REVIEW ON SOCIAL NETWORK ANALYSIS IN DATA MINING

S.Kasthuri[#1], Dr.A.Nisha Jebaseeli[*2],

#1 Research Scholar, Department of Computer Science, Govt. Arts & Science College,
Lalgudi, Trichy, Tamilnadu, India. vishnugka@yahoo.co.in

**2 Assistant Professor and Head, Department of Computer Science, Govt. Arts & Science College,*
Lalgudi, Trichy, Tamilnadu, India. nishamarcia@gmail.com

Abstract: In social network, an exceptional amount of data is presented because of the global utilization of social media analysis, which is of interest to several branches of study like business, sociology, entertainment, psychology, news, politics, and other cultural aspects of societies. Social network analysis has become a very popular field of modern research because it is highly useful for several applications. Data mining techniques analyzes the enhancing reliance on social networks calls to facilitate reforming the unstructured data and place them within a systematic pattern. In this paper, the literature survey of different data mining techniques used by researchers and suggest how this survey and study of the data mining approaches can benefit the importance of social network analysis for business intelligence.

***Keywords:* Data Mining Techniques, Social Network Analysis, Sentiment Prediction Technique, Rapid Clustering Method, Fuzzy c-means clustering algorithm and Random Decision Tree.**

I. INTRODUCTION

In recent times, social networks become more popular online platform to communicate and share the information and it consists of members (characterized as nodes on a network graph) that share one or more specific kinds of special interests, like financial, ideas, visions, values, like, dislike, exchange, friends, conflict, trade, web links, and so on (denoted as links on a network graph) [1]. Online social networks are built on the concept of tradition networks, but without relying on the face-to-face initial. Boyd and Ellison (2007) describe as web-based services (social network sites) that facilitate people: Initially, make a public profile contained by a bounded system; and then, communicates with a list of other users to share a connection; and finally, analysis and traverse their list of connections those who made by others within the system [2]. In order to survey the different aspects of social networks it is now required to design some good business model frameworks by using advanced data mining tools and techniques. Data mining process have to identify the patterns within a massive amount of dataset to obtain useful information. It involves the data to be consolidated in a data warehouse. Various data mining algorithms are applied so as to find out meaningful insights within the data. These insights are utilized for taking timely and accurate decisions by market analysis, business process, various companies, etc. Data mining is also called knowledge discovery and data discovery [3]. Analyzing the social network data has the potential of revealing information of great value. Social network analysis provides a systematic technique to identify analysis, visualize and support processes of knowledge sharing in social networks. It could be determined various aspect of queries like network dynamics, how the information flows, and important members or links. Many data mining methods have been involved in order to overcome the problems like noise, size, and dynamic nature of the social media data. Due to the huge size of data in the social media, an automatic data processing is required so as to analyze the data within a particular time span. The dynamism in the social media data shows the way to the rapid evolution of the data sets over time; such dynamic data can be effortlessly handled by various data mining techniques [4]. In this paper, it presents the literature analysis of available data mining techniques to mine social network data.

II. LITERATURE REVIEW

In this literature survey has been clearly observed that several researchers perform how these interests of the users can be analyzed for future conclusions and utilizations and many different methods proposed.

2018, Mohammed H, et al., [5] introduced hybrid system which utilizes text mining and neural networks for sentiment classification and also compared the performance of deep learning algorithms and different machine learning techniques. The dataset used in this work contains more than 1 million tweets collected in five domains. The system was trained using 75% of the dataset and was tested using the remaining 25%. The experimental results illustrate a maximum precise rate of 83.7% with sensitivity 87.1% and specificity 79.3%, and that proves the efficiency of the hybrid learning approach used by the system over the standard supervised approaches.

2018, Cem Baydogan, et al., [6] implemented Konstanz Information Miner (KNIME) utilized for open source data analytics, which is a powerful data mining and machine learning tool with its features and several visualization tools, was used on twitter data. Ten thousand Twitter data were used, the information access and interpretation steps used in the literature have been investigated in detail. The sentiment analysis work was conducted on Twitter data. Obtaining Twitter data, clearing and transforming data into numerical form, and then extracted into the meaningful results and interpreting them are performed. Machine learning algorithms are performed by using KNIME software. Decision Tree Learner and K-NN algorithms provided the sensitivity, accuracy, and selectivity values that have been observed in two different experimental sets. The results were obtained and compared with the help of experiment result tables.

