**Knowledge Discovery from Social Media**

**Rupa G Mehta1, Bharat Tidke2**

Associate Professor1, Phd Scholar2

Department of Computer Engineering

Sardar Vallabhbhai National Institute of Technology

Surat, India.

**Abstract:** Social science is the study of society and human behaviors. Social media a group of online communication channels committed to community driven input, interaction, information sharing and relationships. Modern social scientists refer social media as a source of information which has significant impact on human social life. Websites and applications dedicated to Social network, Discussion Forum, Social publishing, Online review, Media sharing networks and Ecommerce are among various social media.  Their social media activities, connections (links), and products are collected and saved which can be analyzed by scientists. This has opened a new era of research called Social Data Mining (SDM), which talks about the process of representing, analyzing, and extracting known and hidden but actionable patterns from huge social media data. The Knowledge extracted from social media has been a strong pillar for decision makers and scientists in their major activities. Initiatives have been taken by various social media platforms to collect and share their data with other platforms whereas users are inspired to share their information and ideas on social media platforms, ultimately businesses and researchers utilize these social media data for their analysis. This paper reviews different aspects of social media platforms to provide a fair understanding of various techniques, opportunities, issues and challenges while dealing with social media data.

**Keywords:** Social science, Social media data, Social data mining, Big data.

1. **Introduction**

Social science is the study of society and human behaviors includes psychology, anthropology, communication, economics, human geography, ethnography, et al., where everything comes down to people. Unfortunately, in recent scenario the modern social media, mobile or web strategy involved in communication has been technology concentric makes data to grow rapidly that generates huge unstructured data. This gave sudden rise to concept of Big Data [7]. Rob Kitchin describes Big Data with the following characteristics: huge in volume, high in velocity, diverse in variety, exhaustive in scope, fine-grained in resolution, relational in nature, and flexible in trait [11]. Social media data is not only big but linked, noisy, highly unstructured and incomplete, which may lead to failure or inappropriate understandings about the data by idealistically applying existing data mining techniques or conventional theories of social science. Modern technology combines the tools and methods to analyze Big Data with the platforms which generate the Big Data itself which lead to trace social lives of ordinary people and get transformed into behavior patterns [25].

**Figure 1. Leading Social networks worldwide in terms of active user in 2016**

In the view of the technologist, social media helps to build relationships where two or more concepts, objects, or people are connected, or in state of being connected which is dynamic in nature and multidimensional too. The study of social media can make you feel both empathy and harmony. Where empathy is the source of inspiring desired actions, reactions, and transactions related to all those involved in commerce and relationship models.

SDM can help to shape public policy, to drive business decisions for supermarkets, to address obesity and food health, to study and predicting probable frauds and fraudents for the secured society, to analyze meteorological data weather prediction and to avoid disasters due to natural calamities etc.

1. **Social Media data and Social Theories**

The complex link between the data for each specific individual makes it Big in nature. The link makes it patently not independent and identically distributed, which contradicts one of the most enduring assumptions of traditional data mining methods. The passive content consumers and active content producers are responsible to generate continuous noise along with other traditional noise issues. The spammers are responsible to spread malicious or unwanted messages [8]. Quantification of the relationship may lead to fuzzy relationship [5]. User generated content in social media is often highly unstructured due to usage of various electronics devices for accessing different social media sites. The incorrect sentence formation in the messages, user defined abbreviation makes the information more unstructured [15]. For privacy issues only partial data about the users are visible to others make the data incomplete.

The ubiquitous presence of Social relations in social media data substantially differentiates it from data in traditional data mining and social sciences. The directed network approach is used to represent the social data, where all users are represented by nodes and the relationship among various users are represented by signed or unsigned links of the graph. The complex relationships between different entities are converted into simple binary relationships (pair wise relationship). Researchers exploit the positive link to relate people to their friends, fans, followers, and collaborators whereas negative link indicates foe, enemy and dislike which helps in studying controversy and disagreement [7].

