# Analysis of Mining Social Media – A Learners Perspective

**Gunasundari P , Research Scholar, Department of Computer Science, Bharathiar University, Coimbatore, India, gunasundari82@gmail.com**

**Dr R Thirumalai selvi, Assistant Professor, Department of computer science, Government Arts College (Men) Nandanam, Chennai, India, sarasselvi@gmail.com**

**Abstract In the world of digitalization the data play a key role. The data may be in structured or unstructured. The structured data uses data mining techniques to find the unknown pattern from the known data. But, the social media has huge data due to its rapid growth, the data were dynamic and unstructured. Due to this traditional data mining techniques will not be appropriate. The combinational approach of data mining and social media will provide the user to gain an insight and prominent idea how can be mined. Social media provides each individual to connect with the others depending on their interest. Every individual are accessing Face book, Twitter, LinkedIn, academicia.edu, Google+ for sharing their views and thoughts, day-to-day happenings with any one or more of the above sites. This paper give an idea of the how those sites are classified based on their size, data, research focus, design issues and the types of the sites, types of users and the common approaches on social networks which will help the researchers how the social media, social networking websites structurally classified, studies the existing data mining techniques along with the performance metrics used in past researches and tools for retrieving social media data.**

***Keywords – Academic Social Network; Face book; Google+;Graph; Homophily network; Social media, Recommender System; Social Network portal; Social Network Analysis; Mobile SNS; Social Network Measures.***

## I. INTRODUCTION

"Social media" as what the second word in its name implies -- a form of media. Essentially, social media is a platform for broadcasting information, whereas social networking is a platform for communicating with one another. Social media is a communications channel, whereas in social networking, the communication has a two-way nature [22]

Data mining is the process of extracting the unknown, unsuspected patterns from the known dataset. The data analysis performed using data mining techniques are known as secondary data analysis. Data used for data mining process is a structured data and it is stored in various repositories.

Social media data mining is used to uncover hidden patterns and trends from social media platforms like Twitter, LinkedIn, Face book, and others. This is typically done through machine learning, mathematics, and statistical techniques. [23]

Social network is a term used to describe web-based services that allow individuals to create a public/ semi-public profile within a domain such that they can communicatively connect with other users within the network [3]. Social network has improved the concept and technology of Web2.0, by enabling the formation and exchange of User-Generated Content [8]. Social network is a graph consisting of nodes and links used to represent social relations on social network sites. The nodes include entities and the relationships between them and form the links [2].

Social networks are important sources of online interactions and contents sharing [16], subjectivity, assessments, approaches, evaluation, influences, observation, feelings, opinions and sentiments expressions borne out in text, reviews, blogs, discussions, remarks, reactions, or some other documents [10].

## II. SOCIAL NETWORK ANALYSIS

A class 10 student of St. Aloysius High School Mangalore has created a social networking website of his own. With this step, he has shaved off a good five years from the age at which Mark Zuckerberg started Face book, which is the biggest social media website in the world presently. [13]

A social network is a structure make up of individuals (or organizations) called nodes which are tied (connected) by one or more specific types of interdependency, such as friendships, kinship, common interest, financial exchange, dislike, relationship of beliefs, knowledge or prestige [18].

Social Network Analysis (SNA) is used analyze the interpersonal relationships within an organization or community and can provide rich and systematic descriptions and interpretations of complex social relationships. SNA focuses on the interconnections of the

 DOI : 10.35291/2454-9150.2020.0017 

actors, instead of on the peculiarities of the actor themselves.

In social network analysis the main task is usually about how to extract social network from different communication resources. [7][12]

This paper has specifically focused on the techniques used to mine social network data. The major focus is on the various categories of web mining such as web usage mining, web structure mining and web content mining. [25]

This paper presented a study on data mining application, data mining techniques with special reference to the fields of medicine or healthcare and diseases. [26]

This paper used twitter data set to analyze the sentiments of the users. The J48, Decision Tree and Navies Bayes are classifier used. The 10-fold cross validation methods used along with performance metrics accuracy, F-measure, ROC area. DT classifiers outperform navies bayes in this context. Weka tool is used for classification and cross validation[27]

An experimental setup was made for the comparison of various classification algorithms were performed using MATLAB. Out of 4 algorithms among 6 datasets, the SVM and KNN bagged the first and second place using average accuracy performance metrics [28]

## III. SOCIAL NETWORK CLASSIFICATION [3]

The social network can be classified into three main categories such as small world network, scale fee network and homophily network [6].

