ISSN 2395-1621

Social Media Mining

[#1]Monika V. Dongre, [#2]Paridhi D. Agrawal

[1]monika.dongre22@gmail.com
[2]paridhiagrawal17@gmail.com

[#12]Department of Computer,

Savitribai Phule University
Pune, India.

ABSTRACT

The impact and influence of social media have been very significant and tremendous from the last few years. The access to social networking sites like Yahoo, Google+ Facebook, LinkedIn and Twitter with the amalgamation of the internet and new technologies has become very easy and effortless. People are relying on and are becoming completely dependent on social media. Social Media provides vast, plentiful, voluminous information regarding various subjects and fields. The heavy dependency and use of social networking sites result in the generation of massive data, mainly characterized by three computational issues and factors namely; extent, noise, and range. Handling such a massive social media data is a really complex task. For this purpose, various techniques, algorithms, computational methods are applied to handle and analyse the data and extract useful information like the trends, patterns, and knowledge. This paper primarily focuses on discussing and shedding a light on the various techniques, algorithms, issues and applications that are used to mine diverse aspects and attributes of the social media and network.

Keywords: Social Media Mining(SMM).

ARTICLE INFO

Article History

Received: 18th March 2017

Received in revised form :

18th March 2017

Accepted: 22nd March 2017

Published online :

4th April 2017

I. INTRODUCTION

Social media can be described as a cluster of Internet-based applications that emphasizes, permits the generation and trade-off of user-generated data and content. Social media provides users a means of communication, to connect with each other, and with people living in different parts of the world in an unprecedented way, scale and time. The social network is an emerging term that is used to articulate and explicitly define web-oriented services that allow peers or individuals inside a sphere or domain to make a semi-public/public profile so that they can communicate effectively and link with new peers within the network. In a simpler language, we can say that social network represents a graph that consists of links and nodes, used to signify social relationship on social media websites. *The nodes in the graph represent the individuals and the associations among them are exposed by the links connecting the entities (*see Fig 1.). In simpler terms, SMM is nothing but the mining of information from social media.

Data Mining primary imperatives and objectives are to find useful and unknown information from large data sets, extract patterns and trends and acquire shrewd and insightful knowledge [11]. Various data mining techniques for collecting, searching, analysing data in database to discover patterns and relationships, information retrieval, AI, statistical methods and computations are applied for mining social media data. The step-by-step techniques include *data pre-processing, data exploration, and data understanding* processes in data investigation. Twitter generates over 400 million Tweets daily and Facebook approximately over 1 billion posts [8]. Mining the social media can provide researchers the capability of analysing, understanding new technologies and phenomena that will provide better services create and explore innovative opportunities and make important contributions towards the social media research and development.

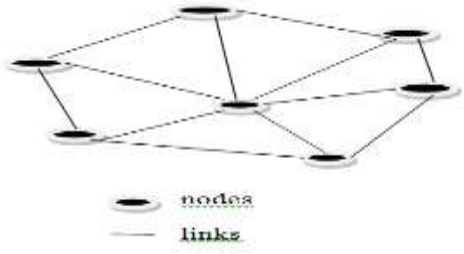

Fig 1. Social Network consisting nodes and links.

II. Social Media Classification

The different types of social media are Social News, Wikis, Social bookmarking, Media Sharing, Opinion, reviews, and ratings, Microblogging, Blogging, Online Social Networking, Answers, Online Shopping sites.

Fig 2. Social Media

III. Different Mining Methods And Algorithms

1. Social Network Analysis (SNA) and Graph mining

SNA fundamentally aims and focuses on studying relationships between peers or individuals, instead of their properties or attributes [1]. It investigates social structures through the use of graph theories and networks [2]. Graph theory is the primitive method in SNA. This method and technique are used in SNA in a mandate to find out and govern significant features of the network such as the nodes and links. Graph mining is used to find out relationships as well as data and content. An example of graph mining or graph theory application on Facebook is, "Places or restaurants visited by friends in Mumbai to hang out".