Many theories in social science like correlation, balance and status have been proven to be pertinent to directed and social media network. As per Social correlation theory, contiguous users in a social network have similar behaviors or attributes. These phenomena clarify user’s inclination to connect or follow with others having certain similarity or sharing the same surroundings. Balance theory guides in justifying instinctive statements like “the friend of friend is friend,” “the friend of enemy is enemy,” “the enemy of my friend is my enemy,” and “the enemy of my enemy is my friend” [27] [13]. Status theory utilize links to signify the relative status superior or inferior among two users and assist in proving the statement like “If A is inferior than B and B is inferior than C than A is inferior than A” [13].

Social Media network is a composition of user information, link information and the content on the network produced by various users has open the research horizon for the researchers and scientists for three categories of tasks, user based tasks, link based task and content based task. The extensive literature is reviewed for different social media and SMM task is discussed in the following section.

1. **Literature Survey**

Evolution and acceptance of online social media in past few years has altered the Internet environment, as it advances to a further collaborative environment [1]. Millions of web user’s uses social networking sites not only for communication, but to create social communities and also for uploading and storing social media content in profound ways. Businesses and research community uses social media data generated from online interactions and from interconnected correlation links from various social networking sites such as Facebook, Twitter, LinkedIn, Instagram, Pinterest, etc [1][2]. The capability to store large and variety of data from different platforms using various technological means has implicitly makes a revolutionary changes in social network research, which ultimately leads to evolution of a computational social science.

Social media can be categorized into two types. First in terms of their context e.g platforms such as social networking sites was clearly designed for social interactions. Other platforms like e-commerce, opinion sharing or data sharing sites normally used for a various applications which share multimedia contents in compliance with comprehensive intend of social interaction. These platforms contain text, images, audio or video which leveraged for a wide variety of purposes [6].

Social media data sourced from various platforms including Social networks sites, Discussion Forum, Social publishing, Online review, Media sharing networks and Ecommerce. It provides many features like every user on these platforms are provided with free web space to read, write or to share their content and creating circle of friends. Similar to search engines like Google these social media sites have their own search engines to locate people, news or multimedia contents. It permits users to build their own profiles which allow their friends or associates to recognize them. Social networking sites like Facebook or Twitter are among most used websites on internet having hundreds of millions active users [20]. Online chats become one of the most used features of these social networking sites. Members can also send e-mails to each other. The extensive comparison of all characteristics of social media platforms is depicted in Table 1.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Social Media Platforms** | **Key Features** | **Focus** | **Search engine** | **Profile Information** | **Demographics** | **Networking Transparency** | **Online Chat** | **Pros** | **Cons** |
| **Facebook** | Most widely used social networking site  Fairly easy to get stats about user activity | Sharing news, stories, content | Searches for all the categories. | Detailed  information | 1.5 billion monthly active users  1 million links every 20 minutes | Displays all the posts from the friends and friends’ friends. | Yes | Ability to stay in touch with their friendsand relatives  Easy place for businesses to expand | Becomes pay to play network  Privacy is still a concern |
| **Twitter** | Micro blogging social site that limits posts to 140 characters  Posting information in real time | Sharing news, stories, content | Searches for tweets and people | Minimal Information | 289 million active users  9,100 tweets happen every second | Displays only tweets from the people followed by the user and any re-tweets for these tweets given by other users | No | Real time updates  More Effective Ad option  Quickly absorb information | Limited amount of space |
| **Youtube** | Easy to use,  channels allow you to create libraries of videos | Sharing of informative and entertainment videos | Searches for Videos | Minimal Information | Over billion active users  60 hours of video are uploaded every minute | Display subscribed channels videos | No | No limit for uploading videos  Information can be shared in term of videos | Searching videos from large volume is tedious task  Unwanted links or videos |
| **Instagram** | Image-friendly  Fully integrated with Facebook, Twitter and other major platforms | Sharing contents | Searches for all the categories | Detailed  information | 600 million monthly active users | Display pictures of friends and friends’ friends. | Yes | Allows for powerful showing (rather than telling) of impact  Good marketing platform with visual effect | Requires an understanding of good visual content  Gap in information dissemination |
| **Google +** | Allows for brands and users to build circles | Sharing news stories content | Searches for all the categories | Detailed  information | 300 million active users | Displays all the posts from the friends and friends’ friends. | Yes | Simple but powerful UI  Linked to all Google products | No Contests or Promotions |
| **LinkedIn** | Build connections and professional relationships  Effort needed to maintain a presence is less than other platforms | Sharing news and information related to companies | Searches for People and companies | Minimal Information | 380 million users  79% of users are 35 or older | Display information of followed person or company | No | Reliable network for professional use | Requires larger time commitment  Only 1% of user are active |
| **MySpace** | * Encourage [user-generated content](https://blog.hootsuite.com/how-to-use-user-generated-content-for-marketing/) | Sharing news stories content | Searches for all the categories | Detailed  information | 50 million users worldwide | Displays all the posts from the friends and friends’ friends. | Yes | Simple Design and user friendly | Not attractive as compare to Facebook |