*Regular Network* - Regular network has a connection only to 4 nearest neighbors. A small world network has an extra long-distance connection. Each point has an extra connection to a more distant point, minimizing the number of links needed to reach across the network. Random network connections to 4 other points. Each point has an extra connection to a more distant point, minimizing the number of links needed to reach across the network.

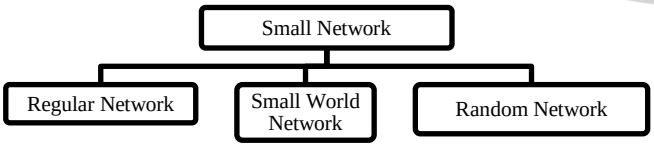


FIGURE 1: CLASSIFICATION OF SMALL NETWORK

*Small World Network* – network whose degree of distribution follows a power law

*Homophily Network* – network that can be referred as birds of a feather.

The figure 1 represents the classification of small networks of social networks.

## IV. COMMON TERMS AND CONCEPT IN SOCIAL NETWORK ANALYSIS

There are some common concepts of social network that are very important and useful to understand about the structural forms of social networks [4]

*Dyad:* A pair of actors (connected by a relationship) in the network

*Triad:* A subset of three actors or nodes connected to each other by the social relationships.

*Ties:* Ties or links connect two and more nodes in graph. Many human behaviors such as advice seeking, information-sharing and lending money to somebody are directed ties while co-memberships are examples of undirected ties [

*Betweenness:* The extent to which a node lies between other nodes in the network. These measures take into account the connectivity of the node's neighbors, giving a higher value for nodes which bridges clusters.

*Centrality:* The measures of centrality identify the most prominent actors, especially the star or the "Key" players that is those who are extensively involved in relationships with other network members. The most important centrality measures are Degree centrality, Betweenness centrality and closeness centrality.

*Closeness:* The degree of an individual is near all other individuals in a network (directly or indirectly). It reflects the ability to access information through the "grapevine" of network members. Thus, closeness is the inverse of the sum of the shortest distances between each individual and every other person in the network.

*Clique:* A clique in a graph is a sub-graph in which any node is directly connected to any other node of the sub-graph.

*Clustering Coefficient:* A measure of the likelihood that two associates of a node are associates them. A higher clustering co-efficient indicates a greater 'cliquishness'

*Cohesion:* The degree to which actors are connected directly tied to every other individual, 'social circles' if there is less stringency of direct contact, which is imprecise, or as structurally

*Density:* Density is a measure of the closeness of a network. Given a number of nodes, the more links, between them, the larger the density.

*Path length:* Nodes or actors may be directly connected by a line, or they may be indirectly connected through a sequence of lines. A sequence of lines in a graph is a walk and a walk in which each point and each line are distinct called a path. The length of a path is measured by the number of lines which makes it up.

*Reach:* The degree any member of a network can reach other members of the network.

In addition to social network extraction, there are other measurements that can be used for social network analysis as well. For example, degree centrality in a social network is used to measure the betweenness and closeness of the social network. [14].

 DOI : 10.35291/2454-9150.2020.0017 

## V. TYPES OF SOCIAL NETWORKING PORTALS (SNP)

This section attempts to order the current range of social networking portal. They are basically organized around the profile, content, micro-blogging, video sharing, etc. [5]

*Profile-based SNP* – primarily organized around members profile pages. E.g. Bebo, Facebook, MySpace.

*Content-based SNP* – the user's profile remains an important way of organizing connections, but plays a secondary role to the posting of content. E.g. Flickr – photo sharing site, YouTube – video sharing site.

*White label SNP* – users can create their own small scale social networking sites which support specific interests, events or activities. E.g., People Aggregator, Ning

*Multi-User Virtual Environment* – a virtual world, allows users to interact with each other's avators – a virtual representation of the site member. E.g. SecondLife

*Mobile SNP* – offers mobile phone versions of their services, allowing members to interact with their networks via phones. E.g., MySpace, Twitter, Face book.

