on the process of submit the data into database.

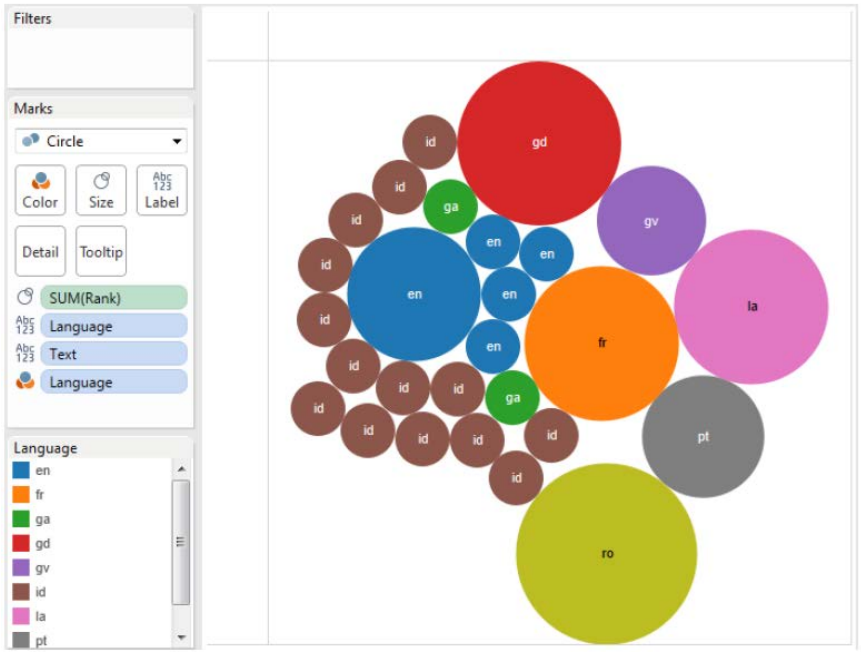

Fig. 6: User of Twitter based on the countries, source: Twitter (Air Asia QZ88501)

In this step it needs to do field analysis based on Tweet Text and User Name. The process of analysis toward Tweet Text will show the possibility of the countries from language of the user. The countries have been symbolized with code of two letter country based on ISO639-1. The result of analysis on language identification showed three columns, namely text, language and rank. Field text showed Tweet or text which up-data by the user of Twitter, field language showed the possibility of the country and field rank showed the rank of the possibility of the country.

Furthermore, the field which will be analyzed is User Name. Process of analysis on user name will show sex of the user, characteristics of the user and age of the user of Twitter. In this case, the result of the analysis can be called as user demo-graphics (Fig. 6).

Present data: Visualization is certain way to convert data into visual format or table until characteristics of the data and relation between items of the data or attributes can be analyzed or reported. Visualization of the data is one of technique that appropriate and interest to use in exploring the data. Moreover, the visualization can be used to describe general pattern that happen, trend which comes as hot issues or the other phenomena.

After process of analysis finish, the next step is visualization of the data. It is used to get interesting form and easy to understand the data as certain information. To look the data which produces as information on easy way, it should be visualized into bubble graph form and the others based on needs. Like previous aims of the research, the result of this research/finding is to make crawling data of Twitter by using Application Programming Interface (API) and it has been provided by Twitter. Based on data of Twitter, it will be processed to be information in which it can be used as object of the research.

The information is also used to reflect how behavior of society toward a phenomenon on social media which happen in the human life. It shows what the phenomena have influence toward global society. The influence of the phenomena can be seen from the tweets which up-date by the user of Twitter. In this case, the language uses will be used to look the possibility where the users come from (country). Based on the analysis on data of Twitter in which it discuss phenomena crash of Air Asia QZ8501, it can be seen that the most of users on Twitter whom give attention come from Indonesia.

Out of the language uses, behavior of society can be look from the age of the Twitter user. Based on the one phenomenon, it can be seen that frame of the users age in which the frame show the active users on the Twitter. Moreover, sex, personality and organization which includes on the user can be known through the result of this research.

In this research, the writer conducted a research about crash of Air Asia QZ8501 which happened at 28th January 2014 in strait of Karimata Indonesia. Moreover, the writer will show where come from (country) the users that look the phenomenon based on language uses on their tweets.