**Table 1. Summary of various characteristics of social media sites**

**3.1 User based analysis**

Users on social media are from various communities and background having different taste and choices, while active on social networking sites. Finding right user for specific analysis or recommendation can helps the organization to better understand their customers. User based analysis mainly divided into three categories. First is group of users or community where peoples are grouped based on similar interest. They are also related to each other based on some forehand knowing criteria’s like any other communities in the real world e.g. college group. Second is classification of users based on different characteristics such as behavior, preferences, etc. Third group is of spammers. As social networking sites do not give well-built authentication mechanisms, and it is quite easy to masquerade as a user and creep into a person’s network of trust whether he is friend or not known as spammers [27].

**Figure 2. Various methods for Social Media Mining**

**3.2 Link Based Analysis**

Social networking sites can be formulated as graphs, where each vertex reflects as user and edges reflects as relationships between many users. Social networking sites are active networks which changes dynamically with time, incoming new vertexes and edges are regularly combined with existing graphs. To explore relation based structural analysis which mainly based on link prediction, influence analysis, community discovery using correlational links, social tie prediction as well as tie strength prediction, etc. Link prediction used to predict the probable forthcoming connection between two vertexes. Various methods have been proposed for link prediction such as Linear Algebra [25], feature-based classification [10] and probabilistic methods [14]. Linear Algebra determines similarity among two vertexes based on singular similar matrix. Feature-based classification normally chooses a set of features for a vertex and exploits the current link information to create binary classifiers for predicting future link. Probabilistic methods objective is to develop models for connection probabilities between vertexes in Social networking sites.

**3.3 Content Based Analysis**

Social media and networking sites contains textual data in the form of messages, opinions, multimedia data such as images, videos, locations etc. Almost every content analysis such as data analysis, text analysis, and multimedia analysis can be translated to social media analytics. In general, content analysis can be define as process of summarizing, quantitative analysis of messages or text that relies on the scientific method includes attention to objectivity, intersubjectivity, apriori design, reliability, validity, generalizability, replicability, and hypothesis testing [16].

The formation of an efficient process is important in collecting, analyzing, and organizing unstructured revealing information present in social media into [intelligible](http://www.thesaurus.com/browse/intelligible) concepts for various decision support applications, such as Feature selection [23], Sentiment analysis [10], Crowdsourcing [6], Profiling [25], Trend and topic detection [15], Social recommendation [26], etc. However, social media content based analytics deals with many unique challenges. First, there is big and continuously growing social media data that must be analyze within a limited time constraint. Second, social media data includes noisy data as it is unstructured data by nature. Third, social networks are dynamic and continuously changing. Techniques for content analytics in social networks include association, summarization, classifications and clustering on textual as well as multimedia data. All of above techniques can be useful in many applications which have been explored in next section.

1. **Applications of Social Media**

Various industries from different part of the society are viewing social media as big platform to reach users to communicate or for promoting their brands. Users also prefer to connect with their brands they think about using social media. The main goals of various sectors for using social media are

* Receiving more traffic to their respective website
* Build and improve associations with customers
* Increase visibility to raise number of viewers

Ultimately, social media marketing is more effective for some industries than others, but all industries can take advantage of it if they approach it with the right goals and the right strategies.

* **Finance and Marketing**

Social media drastically changes banking relationships by remodeling various services for example users now allow to send or receive money from others via online platforms. Modern financial technology based companies use social media data to improve their facilities by providing people with credit cards or getting loan. Social media marketing are on top priority for company’s general marketing policy, allowing them to agree on data-driven approach where they understand mindset of dynamically changing social environment [24].