The different graph algorithms are as follows:

1. Graph/Tree Traversal Algorithms
   - I. Depth-First Search (DFS)
   - II. Breadth-First Search (BFS)
2. Shortest Path Algorithms
   - I. Dijkstra's Algorithm
   - II. Bellman-Ford Algorithm
   - III. Floyd-Warshall Algorithm
3. Minimum Spanning Tree Algorithms
   - I. Prim's Algorithm
   - II. Kruskal's Algorithm
4. Maximum Flow Algorithms
   - I. Ford-Fulkerson Algorithm
5. Matching Algorithms
   - I. Bipartite Matching
   - II. Weighted Matching

1.

2. Text Mining

Text Mining is another useful, emerging and significant technique that aims to extract information from structured, semi- structured or unstructured data. A social network consists of huge amount of texts, posts, pictures, messages that are daily generated. Text mining techniques are useful in many ways like the automatic classification of textual data, automatic email processing like junk, spam, and important emails are separated, we get the information and knowledge about website automatically, sentiment analysis [6], and biomedical text mining. The two text mining methods are as follows:

- Cluster Analysis: Semi-automatic or automatic analysis of the voluminous data to mine the previously never seen before patterns like sets or groups of data known as cluster analysis.
- Anomaly Detection: It can be referred to searching of events or items which do not approve to an anticipated event.

.

3. Aspect-based/feature based Opinion Mining

In the aspect based or feature based mining, the range of unit customers has revised or studied is mined because not all aspects/qualities of an object are consistently revised by customers. It then becomes necessary to summarize the aspects studied to analyse the division of the complete evaluation whether they are affirmative or undesirable. Some reviews are vague, equivocal and uncertain and thus sentiments expressed on some objects are simpler to examine than others. In the aspect-based opinion, the main problem lies mostly with forum discussions and blogs rather than in service or product reviews. The object/aspects (which might be a PC device) studied is either *'thumbs up'* or *'thumbs down',* thumbs up denotes an affirmative review whereas thumbs down denotes an undesirable review. In blogs and forum discussions, aspects and objects are not known and there are great levels of inconsequential data which create noise. Hence, it is necessary to recognize review sentences in each assessment to decide if the review sentence is affirmative or undesirable. This helps to summarize aspect-based opinion which improves the complete mining of service or product review.

IV. Tools Used For Social Media Mining

The various tools used to mine social media are as follows: Centrifuge, Commetrix, Cuttlefish, Cytoscape, Egonet, Gephi, Graph-tool, GraphChi, Graphviz, InFlow, JUNG(Java Universal Network/Graph Framework), Keynetiq, MeerKat, Netlytic, NetMiner, Network Workbench, NetworKit, NetworkX, NodeXL, Pajek, and Polinode.

V. Applications of Social Media Mining

1. Web Mining

Web mining involves the application of data mining strategies to discover trends and patterns from the World Wide Web. Web mining can be firmly divided into three different types –- The Web usage mining, Web content mining, and Web structure mining.

2. Facebook API

Search API: The graph API is an HTTP-based API that provides admittance to the Facebook social graph, unvaryingly represented objects in graph and association between them.

FQL: Facebook Query Language allows you to use a type of interface that is just like SQL interface to query the data exposed by the graph API.

Dialogs: Facebook affords a numeral of dialogs for Facebook Login, posting on a person's timeline, blocking, poking, and sending requests.

3. Applications of SNA

SNA [2] is basically used for identification of teams, individuals, and units who play the centre parts or centre roles. It is used to make stronger the already existing communication links and networks. It can also be applied to improve the techniques, algorithms, and refine strategies to facilitate innovation, creativity and empower learning.

4. Community Analysis

A real-life community can be described as the form of individuals with mutual commercial, communal, or radical interests/characteristics, habitually existing in qualified juxtaposition. Information about a community entails – 1) a combination or combination of at the minimum two nodes involving similar inquisitiveness and – 2) communications and collaborations with detail to that inquisitiveness. There exist two kinds of clusters in social media – Explicit Clusters: shaped by user contributions – Implicit Clusters: formed implicitly by social collaborations (individuals calling New York from Toronto) the telephone operator studies them as one community or cluster for advertising determinations [3].