*Micro-blogging / Presence updates* – Micro- blogging services allows the users to publish short messages publicly or within contact groups. E.g., Twitter

*Academic SNS* – Each site offers its own combination of tools and capabilities to support research activities. E.g., Academic.edu, Research gate, Mendeley

## VI. COMMON APPROACHES ON SOCIAL NETWORKS

There is little common approach through which social networks can be approached. [11]

TABLE 1: APPROACHES OF SOCIAL NETWORKS

| Common Approach | Classification under common approach |
|---|---|
| Graph Theory | Community Detection |
| | Recommender System |
| | Semantic Web |
| Opinion Analysis | Aspect-Based / Feature-Based Opinion Mining |
| | Homophily Clustering |
| | Opinion Definition and Opinion Summarization |
| | Opinion Extraction |
| Sentimental Analysis | Sentiment Orientation |
| | Product Rating & Reviews |
| | Reviews and Rating (RnR) Architecture |
| | Aspect Rating Analysis |
| | Sentiment Lexicon |
| Unsupervised Classification | Semi-Supervised Classification |
| | Supervised Classification |
| Topic Detection and Tracking (TDT) | Transaction – based Rule Change Mining (TRCM) |

## VII. RESEARCH OPTIONS OF SOCIAL NETWORK PORTALS

Research on social network can be divided into following categories as network data, research focus and design issues [17 ].

*Based on the network data:*

The figure 2. Depicts the classification of social network based on the network data.

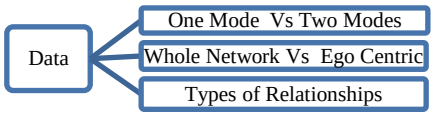


FIGURE 2: CLASSIFICATION BASED ON NETWORK DATA

One mode is a communication between a people to person. Two modes is a communication between a person and an event.

Whole network / complete / social centric focuses on large group of people. Tracks the quantifies relationships between people in a group and studies about the patterns of interactions show theses patterns affect the group as whole.

Network data based on relationship may be based on communications, affection, advice, proximity, power and multiplicity of relationship.

*Based on Research Focus:*

The social networks can also be classified on the research focus at individual level, system level, network structure and computational model.

The outcome of the individual level is impact of centrality, degree, reachability, betweenness[burst 2005]. The impact of system level research is structure of overall network that density of connection, impact of centralization on signal aggregate.

Network level research focuses on Middle level features (e.g., triads, quads, etc.,), description of overall structures i.e., degree of separation, clustering for small world network, degree distribution for scale free network. The figure 3 infers the classification of social network based on the research focus.

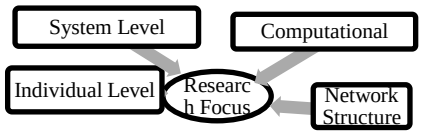


FIGURE 3: CLASSIFICATION BASED ON RESEARCH FOCUS

*Based on the Research design:*

Research design can be classified into statistical and design challenges. Statistical challenges deals with interdependencies of observations and design challenges deals with longitudinal data. The figure 4 explains the diagrammatic view of research design classification.

 DOI : 10.35291/2454-9150.2020.0017 

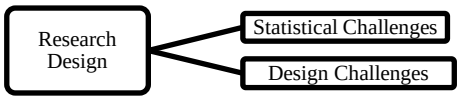


FIGURE 4: CLASSIFICATION BASED ON RESEARCH DESIGN

## VIII. DISCUSSION

Wagner et al., [15] said that with the emerging usage and popularity of online social media services, people now have accounts on several and diverse services like Twitter, YouTube, Facebook, and LinkedIn. Popularity of academic social networking sites (SNSs) is increasingly exponentially. Academic users at different levels of their career and form different disciplines are today becoming more interested in them. They are using those academic SNSs for several goals, and thus in different ways constituting some patterns of opinions. This paper aimed to describe usage patterns of an academic SNS (namely Academica.edu) through different academic user groups. To achieve this, user profile data were gathered directly from Academic.edu Website. Recommender systems are one way of helping users to deal with abundant data by recommending items that match user's personal interests. Humphreys [1] stated that social network applications have now been migrated from computer to the mobile phone, network information and communication can be integrated into public space: and theses new services developed for mobile phone allows users to create, develop and strength their social ties. Mobile devices have an enormous advertisement potential. Besides being extremely popular, most peopled carry them all the time, enabling personalized advertising [9].

