A survey and perspective analysis of text mining techniques in social media web portals

[1]Suriyapavithra. K.V., [2]N.Mohamed Farook Ali

[1]M.Phil. Scholar, [2]Assistant Professor
[1]Department of Computer Science,
[1]Vidyasagar College of Arts and Science, Udumalpet, India

***Abstract*—Social media data analysis become a vital research and business activity in the text mining. Social Media sites such as Facebook, Twitter, LinkedIn and Google+ contain large volume of unprocessed raw data. By analyzing this data new knowledge can be gained. Most of the social media data are unstructured and also dynamic. So the traditional text mining techniques will not suitable to handle these unstructured data. In this paper we discuss about text mining, areas of text mining, social media data, text mining techniques. In this paper a survey of the works done in the field of social network data analysis and techniques followed to perform the text mining on the social network data. This survey identify and analyse the text mining techniques and approaches used to analyse the social media data and finds the base for the text mining research in unstructured social media data.**

***Index Terms*— Text mining, Social media, Social network data analysis**

I. INTRODUCTION

Data mining is a powerful tool that can help to find patterns and relationships within our data. Data mining discovers hidden information from large databases.[1] The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Social networks can be used in many business activities like increasing word-of-mouth marketing, marketing research, General marketing, Idea generation & new product development, Co-innovation, Customer service, Public relations, Employee communications and in Reputation management.

There are various data mining techniques such as Characterization, Classification, Regression, Association, Clustering, Change Detection, Deviation Detection, Link Analysis and Sequential Pattern Mining.

Text mining process performs the structuring the input text, deriving patterns within the structured data, and final evaluation and interpretation of the output in the following areas [2]:

a. Information Retrieval
b. Data Mining
c. Natural Language Processing
d. Information Extraction

a. Information Retrieval: Traces and recovers the specific information from the unstructured or semi structured texts.
b. Data Mining: Discovers the hidden and also unknown patters from the data to predict the behaviors and future trends.
c. Natural Language Processing: NLP is a part of text mining used to perform linguistic based analysis. NLP plays vital role in social media monitoring the user posted content.
d. Information Extraction: Extracting structured information from unstructured. In most of the cases, this activity includes processing human language texts by means of NLP.

II. BACKGROUND

Social network is a web portal which enables the users to interact with each others through comments, post messages in wall, images and videos. Data representation in social media consists of nodes and links in the data graph. Nodes denote entities and links denotes the relationship among entities. In the Facebook, friends, relatives and colleagues represents nodes and that are connected individually through links.

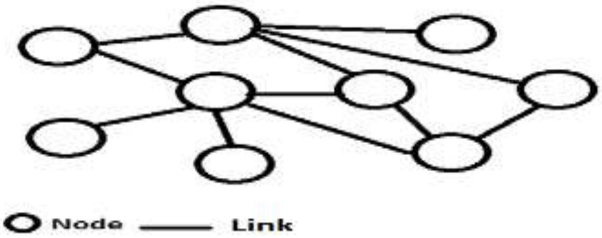

Fig.1. Links and nodes in social media

In this era, unstructured data like social media data gains more important and popularity than structured data [4]. Text mining makes it easy to obtain a meaningful and structured data from the irregular data patterns [5], [6], and [7]. It is really not an easy task for the computers to understand the unstructured data [8], [9] and make it structured.

In the organizations, most of the existing data in a text format, so text mining plays important role to extract the features from these semi or unstructured data [10]. Text mining is used for structuring the unstructured text data. In general, the data related to social media sites is not collected for the research purpose [11], it is mandatory to change the structure of the data coming from the social media. 80% of the available text on the web is unstructured while only 20% is structured [12].

III. LITERATURE REVIEW

From the last decade onwards social media have become not only popular but also affordable and universal communication means which has thrived in making the world a global village. Social media especially Twitter and Facebook play an energetic role in publicising the peoples feel about certain products / services, issues and events in many aspects of life.

Pang, B., Lee, L [13], Millions of people access social media sites such as Twitter, Facebook, LinkedIn, YouTube and MySpace to search out for information, breaking news and news updates. Most of the updates are often posted by unfamiliar people they have never and may never have contact with. Consequently information gathered on social media is sometimes used to make valuable decisions. While some groups are retrieving information from social media sites, others are posting information for the use of other internet users.

Stieglitz, S., & Dang-Xuan, L.[14], stated that social media portals provide a space to the persons to post their views, opinion and also interact with other persons. The best thing which these sites are doing is that it has become easy for the individuals to understand a particular person depending upon his or her activities. These activities allows to come closer to each others from various culture and region with better understanding of each other's reactions, sensitivities and areas of interest.

My Personality project [15] takes the 250 user instances from the facebook with 10,000 status updates. The study has the following two interconnected objectives: (1) having knowledge about the relevant personality-correlated indicators that presents user data implicitly or explicitly in Facebook, and (2) identifying the feasibility of prognostic character demonstration so that upcoming intelligent systems could be supported. The study emphasized on the promotion of pertinent features in a model, through which the enhanced output of the classifiers under evaluation could be observed.

Rahman,M.[16] investigated images for the advertisement of their products and for the decision-making process. For the recovery on Facebook user database, Facebook API performs Application Secret key and Facebook API Key are executed by Facebook API. As a result, WEKA files and data mining techniques are supported to collect certain data into the secondary database, while the text data is represented by the detached data.