2018, Atharva Patil, et al., [7] presented a methodology for performing sentimental analysis in restaurants on the opinions and determine their polarity references. The proposed methodology analyze the opinions that are posted on different social media platforms and apply the knowledge of Data mining combined with Sentimental Analysis with the support of a modified K-means algorithm with dynamic thresholding. Followed by clustering on the positive and negative feedbacks obtained from the previous process, to identify the broad topics of that organization that the feedbacks target. Finally, the resultant supports to improve their current processes based on the opinion received. The opinions are gathered from different sources to suffice the needs.

2017, Ardra, et al., [8] analyzed the behavioral pattern of youth that can be used to find the trending interest using the data from twitter. Companies or businesses know how to find out the positive and negative opinions of their brands; can evaluate their overall performance, especially about their online presence. Mining and sentiment analysis helps to improve company sales and their marketing strategies. Individuals are also able to obtain certain benefits particularly when making a brand for themselves. Famous authors celebrities, Artists, and all other popular individuals benefit from sentiment analysis. They get to know how people react, how they inspire the public, both negative and positive, to any recent move and the general attitude of people towards them. The experimental result of this approach illustrates the feedback of youth/publics on various matters and persons in different areas and verified the effectiveness. This technique proposed may support people to take decision and performance improvement.

2017, Firoj Fattulal Shahare, et al., [9] developed an innovative technique for the news events based on the social media big data to do the sentiment analysis. Aim main target to process on News data and find out what reaction from this data in the form of emotion. Proposed a replacement methodology to try and do the sentiment analysis for news data a lot of specially ,supported the social news and social media information (specifically emotions text), a Levenshtein technique is made to together categorical its emotions and linguistics, that lays the muse for the happening sentiment analysis. The word feeling computation algorithmic rule is planned to get the beginning word feeling that area unit more refined through the quality emotion wordbook. With the word emotions are able to reason each sentence sentiments. The proposed method utilizes Levenshtein algorithm and Naïve Bayes to determine the sentiment into different types from given social media news data. Levenshtein algorithm provides easy way to text processing on data. Its work fast and provide maximum level of accuracy to processing huge amount of data.

2017, Prof. Shilpa V, et al., [10] analyzed social networking data set for pattern recognition as it has emerging application areas in data mining. Facebook 100 dataset and applied Bisecting K-Means algorithm on it, with the intention that would get better clustering outputs. Bisecting K-Means represents first bisects of the data into two elements and chooses the part with larger number of elements after that apply clustering on it again. This process remains till N Number of clusters. The sample dataset applied to find out the desired results. The comparison of Bisecting K Mean algorithm and other data mining algorithm are capable of finding out the different pattern from social networking dataset. To find out the best possible clustering for the FB 100 dataset and in the most efficient manner, and it highlights the formation of association rules between the attributes and explores the association rule between discovers the patterns and different parameter.

2017, Mrs. Kulkarni Varsha, et al., [11] evaluated and compared two data mining algorithms such as Support Vector Machine (SVM) and Polarity Algorithm. The proposed method utilizes sequence of data mining algorithms such as Stop words Cleaning, Tweet collection, Frequency computation, Tokenization, College based Sentiment analysis, Tweet based Sentiment analysis, feature Based frequency, Low negative polarity, maximum Neutral polarity, Rank Colleges based on maximum Positive polarity, and then Rank the college based on Tweets. The drawback of survey monkey and glass door applications overcomes collaborative filtering and a manual process. To analyze the real time data and performs a sequence of SVM and sentiment analysis to produce conclusions of which Engineering College is better way to using Twitter data. The experimental evaluations have been performed with single feature and multiple features such as Research, Infrastructure, Placements, etc. and showed the efficiency of real time data.

2017, D.I. George Amalarethinam, et al., [12] presented the effective sentiment prediction technique in Big Data, using Spark. Naïve Bayes prediction algorithm is parallelized using PySpark. The simulation results obtained from the proposed work were analyzed to exhibit high levels of scalability in terms of both accuracy and time, in such case it observed that even with the enhance in the data size, the time taken for processing showed very less variance. A comparison of Naïve Bayes and SVM and the proposed technique has been identified to provide better accuracy levels. The major benefit of the proposed technique does not involve any additional corpus, therefore making the system less dependent. The proposed technique examines the text in terms of positive and negative opinions. The accuracy comparison of the proposed technique was also carried out and it was identified that the proposed parallelized Naïve Bayes technique offers both faster and more accurate results compared to sequential Naïve Bayes technique and SVM technique.