## IX. TYPES OF SOCIAL MEDIA [19]

1) Social Networks – Connect with people(Face book, Twitter, LinkedIn)

2) Media sharing networks –share photos, videos and other media ( Instagram, Snapchat, YouTube, Vimeo)

3) Discussion forums – share news & ideas (Reddit, Quora, Digg)

4) Bookmarking & content circulation networks – discover, save and share new content (Pinterest, Flipboard)

5) Consumer review networks – find and review business (Yelp, Zomato, TripAdvisor)

6) Blogging and publishing networks – publish content online(WordPress,Tumblr,Medium)

7) Internet-based networks – share interests and hobbies(Goodreads, Houzz,Last.fm)

8) Social shopping networks – shop online(Polyvore, Etsy, Fancy)

9) Sharing economy networks - trade goods and services(Airbnb,Uber,Taskrabbit)

10) Anonymous social networks – communicate anonymously (Whisper, Ask.fm,After school)

## X DATA MINING TECHNIQUES FOR SOCIAL MEDIA

As per the study presented by [19], 19 data mining techniques which are applied by the researchers in the field of social media are listed below

- AdaBoost
- Artificial Neural Network (ANN)
- Apriori
- Bayesian Networks (BN)
- Decision Trees (DT)
- Density Based Algorithm (DBA)
- Fuzzy
- Genetic Algorithm (GA)
- Hierarchical Clustering (HC)
- K-Means
- k-nearest Neighbors (k-NN)
- Linear Discriminant Analysis (LDA)
- Linear-Regression (Lin-R)
- Logistic Regression (LR)
- Markov
- Maximum Entropy (ME)
- Novel , Wrapper
- Support Vector Machine (SVM)

## XI. OVERVIEW OF SOCIAL MEDIA RESEARCH METHODS

Content analysis – used for labeling audio, text, and/or visual communication from social media in a systematic manner, and can produce numeric output.

Thematic analysis – the process of locating patterns within coding, data familiarization and developing and revising themes.

Social network analysis – social net work analysis can be used to map and measure the relationship between WebPages, individuals, organizations, and information and/or knowledge entities.

Machine learning – a type of artificial intelligence which allows the computer to learn with being programmed. The task of human labeling a subset of data and make the computer to learn and the remaining was coded to learn.

Semantic analysis (Linguistics)- the analysis which examines the meaning of the language used and the relationships that occurs between the words, phrases and clauses.

## XII. AN OVERVIEW OF TOOLS FOR RETRIEVING SOCIAL MEDIA DATA

| Tool | OS | Download and/or access from | Platforms* |
|---|---|---|---|
| Audiense | Web-based | https://audiense.com/ | Twitter |
| Brand24 | Web-based | https://brand24.com/features/#4 | Twitter, Facebook, Instagram, Blogs, Forums, Video |

 DOI : 10.35291/2454-9150.2020.0017 

| | | | |
|---|---|---|---|
| Brandwatch | Web-based | https://www.brandwatch.com/ | Twitter, Facebook, YouTube, Instagram, Sina Weibo, VK, QQ, Google+, Pinterest, Online blogs |
| Chorus (free) | Web-based | http://chorusanalytics.co.uk/chorus/request_download.php | Twitter |
| COSMOS Project (free) | Windows & MAC OS X | http://socialdatalab.net/software | Twitter |
| Echosec | Web-based | https://www.echosec.net | Twitter, Instagram, Foursquare, Panoramio, AIS Shipping, Sina Weibo, Flickr, YouTube, VK |
| Followthehashtag | Web-based | http://www.followthehashtag.com | Twitter |
| IBM Bluemix | Web-based | https://www.ibm.com/cloud-computing/bluemix | Twitter |
| Keyhole | Web-based | https://keyhole.co/ | Twitter, Instagram, Facebook |
| Mozdeh (free) | Windows (Desktop advisable) | http://mozdeh.wlv.ac.uk/installation.html | Twitter |
| Netlytic | Web-based | https://netlytic.org | Twitter, Facebook, YouTube, RSS Feed |
| NodeXL | Windows | https://www.smrfoundation.org/nodexl/ | Twitter, YouTube, Flickr, Wikipedia |
| NVivo | Windows and MAC | http://www.qsrinternational.com/product | Twitter, Ability to import |
| Pulsar Social | Web-based | http://www.pulsarplatform.com | Twitter, Facebook topic data, Online blogs |
| Social Elephants | Web-based | https://socialelephants.com/en/ | Twitter, Facebook, Instagram, YouTube |
| Symplur (Healthcare focus) | Web-based | https://www.symplur.com/ | Twitter |
| SocioViz | Web-based | http://socioviz.net | Twitter |
| Trendsmap | Web-based | https://www.trendsmap.com | Twitter |
| Trackmyhashtag | | https://www.trackmyhashtag.com/ | Twitter |
| Twitonomy | Web-based | http://www.twitonomy.com | Twitter |
| Twitter Arching Google Spreadsheet (TAGS) (free) | Web-based | https://tags.hawksey.info | Twitter |
| Visibrain | Web-based | http://www.visibrain.com | Twitter |
| Webometric Analyst (free) | Windows | http://lexiurl.wlv.ac.uk | Twitter (with image extraction capabilities), YouTube, Flickr, Mendeley, Other web resources |