Fig. 7: Sex of the user on Twitter, sources Twitter (Air Asia QZ88501)

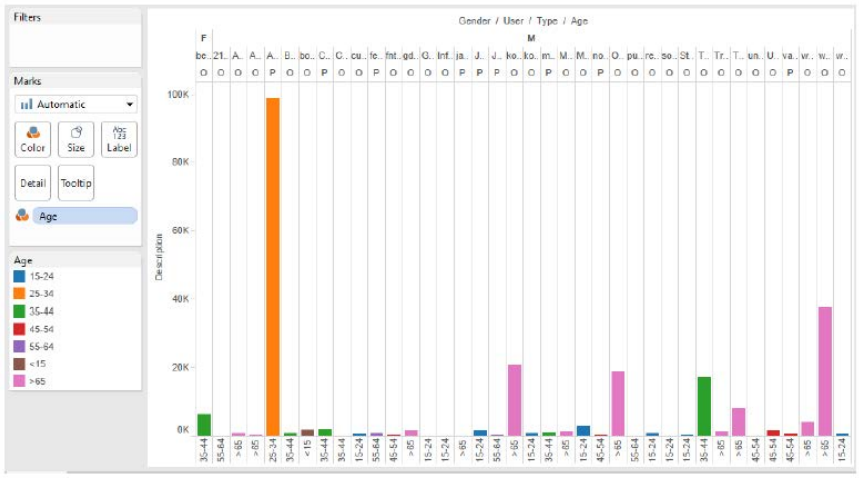

Fig. 8: Age of Twitter user, Twitter (Air Asia QZ88501)

It is not juts country, it is also used to show frame of age which active to observe the phenomenon from the users. Moreover, sex and characteristic of the users whether they come from organization or person are showed in the result of this research.

Out of the identification toward language uses, this research is also used to show characteristics, frame of ages, sex of the users.

The results/findings of this research can be showed into Fig. 6-8. Frame of age which discuss crash of Air Asia QZ8501 are 25-34 years old and the sex are dominated by male.

CONCLUSION

Based the research on social media analytics, it is Twitter, it can be concluded; Crawling process toward data of Twitter by using Application Programming Interface has been done successfully and it showed that informative data processed through capture, understand and present.

By having the data, it can be identified sex, age, characteristics of the active user and the country which has the user of Twitter, it is known from the language uses. The society gets response toward phenomena. The

respond can be measured to know how the extent of the phenomena effect for them. This research can be used to look how the extent of someone influence or certain issues and the current phenomena.

REFERENCES

01. Global Web Index, 2014. Survei data global web index. Global Web Index, London, UK.
02. Holsapple, C., S.H. Hsiao and R. Pakath, 2014. Business social media analytics: Definition, benefits and challenges. Proceedings of the 20th Americas Conference on Information Systems (AMCIS2014), August 7-9, 2014, Association for Information Systems, Savannah, Georgia, USA., pp: 1-12.
03. Carneiro, H.A. and E. Mylonakis, 2009. Google trends: A web-based tool for real-time surveillance of disease outbreaks. Clin. Infect. Dis., 49: 1557-1564.
04. Corley, C.D., D.J. Cook, A.R. Mikler and K.P. Singh, 2010. Using Web and Social Media for Influenza Surveillance. In: Advances in Computational Biology, Arabnia, H. (Ed.). Springer, New York, USA., pp: 559-564.
05. Culotta, A., 2010. Towards detecting influenza epidemics by analyzing Twitter messages. Proceedings of the 1st Workshop on Social Media Analytics, July 2010, ACM, Washington, USA., pp: 115-122.
06. Paul, M.J. and M. Dredze, 2011. You are what you tweet: Analyzing twitter for public health. Proceedings of the 5th International AAAI Conference on Weblogs and Social Media, July 5, 2011, AAAI, Menlo Park, California, pp: 265-272.
07. Cranshaw, J., R. Schwartz, J.I. Hong and N. Sadeh, 2012. The livehoods project: Utilizing social media to understand the dynamics of a city. Proceedings of the 6th International AAAI Conference on Weblogs and Social Media, June 4-8, 2012, Trinity College, Dublin, Ireland, pp: 58-65.
08. Long, X., L. Jin and J. Joshi, 2012. Exploring trajectory-driven local geographic topics in foursquare. Proceedings of the 2012 ACM Conference on Ubiquitous Computing, September 2012, ACM, Pittsburgh, Pennsylvania, pp: 927-934.
09. Stieglitz, S. and L. Dang-Xuan, 2013. Social media and political communication: A social media analytics framework. Social Network Anal. Mining, 3: 1277-1291.
10. Zeng, D., H. Chen, R. Lusch and S.H. Li, 2010. Social media analytics and intelligence. IEEE. Intel. Syst., 25: 13-16.
11. Sterne, J. and D.M. Scott, 2010. Social Media Metrics: How to Measure and Optimize your Marketing Investment. John Wiley, Hoboken, New Jersey, USA.,.
12. Bradley, A.J., 2010. Becoming a social organization: Taking a strategic approach to social media. Gartner Inc., Stamford, Connecticut.
13. Fan, W. and M.D. Gordon, 2014. The power of social media analytics. Commun. ACM., 57: 74-81.
14. Fan, W., L. Wallace, S. Rich and Z. Zhang, 2006. Tapping the power of text mining. Commun. ACM., 49: 76-82.
15. Fielding, R.T. and R.N. Taylor, 2002. Principled design of the modern web architecture. ACM. Trans. Internet Technol. (TOIT.), 2: 115-150.