2016, Vidhyabhushan Dasondi, et al., [13] proposed automatic text classification technique helps to recognize the positive feedback and the negative feedback of the communicated text data. The proposed technique is used to apply on the twitter dataset both designing and modeling of the proposed classifier. To evaluate the proposed technique analyzed each word in both the available classes. Hence the probability of words in a class is computed first and then the importance of the word for generating a sentence is also measured. Using the computed probability the combine weights are developed and that provide how effectively a word is occurred in positive sentences and negative sentences. Finally, the developed data model for the classification of real world text is performed. The proposed technique is implemented using the JAVA technology and also their performance corresponding to the error rate, accuracy, and the space complexity of the proposed method is evaluated and reported.

2016, Ahmed Alsayat, et al., [14] proposed an innovative technique to examine the social media data. The method used K-Means algorithm along with GA (Genetic algorithm) and Optimized Cluster Distance technique to cluster the social media community based on leadership, follower and attitude scores. The experimental results of the proposed algorithm outperform compare than other existing techniques. Using this method presents a use-case of the method to further describe user community by getting more insights from clustering results and assigning self-explanatory labels to each cluster. The performance of proposed algorithm is to valuate various cluster validation metrics. The experimental analysis illustrates that the proposed method gives a novel use-case of grouping user communities based on their activities and also provides better clustering results. Proposed approach is optimized to provide scalable performance for real-time clustering of social media data.

2016, Ms .Varsha Bairagi, et al., [15] proposed fuzzy rule based text classification technique by which the bulling text or social cheating text is recovered from the communication. During investigation that is found the fuzzy based classification technique apt for analyzing text. The proposed technique is considered as a binary text classification technique that first analysis the previous patterns from the available samples data. Initially, the input training samples are processed for finding the fuzzy probability distribution for both the classes. Using the evaluated probability the test samples or new input text are categorized. The proposed technique is implemented using the visual studio development environment and comparative analysis between the traditional classification techniques and the proposed classification technique were performed efficiently. For comparing and evaluating the performance two additional classifiers like Bays classification and the KNN classification technique is implemented. The performance analysis of the implemented classifiers is performed some metric levels such as the error rate, accuracy, memory consumption and time consumption. In accordance with the obtained performance the proposed classification technique is accurate and efficient performance than the traditional classification technique. Therefore that can be adoptable for other text based classification technique.

[2018], Anna Cinzia Squicciarini et al. [16] developed and analyzed Adaptive Privacy Policy Prediction that has a free privacy mechanism for personalized policies. The proposed system forces the user to upload the images derived from the individuality and satisfaction of the metadata. It includes two components: Adaptive Privacy Policy Prediction for Core and also the same system with Social. For example, if a customer is uploading an image and it will be immediately transmit to the Adaptive Privacy Policy Prediction with Core. This core component arranges the images and determines whether there is a need for demand for the other social component. The judgment policy being mistaken is the disadvantage because of not having knowledge of the Meta data information regarding the images. Both the Meta data and log data information are guided by the classification technique.

2015, Paulo R, et al., [17] presented a Conversation Classifier based on Multiple Classifiers, to detect Life Events on Social Media. The experimental evaluation on two datasets, one in English and another in Portuguese, has demonstrated that the proposed system combining classifiers is promising, but the higher diversity in the pool created from the English set makes the gap of accuracy much wider when compared with single classifiers. The experiments show that multiple classifiers are promising for this issue, being able to present an increase of about 45% in the F-Score.

2015, Sharmishta Desai, et al., [18] implemented decision tree algorithm and it has suitable for social media data because it provides more accurate results than other techniques. It has proved with results that decision tree algorithm is more. The decision tree algorithm has been implemented and compared with other machine learning algorithms like Adaboost, NaiveBayes etc. The proposed algorithm gives accurate results as compared to other algorithms. Moreover, different decision tree technique of RDT (random decision tree) algorithm performs better than other decision tree algorithms such as C4.5 or ID3. Different phases of social media data mining are also identified and suggested that data cleaning is one of essential phase for social data mining.

2015, Puneet Garg, et al., [19] analyzed LinkedIn data with different clustering techniques and address some common issues such as similarity computation and normalization of messy LinkedIn data. Moreover normalize geographic data present in dataset with the support of geo-coordinates, after that progressed on geo clustering of LinkedIn connections and their visualization on Google Map. LinkedIn data is extracted with the approval of LinkedIn API and the normalized data of removing redundancies. Additionally, the normalized data performed in accordance with locations of LinkedIn connections utilizing geo coordinates afforded by Microsoft Bing. After that, clustering process of this normalized data set is finished the process according to company names, job title, and geographic locations analyzing through Hierarchical, K-Means clustering, Greedy algorithms and clusters are visualized to perform a better insight into them.