*Some tools may allow access to other platforms and the ability to import your own data.

## XIII. DATA MINING TECHNIQUES FOR SOCIAL MEDIA

There are various data mining techniques have been developed to overcome the problems such as noise, size and dynamic nature of social media data. The different types of techniques are as follows: [24]

1) Unsupervised classification
    a. Sentiment lexicon
    b. Sentiment orientation
    c. Opinion definition and summarization
    d. Basic clustering techniques
    e. Opinion extraction
2) Semi-supervised classification
3) Supervised classification
    a. Support vector machine
    b. Neural networks
    c. Navies Bayes
    d. K-nearest neighbor
    e. Decision trees
    f. CHAID (Chi-Square Automatic Interaction)
    g. Text mining

## XIV ISSUES IN MINING SOCIAL MEDIA

1) Community analysis
2) Sentiment analysis and opinion mining
3) Social recommendation
4) Influence modeling
5) Information diffusion and provenance
6) Privacy, security and trust

## XV. FINDINGS

The main research focus of this is social media mining. Here the most frequently used social media tool for retrieving data, social media mining classifiers such as SVM, NB, DT and LR were compared along with the performance metrics such accuracy, precision, recall and ROC used. In most of the case the existing dataset were used and the classifiers are chosen depending upon suitations. No one common classifier has performed good for all the cases. The tools Weka, R, MATLAB were used for evaluation

## XVI CONCLUSION

This paper gives us the clearly overview about the Social media, social networks, classification of social network, types of social network, common terms and approaches, data mining. The following section provided the related works towards social network, data mining and it mixture. At the end it has provided an insight idea about the various data mining algorithms and how they are implemented, compared based on the performance metrics. Due to dynamism, unstructured type of data researchers have more opening in the fields social media mining with various

 DOI : 10.35291/2454-9150.2020.0017 

approaches such as machine learning, deep learning, data science, big data, etc,. Hopefully the future work can be exploded with various algorithms and implication of those with real time data.

## REFERENCES

[1] Abba almu et al., "Effect of Mobile Social Networks on Secondary School Students" International Journal of Computer Science and Information Technologies, Volume 5(5)2014,6333-6335

[2] Borgatti, S.P,: "2-Mode concepts in Social network analysis" Encylopedia of complexity and system science, 8279-8291,2009

[3] Chen, Z, S, Kalashnikov, D. V. and Mehrotra, S. Exploring context analysis for combining multiple entity resolution systems In Proceedings of the 2009 ACM Internationl Conference on Management of Data (SIGMOD'09), 2009

[4] Hilal Ahmad Khanday, Dr. Rana Hashmy," Exploring Different Aspects of Social Network Analysis using Web Mining Techniques" International Journal of Emerging Trends and Technology in Computer Science, Volume 4, Issue 2, March-April 2015, ISSN 2278-6856

[5] http://fraser.typepad.com/socialtech/files/social-networking-overview.pdf

[6] http://www.hks.harvard.edu/netgov/files/NIPS/Lazer_NIPS2008_workshop.pdf

[7] Jin, Y. Z., Matsuo, Y., and Ishizuka, M. "Extracting Social Networks among Various Entities On the Web" In *Proceedings of the Fourth European Semantic Web Conference, 2007*.