A study by Al-Daihani, S. M., [17] applied the text mining approach on a large dataset of tweets. The complete Twitter timelines of 10 academic libraries were used to collect the dataset for this research. Nearly 23,707 tweets formed the total dataset, where there were 7625 hashtags, 17,848 mentions, and 5974 retweets.

A study by Mosley Jr, R. C [18] discussed the clustering technique, the execution of correlation and association analyses to social media. The investigation of insurance 68,370 tweets was carried out to assess this matter. Clustering analysis and Association analysis are applied in these tweets. Clustering analysis used to cluster the tweets based on their similarities or dissimilarities. Association Analysis is to discover the occurrences of particular composed words.

Kermanidis,K.L. [19] discussed the influence of tweets' sentiment on elections and the impact of the elections' results on web sentiment.

Ahmad N, Siddique [20] discussed psychoanalytic profiling of the tweets and analyse the personality behaviour of the users. Twitts are separated in dominance, influence, steadiness and compliances category and perform the analysis using Rapidminer and R language. The results shows that most of the tweets fall under dominance category.

Rizwana Parwin[21], People did not consider spellings and grammar while communicating with each other using social networking websites (Facebook, Twitter). It makes tedious task to extracting logical patterns with accurate information.

A Razia Sulthana et al [22] performs sentiment analysis in tweets on Hillary and Trump, linear regression approach predicts the polarity of the tweets better than SVM and Naive bayes. Linear regression with 10 fold cross validation applied to get the 85.23%.

IV. CONCLUSION AND SCOPE

This paper provides a more current evaluation and update of social network analysis research available. Literatures have been reviewed based on different aspects of social network analysis.

This paper studies the application of the techniques and concept of text mining for social networks analysis, and reviews the related literature about text mining and social networks. Social networks investigation carried out through the techniques of web mining is an interesting field of research. However, there are many challenges in this research field to be resolve with improvement

References

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

[2] https://data-flair.training/blogs/text-mining/

[3] Sukanya, M., & Biruntha, S. Techniques on text mining. InAdvanced Communication Control and Computing Technologies (ICACCCT), 2012 IEEE International Conference on (pp. 269-271). IEEE.

[4] Chakraborty, G., & Krishna, M. (2014). Analysis of unstructured data: Applications of text analytics and sentiment mining. In SAS global forum (pp. 1288-2014).

[5] Grimes, S. Unstructured data and the 80 percent rule. Carabridge Bridgepoints.

[6] Hung, J. L., & Zhang, K. Examining mobile learning trends 2003–2008: A categorical meta-trend analysis using text mining techniques. Journal of Computing in Higher education, 24(1), 1-17.

[7] Feldman, R., & Dagan, I. (Knowledge Discovery in Textual Databases (KDT). In KDD (Vol. 95, pp. 112-117).

[8] Rajman, M., & Besançon, Text mining: natural language techniques and text mining applications. In Data mining and reverse engineering (pp. 50-64). Springer US.

[9] Gök, A., Waterworth, A., & Shapira, P. (2015). Use of web mining in studying innovation. Scientometrics, 102(1), 653-671.

[10] Fan, W., Wallace, L., Rich, S., & Zhang, Z. Tapping into the power of text mining.

[11] SØRENSEN, H. T., Sabroe, S., & OLSEN, J. A framework for evaluation of secondary data sources for epidemiological research. International journal of epidemiology, 25(2), 435-442.

[12] Zhang, J. Q., Craciun, G., & Shin, D. When does electronic word-of-mouth matter? A study of consumer product reviews. Journal of Business Research, 63(12), 1336-1341.

[13] Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis; Foundations and Trends in Information Retrieval; Vol. 2, Nos. 1–2, 1–135, 2008.

[14] Stieglitz, S., & Dang-Xuan, L. (2013). Social media and political communication: a social media analytics

[15] Celli, F., Pianesi, F., Stillwell, D., & Kosinski, M. (2013, June). Workshop on computational personality recognition (shared task). In Proceedings of the Workshop on Computational Personality Recognition.

[16] Rahman, M. Mining social data to extract intellectual knowledge. arXiv preprint arXiv:1209.5345.

[17] Al-Daihani, S. M., & Abrahams, A. (2016). A Text Mining Analysis of Academic Libraries' Tweets. The Journal of Academic Librarianship, 42(2), 135-143.,2016

[18] Mosley Jr, R. C. Social media analytics: Data mining applied to insurance Twitter posts. In Casualty Actuarial Society E-Forum, Winter 2017 Volume 2 (p. 1).

[19] Kermanidis,K.L., & Maragoudakis, M. Political sentiment analysis of tweets before and after the Greek elections of May 2012. International Journal of Social Network Mining, 1(3-4), 298-317.

[20] Ahmad N, Siddique J. Personality Assessment using Twitter Tweets. Procedia Computer Science. 2017 Jan 1;112:1964-73.

[21] Rizwana Irfan, A Survey on text mining in social networks, The Knowledge Engineering Review, Cambridge University Press, 2018

[22] A Razia Sulthana et al, Sentiment analysis in twitter data using data analytic techniques for predictive modelling, National Conference on Mathematical Techniques and its Applications (NCMTA 18), 2018