2014, Fakhri Hasanzadeh, et al., [20] proposed method for social network to detect communities by performing with these techniques of clustering, analysis of communications. Initially, the proposed method focuses on using ontology and existing subject in social network, separate the network to clusters and finally the proposed techniques utilized for the analysis of communication between individuals, find communities. The experimental results illustrate that has been proved a better accuracy to detecting communities and keeping relevant communications around one topic inside a community.

2014, Brinal Colaco, et al., [21] focused on the process of making predictions of private information using public information. In a social network, users have profile data that make certain aspects of their personality predictable. The experimental results show that how Mamdani FIS (Fuzzy Inference System) is an efficient technique to make prediction more accurately leading to accurate private information inference. This technique is compared with the Naïve Bayesian technique on the same dataset. And from the results obtained, it is proved that, before and after sanitization, the fuzzy inference system gives better precision and accuracy. The proposed system checks whether the user profile is susceptible to inference attacks. It will propose sanitization techniques to avoid these attacks. The proposed system will improve the user’s experience of using the social network. With many third part application trying to obtain information of the users, this application will decrease the accuracy of the classifiers that infer private information. This system will try to reduce the inference attacks on social networks.

2013, Hwi-Gang Kim, et al., [22] proposed the ratio of word frequency as a measure to sense social hot topics of each day. Twitter streaming data is used to find out social hot topics and sensation news around the environment. The proposed method contains non-topic keywords that have been used emotional words while recognizing topic words. Moreover, the geographic communities for topics by the proposed method. Author visualized the ratio of word frequency for each topic. Geographic clustering evaluation results were performed using the GUT (Google Fusion Table). The proposed technique provides a simple way but useful approach to analyze real-time streaming data and discover geographic communities.

2013, Neha Mehta, et al., [23] analyzed the web information repository experiential because of its huge size of dataset, but the significant information is hard to found out an unproblematic task. Different penetrating and web mining methods are being utilized by the current day using search engine for the basis of information repossession from the websites. One of the potential techniques is Web document clustering to provide better the effectiveness in information discovery process. The conventional web mining, method of web mining has complexity in managing confront posed by the collecting of data which is uncertain and unclear. FCM technique has the probable to preserve such type of situations competently. In this paper, summarizes the dissimilar characteristics of web data, the web mining basics and boundaries of existing web mining methods. The submission of use of Fuzzy logic with web mining is to highlight its significance in information retrieval. A relative study of dissimilar fuzzy clustering techniques with the predictable clustering technique has been discussed.

2010, J. Prabhu, et al., [24] proposed the clustering technique known as the RCM (Rapid Clustering Method) that utilize Subtractive Clustering combined with FCM clustering along with a histogram sampling technique to provide effective and quick result explorations for large sized datasets. RCM has been applied to cluster the dataset and then analyze the performance and traits in a social network and also used to develop the cross-selling practices using quantitative association rule mining. The proposed clustering algorithm performs better than SCM-FCM and K-Means in a lesser time period and also by preserving the level of accuracy.

III. CONCLUSION

Data mining provides a variety of systems for identifying cooperative learning from vast datasets and an extensive range of methods for detecting useful knowledge from massive datasets such as patterns, trends, and rules. Different data mining methods have been used in social network analysis as focused on this paper. In this paper, the current evaluation and update of social network analysis were discussed and reviewed based on different aspects of social network analysis. This work aids to comparative studies of the techniques and idea of web mining for social networks analysis, and reviews the connected literature concerning web mining and social networks. Data mining techniques have been faces many challenges during this analysis area to be resolve with aggressive improvement.

IV. REFERENCES

[1] Zafarani R, Abbasi MA, Liu H. Social Media Mining an Introduction. Cambridge University; 2014.

[2] 3. Xu G, Zhang Y, Li L. Web Mining and Social Networking Techniques and applications, 1st edition. Springer; 2011.

[3] M. Vedanayaki, “A Study of Data Mining and Social Network Analysis”, Indian Journal of Science and Technology, Vol 7(S7), 185–187, November 2014.

[4] Adedoyin-Olowe, M., Gaber, M. M., & Stahl, F. (2013). A survey of data mining techniques for social media analysis. arXiv preprint arXiv:1312.4617.