* [**Retail**](http://www.cbsnews.com/8301-505125_162-49140989/5-businesses-that-will-live-or-die-by-social-media/?tag=bnetdomain)

Retail businesses like ‘Groupon’ and ‘Living Social’ flourish on distribution through social media. But conventional retailers are coming in the competition as well. Customers can contribute by their feedback to improve quality of services [21].

* [**Education**](http://mashable.com/2010/09/29/social-media-in-school/)

Social media are more popular among students becomes a great platform for learning, eager about enchanting on assignments. Professors, supporting staff and management can also enjoy creativity provided by social media as well, collaborate and distribution of information using social media is a new a way to reach masses [28].

* [**Travel and Hospitality**](http://www.customerserviceinthecloud.com/blog/2011/08/26/top-5-industries-using-social-media-for-customer-service/)

Travel and hospitality industry are very competitive businesses and uses social media as medium to reach people from diverse background to promote their sales. Additionally, travelers also shared their complaints and feedback which could be helpful in developing and maintaining a good reputation for the company [12].

* **Agriculture**

Youtube is one of the most famous platforms among farmers as they can get information about different aspects of agriculture online. [22].

1. **Conclusions and Future Research Directions**

Social media data is driven by statistical and analytical technologies to obtain information for various decisions making. Social media data have the capability to solve business and scientific analysis problems by providing opportunities and thoughts. This review is suitable for researchers, analyst and practitioners in the area of social media mining to support in realizing directions, challenges and assumptions while dealing with large scale social media data.

There are still some issues and open research challenges on which one can focus to explore social media data in more appropriate way in future. In the remainder section, summary of this issues and possible research directions has been discussed.

* **Escaping from disintegration of social correlations through accessible cross-platform interrelationship**

A major barrier in utilization of social network data is the disintegration of community of social network users into various proprietary and concealed social networks. This scenario is associated by the matter that every new social media application inclined to develop their individual social network. Instead they can construct their network based on existing rich and interrelated data. Reliable mechanisms to preserve privacy are an essential prerequisite [17].

* **Community finding and Sentiment analysis in large scale online social networks**

Social networks data will keep on rising in terms of size, velocity and variety, designing specific relationship and finding communities from large scale social networks is lively research challenge. Similarly sentiment analysis remains fancy research challenge while dealing with various characteristics of social networks data like volume velocity and variety [3].

* **Social Networks Analysis for providing security in different domains**

Social Networks can be useful media for providing security for various domains by gaining knowledge from different online interactions on social networking sites. Analysis of terrorism [4], is one such applications where social media can play vital role by analyzing behavior of suspected people on social networking sites.

* **Ethical Issues in a social networks**

Social networking sites such as Facebook and MySpace faces serious ethical issues of sharing personal information, child exploitation, etc. that need to be considered and solve appropriately [5].

* **Spam detection and adversarial interactions in social media**

The detection of spam user or opinion need to be studied to stop irrelevant users and information spreading in online social networks [9] [18].

* **Personalization of users based on online social interaction**

Users with specific taste, preferences or developing trends must be located based on some appropriate content using personalization algorithm and techniques with improved accuracy, while recommending products or events [29].