5. Sentiment Analysis:

This involves analysis of sentiments, for example, analysis of a movie review for examining and estimating how favourable or unfavourable a review is for a movie. This kind of analysis may require a labelled data set or labelling of the affectivity of words.

VI. CHALLENGES AND ISSUES IN SOCIAL MEDIA MINING

1. Evaluation Dilemma

As the data is voluminous and plentiful, evaluation dilemma arises [9]. We need to find and explore new ways for the evaluation of the data.In traditional data mining; tested and training datasets are used for performance comparing and to validate findings. This is useful to develop and evaluate models against some ground truth. However, the traditional data sets may not be viable in the case of social media data and thus will raise a question in front of the researchers to consider and find other ways that will help to validate their claims in absence of unrecognized ground truth.

2. Big Data Paradox

The data generated by the social media sites every day can be measured in terms of petabytes (for example Facebook, twitter, and quora). But what exactly is Big Data? Is it 3Vs, 4Vs, 5Vs..... nobody can measure and keep a track on how big the data is and in the upcoming years what will be the size of this data. Handling such a big data paradox is a major challenge.

3. Noise Removal Fallacy

Removal of noise from the data can sometimes render the data from social media sites to become unimportant or useless. Posts such as dish the people liked when they visited the restaurant or what a person ate when he returned home after work are some of the sources of noise in social media data. The naturally connected structure of social media data leads to complications of the tasks and causes the researchers to examine and approach noise-removal in a pretty different way than they would with attribute- key value pair data.

4. Detecting Deception

Detecting Deception is another major issue and challenge in social media mining. Fake information can be generated with the intention to deceive people on social media. This fake information may spread through social media in the same way as valid information.

VII. CONCLUSION

This paper effectively covers all the important techniques, algorithms, the challenges and the issues related for mining the social media. Various data mining techniques have also been used in the analysis of social media data and network. The application of these techniques has produced substantial and successful outcomes that have helped to gain insightful knowledge of the pervading social media. Various levels of success have been accomplished by making use of either one or integration of one or more techniques.

The diverse results of these experiments have established the relevancy and usage of various data mining techniques in retrieving patterns, trends, and knowledge from the enormous amount of social media data. In the upcoming years, more new techniques will be discovered, but with this, there also lies the need to overcome the new challenges, drawbacks, and issues. In future, we intend to work on the new and the old techniques and thus investigate the wide social network.

REFERENCES

[1] Gloor, P.A., Krauss, J., Nann, S., Fischbach, K., and Schoder, D, "Web Science 2.0: Identifying Trends through Semantic Social Network Analysis". Computational Science and Engineering, 2009.

[2] C.K.-S. Leung and C.L. Carmichael, "Exploring social networks: a frequent pattern visualization approach". in Proc. IEEE SocialCom, 2010.

[3] N. Mishra, R. Schreiber, I. Stanton, and R.E. Tarjan, "Clustering social networks". Proc. WAW, 2007.

[4] Shuo Chang, Vikas Kumar, Eric Gilbert, and Loren Terveen, "Specialization, homophily, and gender in a social curation site: Findings from pinterest". CSCW, 2014.

[5] Wendy Liu and Derek Ruths, "What's in a name? using first names as features for gender inference in twitter". AAAI Spring Symposium: Analyzing Microtext, 2013.

[6] Shandilya, S.K., Jain, S, "Automatic Opinion Extraction from Web Documents". Computer and Automation Engineering, 2009.

[7] Christy M.K. Cheung, Matthew K.O. Lee and Christian Wagner, "Introduction to Social Media and e-business Transformation Minitrack". 47th Hawaii International Conference on System Science, 2014.

[8] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "Twitterrank: finding topic-sensitive influential twitterers". *WSDM*, 2010.

[9] A. M. Kaplan and M. Haenlein, "Users of the world, unite! The challenges and opportunities of social media". Business Horizons, 2010.

[10] Michele Merler, Bert Huang, Lexing, Gang Hua and Apostol Netsev, "Semantic model vectors for complex video event recognition". Multimedia IEEE Transactions, 2012.

[11] F. Bonchi, "Influence propagation in social networks: a data mining perspective". IEEE Intelligent Informatics Bulletin, 2011.