[5] Mohammed H. Abd El-Jawad, Rania Hodhod, Yasser M. K. Omar “Sentiment Analysis of Social Media Networks Using Machine Learning”, IEEE, 10.1109@ICENCO.2018.8636124.

[6] Cem Baydogan, Bilal Alatas,"Sentiment Analysis Using Konstanz Information Miner in Social Networks",IEEE, 2018

[7] Atharva Patil, Karan Bheda , Nishita S. Upadhyay, Rupali Sawant," Restaurant's Feedback Analysis System using Sentimental Analysis and Data Mining Techniques ",Proceeding of 2018 IEEE International Conference on Current Trends toward Converging Technologies, Coimbatore, India 978-1-5386-3702-9/18/2018 IEEE

[8] Ardra, Blessy Merin Varughese, Merline Susan Joseph, Preethi Elsa Thomas, Sherly K K," *Analyzing the Behavior of Youth to Sociality Using Social Media Mining*", IEEE, International Conference on Intelligent Computing and Control Systems ICICCS 2017

[9] Firoj Fattulal Shahare," Sentiment Analysis for the News Data Based on the social Media", International Conference on Intelligent Computing and Control Systems ICICCS 2017

[10] Prof. Shilpa V. Gajbhiye, Prof. Gaurav B. Malode," Enhancing Pattern Recognition in Social Networking Dataset by Using Bisecting KMean", IEEE, International Conference on Intelligent Computing and Control (I2C2), 2017

[11] Mrs. Kulkarni Varsha, Monica R, "Analyzing of Premier Institution using Twitter Data on Real-Time Basis", IEEE, International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS-2017)

[12] Dr. D.I. George Amalarethinam, V. Jude Nirmal ," Real-Time Sentiment Prediction on Streaming Social Network Data Using In-Memory Processing", World Congress on Computing and Communication Technologies (WCCCT), IEEE, 2017

[13] Vidhyabhushan Dasondi, Milap Pathak , Narendra Pal Singh Rathore, "An Implementation of Graph based Text Classification Technique for Social Media",IEEE, Symposium on Colossal Data Analysis and Networking (CDAN), 2016

[14] Ahmed Alsayat, Hoda El-Sayed," Social Media Analysis using Optimized K-Means Clustering", IEEE, SERA 2016, June 8-10, 2016, Baltimore, USA

[15] Ms .Varsha Bairagi, Dr.Namrata Tapaswi ,"Social Network Comment Classification using Fuzzy based Classifier Technique", IEEE, Symposium on Colossal Data Analysis and Networking (CDAN), 2016.

[16] Anna Cinzia Squicciarini, "Privacy Policy Inference of User-Uploaded Images on Content Sharing Sites", IEEE Transactions On Knowledge And Data Engineering, vol. 27, no. 1, January 2015.

[17] Paulo R. Cavalin ; Luis G. Moyano ; Pedro P. Miranda," A Multiple Classifier System for Classifying Life Events on Social Media", IEEE 15th International Conference on Data Mining Workshops, 2015.

[18] Sharmishta Desai, Dr. S.T.Patil, "Efficient Regression Algorithms for Classification of Social Media Data", EEE, International Conference on Pervasive Computing (ICPC), 2015.

[19] Puneet Garg, Rinkle Rani, Sumit Miglani," Mining Professional's Data from LinkedIn", IEEE Fifth International Conference on Advances in Computing and Communications, 2015

[20] Fakhri Hasanzadeh, Mehrdad Jalali , Majid Vafaei Jahan," Detecting Communities in Social Networks by Techniques of Clustering and Analysis of Communications ", 978-1-4799-3351-8/14©2014 IEEE

[21] Brinal Colaco, Shamsuddin S. Khan," Privacy Preserving Data Mining for Social Networks", IEEE, International Conference on Advances in Communication and Computing Technologies, 2014

[22] Hwi-Gang Kim, Seongjoo Lee, Sunghyon Kyeong, "Discovering Hot Topics using Twitter Streaming Data", IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2013

[23] Neha Mehta, Mamta Kathuria, Mahesh Singh, "Comparison of Conventional & Fuzzy Clustering Techniques: A Survey", International Journal of Recent Technology and Engineering(TM), 2013

[24] J. Prabhu and M. Sudharshan M. Saravanan and G.Prasad," Augmenting Rapid Clustering Method for Social Network Analysis, IEEE, International Conference on Advances in Social Networks Analysis and Mining, 2010.