[8] Kaplan, A M and Haenlein, M : Uesers of the World Unite! The Challenges and opportunities of Social media. Science direct,53,59-68,2010.

[9] Laura Marcia Villalba Monne,"A Survey of Mobile Social networking", Helsinki University

[10] Liu, B.: Sentiment analysis and opinion Mining. AAAI-2011, San Francisco, USA, 2011.

[11] Mariam Adedoyin-Olowe, Mohamed Medhat Gaber, "A survey of Data Mining Techniques for Social Network Analysis"

[12] Matsuo, Y., Tomobe, H., and Nishimura, T. "Robust Estimation of Google Counts for Social Network Extraction" In *Proceedings of Twenty Second Conference on Artificial Intelligence (AAAI-07)*, July 22-26, 2007, Vancouver BC Canada.

[13] Preeti Srivastava, "Social Networking & Its Impact on Education – System in Contemporary Era", International Journal of Information Technology Infrastructure Volume 1, No.2, November-December 2012, ISSN 2320 2629

[14] Raju. E, Sravanthi.K,"Analysis of Social Networks using the Techniques of Web Mining", IJARCSSE,Vol. 2, Issue 10, October 2012,443-450.

[15] Sanjeev Dhawan, Kulvinder Singh, Vandana Khanchi, " Critical analysis of Social Networks with Web Data Mining", IJITKM Special issues (ICFTEM-2014) May 2014 pp 107-111 (ISSN 0973-4414)

[16] Thompson, J B.: *Media and modernity: A social theory of the media*. John Wiley & Sons, 2013.

[17] Vedanayaki .M, " A study of Data Mining and social Network Analysis", Indian Journal of Science and Technology, Volume 7(57), 185-187, November 2014, ISSN (Print) : 0974-6846

[18] Wen-jun, S., and Hang-ming, Q. "A Social Network Analysis on Blogspheres" *In Proceedings of the 15th IEEE International Conference on Management Science and Engineering*, 2008, pp.1769 - 1773, Long Beach, CA, USA.

[19] https:blog.hootsuite.comtypes of social media

[20] MohammadNoor Injadat, Fadi Salo and Ali Bou Nassif, Data Mining Techniques in Social Media: A Survey, Neurocomputing

[21] https://blogs.lse.ac.uk/impactofsocialsciences/2019/06/18/using-twitter-as-a-data-source-an-overview-of-social-media-research-tools-2019/

[22] https://searchunifiedcommunications.techtarget.com/answer/Whats-the-difference-between-social-media-and-social-networking

[23] https://learn.g2.com/social-media-data-mining

[24] Shailja Bhardwaj,"A study of data mining techniques in social media" International Research Journal of Engineering and Technology, Volume 5,issue 5, May2018, Page 2713

[25] S.G.S Fernando et.al "*Empirical Analysis of Data Mining Techniques for Social Network Websites*" in COMPUSOFT, An international journal of advanced computer technology, Volume-III, Issue-II PP:582-592, 2014.

[26] Aarti Sharma, Rahul Sharma,Vivek Kr. Sharma,Vishal Shrivatava, " *Application of Data Mining – A Survey Paper*", in (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2) , PP: 2023-2025, 2014.

[27] Wakade, Shruti et al. "Text Mining for Sentiment Analysis of Twitter Data." (2012).

[28] Sharma, Seema & Agrawal, Jitendra & Agarwal, Shikha & Sharma, Sanjeev. (2013). Machine learning techniques for data mining: A survey. 2013 IEEE International Conference on Computational Intelligence and Computing Research, IEEE ICCIC 2013. 1-6. 10.1109/ICCIC.2013.6724149.

## AUTHORS PROFILE

**Mrs P. Gunasundari** has secured her MCA.,M.Phil., Degree in Computer Science. Currently she is a research scholar in Bharathiar University in Coimbatore. She is working towards her Ph.D. in the area of Data Mining, Social Networking.

**Dr. (Mrs.)R.Thirumalai Selvi** is currently the research supervisor and Assistant Professor in the Department of Computer Science, Govt. Arts College (Men) (Autonomous), Chennai. She has over 20 years of experience in various arts and science colleges and as a research supervisor. she is guiding many PhD students registered under various universities.

 DOI : 10.35291/2454-9150.2020.0017 