**References**

1. Aggarwal, Charu C. "An introduction to social network data analytics." Social network data analytics. Springer US, 2011. 1-15.
2. B. Shneiderman and C. Plaisant, "Sharpening Analytic Focus to Cope with Big Data Volume and Variety" in IEEE Computer Graphics and Applications, vol. 35, no. 3, pp. 10-14, May-June 2015.
3. Balahur, Alexandra, and Guillaume Jacquet. "Sentiment analysis meets social media–Challenges and solutions of the field in view of the current information sharing context." (2015): 428-432.
4. Ball, Leslie. "Automating social network analysis: A power tool for counter-terrorism." Security Journal 29.2 (2016): 147-168.
5. Beasley, Asaf, Winter Mason, and Eliot Smith. "Inferring emotions and self-relevant domains in social media: Challenges and future directions." Translational Issues in Psychological Science 2, no. 3 (2016): 238.
6. Cai, Ji Li Zhipeng, Mingyuan Yan, and Yingshu Li. "Using crowdsourced data in location-based social networks to explore influence maximization." In Computer Communications, IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on, pp. 1-9. IEEE, 2016.[34]
7. Felt, Mylynn. "Social media and the social sciences: How researchers employ Big Data analytics." Big Data & Society 3, no. 1 (2016): 2053951716645828.
8. G. Stringhini, C. Kruegel, and G. Vigna. Detecting spammers on social networks. In ACSAC, 2010.
9. G. Brown, T. Howe, M. Ihbe, A. Prakash, and K. Borders. Social networks and context-aware spam. In ACM Conference on Supportive Cooperative Work, 2008.
10. H. Tang, S. Tan, X. Cheng, “A survey on sentiment detection of reviews”, Expert Syst. Appl. 36, 10760–10773, 2009.
11. Kitchin R, “The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences.” Thousand Oaks, CA: Sage (2014).
12. Leung, Daniel, Rob Law, Hubert Van Hoof, and Dimitrios Buhalis. "Social media in tourism and hospitality: A literature review." Journal of Travel & Tourism Marketing 30, no. 1-2 (2013): 3-22.
13. Leskovec, Jure, Daniel Huttenlocher, and Jon Kleinberg. "Signed networks in social media." In Proceedings of the SIGCHI conference on human factors in computing systems, pp. 1361-1370. ACM, 2010.
14. M. Mathioudakis and N. Koudas, “Twitter Monitor,” in Proceedings of the international conference on Management of data - SIGMOD ’10, p. 1155, 2010.
15. Moreno-Munoz, A., F. J. Bellido-Outeirino, P. Siano, and M. A. Gomez-Nieto. "Mobile social media for smart grids customer engagement: Emerging trends and challenges." Renewable and Sustainable Energy Reviews 53 (2016): 1611-1616.
16. M. Speriosu, N. Sudan, S. Upadhyay, and J. Baldridge. Twitter polarity classification with label propagation over lexical links and the follower graph. In ULNLP, 2011.
17. Neuendorf, K.A., The Content Analysis Guidebook, CA, Thousand Oaks: Sage Publications, 2002.
18. O’Keeffe, Gwenn Schurgin. "Social media: challenges and concerns for families." Pediatric clinics of North America 63, no. 5 (2016): 841-849.
19. Pennington, Diane. "Analysing found non-text social media data: options and challenges." Social Media and Society (2016): 1-3.
20. R. Xiang, J. Neville, and M. Rogati. Modeling relationship strength in online social networks. In WWW, 2010.
21. Rapp, Adam, Lauren Skinner Beitelspacher, Dhruv Grewal, and Douglas E. Hughes. "Understanding social media effects across seller, retailer, and consumer interactions." Journal of the Academy of Marketing Science 41, no. 5 (2013): 547-566.
22. Rhoades, Emily, and Kelly Aue. "Social agriculture: Adoption of social media by agricultural editors and broadcasters." In Southern Association of Agricultural Scientists Conference. Orlando, FL. 2010.
23. Salah AA, Manovich L, Salah AA, et al. (2013) Combining cultural analytics and networks analysis: Studying a social network site with user-generated content. Journal of Broadcasting & Electronic Media 57(3): 409–426.
24. Schwienbacher, Armin. "The internet, crowdfunding and the banking industry." In The Palgrave Handbook of European Banking, pp. 213-229. Palgrave Macmillan UK, 2016.
25. Tang, Jiliang, and Huan Liu. "An unsupervised feature selection framework for social media data." IEEE Transactions on Knowledge and Data Engineering 26.12 (2014): 2914-2927.
26. Tang J, Sun J, Wang C, Yang Z, “Social influence analysis in large-scale networks”. In: Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 807–816, 2009.
27. Tang, Jiliang, Yi Chang, and Huan Liu. "Mining social media with social theories: a survey." ACM SIGKDD Explorations Newsletter 15, no. 2 (2014): 20-29.
28. Yasemin, G and Distance Education Center. "Current State of Usage of Social Media for Education: Case of Turkey." *Journal of Social Media Studies* 1, no. 1 (2014): 53-69.
29. Zafarani, Reza, and Huan Liu. "Connecting users across social media sites: a behavioral-modeling approach." In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 41-49. ACM, 2013.