

A Survey of Various Opinion Mining in Social Media

Jata Shankar Jha^{#1}, Asst. Prof. Vimal Shukla^{*2}

[#]Department of Computer Science & Engineering
KNPCST, Bhopal, India

¹jjatashankar@hotmail.com

²vimalshukla.cse@gmail.com

Abstract— Opinion Mining is a technique of providing and giving opinion on a particular topic so that a final conclusion can be extracted from it. Here in this paper a complete survey of all the technique that is used for the opinion mining. A complete survey of all the technique implemented for the social media opinion is discussed and analyzed here so that various advantages and limitations can be analyzed and hence on the basis of which a new and efficient technique can be implemented in future.

Index Terms—World Wide Web, Opinion Mining, WCM, WUM, Social Network.

I. INTRODUCTION

World Wide Web (WWW) is a digital world which is rapidly and continuously growing on all aspects. It is a massive, huge, diverse and dynamic unstructured data repository. This repository on WWW is used as information repository for knowledge reference. Web faces challenges in the form of large, semi structured web pages and information on web is liable to be diverse in meaning, quality of the information extracted and the conclusion of the knowledge obtained from extracted information. Thus for the proper understanding and analysis the data structure of the Web is important for efficient Information Retrieval.

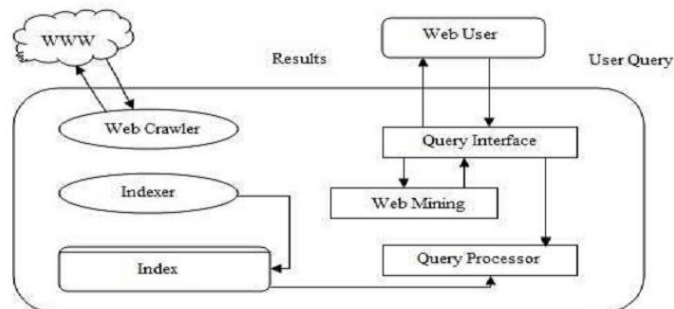


Figure :1 Architecture of Search Engine.

The architecture of a search engine consists of Crawler, Indexer and Ranking mechanism. The crawler traverses the web and downloads the web pages also known as spider or robot. The downloaded pages are then given to a indexing module which parses the web pages and constructs index based on the keywords in the web pages. An index is maintained using the keywords contained in the web pages. User types a query using different keywords on the interface of a search engine provided after which the query processor component matches the query keywords with the index and giving back the URLs of the pages related to the user. The pages are first passed through a ranking mechanism done by the search engines that show the most relevant or high ranked pages at the top and less relevant/ low ranked pages at the bottom [1]. After this process the pages result are shown to the user.

Web mining is defined as mechanism that classifies the web pages and internet users in accordance with the contents of the web page and the behavior of the user adopted in the past on the internet. Web Mining is an application of data mining technique. It is used to find and retrieve information from the WWW automatically [2].

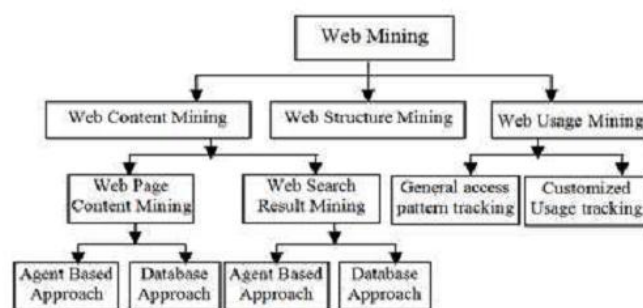


Figure :2 Web Mining Categories

The information that can be retrieved by web mining technique is divided in classes which are:

- Web content from data residing on web pages and inside the documents contained by the web pages.
- Web activity obtained from server logs and browser activity.
- Web graphs obtained from links between the web pages and other various types of data.

Web Content Mining (WCM) concerns with retrieval of information from WWW in a structured form and thereby indexes the information for quick retrieval. It is the process to extract useful information from web documents contents. The documents consists of images, audio, video, text files, structured records etc. WCM technique is used on web documents and the results page obtained from a search engine. WCM basically works upon or say consists of two approaches agent based approach and database based approach. Agent based approach finds relevant information using the characteristics of a domain in meticulous manner while the database approach retrieves the semi structure data from the web.

Web Usage Mining (WUM) records the user's profile and his behavior in a log file inside the web. Web usage mining process derives useful information from the data which is accessed by the user while surfing on the Web. It basically extracts data stored in server access logs, agent logs, client-side cookies, user profiles, meta data etc.

The basic fundamental behind all these process is to retrieve information according to the user's behavior and interests and show them the results. While resolving a query users are displayed with multiple results in the form of URL's of various web pages. For such process various techniques are adopted in the form of multiple algorithms

Defining a social network it is a social structure which consists of individuals or organizations) which are in some manner related to each other. The social network viewpoint provides a set of methods which analyzes the structure of social entities and the theories that explains the patterns observed in these structures. The study of these structures uses social network analysis for identification of local and global patterns, locating influential entities and examining network dynamics. Social network approaches to understand social interaction which should be first visualized and investigated through the properties of relations between units and not unit properties itself. In a social network different types of relations that may be singular or combination form the network configurations and network analytics. Whereas a social networking service provides platform for building

social networks or social relations among users who share interests, activities, backgrounds etc. With the help of social network service each user maintains its profile containing his or her social links and other additional services.

Therefore with the help of Social networks users can create a public profile and can maintain a list of users to share connections and view and cross the connections inside the system. Social network services are web based facilitating the user to interact over the Internet, in the form of e-mail and instant messaging. Social network allow multiple information and communication tools like mobile connectivity, photo, video, sharing, blogging etc.

Social networks provide community services which are group centered and sometimes social networks are also considered as individual centered service. Social networks can be divided into communities consisting of users that may have same like, features, dislikes etc. Considering the main types of social networking services are those which are categorized in groups or communities of users like schoolmates, politicians, celebrities etc. and a recommendation system linked to trust of its users. Some of the widely used Social networks are Facebook, Google+, YouTube, LinkedIn, Instagram, Pinterest, Tumblr and Twitter.

Each user has its own social network that may be online or offline consisting of friends, families and people they are acquainted with. Basic fundamental aim of online social networking is to make users social networks visible to others who are not in his/her immediate network. In a social network people are held together by kinship, friendship, classmates, colleagues, business partners, etc. which is a pre-established interpersonal relationships. The connections between the users are built one at a time. The primary reason behind people joining a social networking site is to maintain old relationships with others and form new ones for expansion of their network.

Social networks are extremely unique in their own way as users collectively identify others if they are fake and also users generally do not compartmentalize their life i.e. they don't have only one social network. Communities in the social network are held together by common interest. The users may have common hobby for which the community members are passionate, a common goal, project, similar lifestyle, geographical location, profession etc. Thereby in social networks, there are two types of users exhibiting different influence and different behavior [3].

As the internet consists of information of various type and kind. To exploit and gain knowledge from this data there are

users who reside on internet and continuously use it. The users share, disseminate and communicate multiple type of information among them. The information being in texts, audio, video, images etc. [2]. The users belong to various communities that consist of similar type of users in behavior which influence each other to share various type of data. The medium for this type of data exchange can be in the form of mails, messages and social networks.

Social Networks can be visualized as an internet service which helps the user to build a social networks over the internet and relations with other users for sharing of interests, backgrounds and even establish real life connections and participate in multiple activities with the users that can be characterized by communities.

Social Network services are web based group centered services in which users share almost every type of data.

II. LITERATURE SURVEY

Wayne Xin Zhao [4] et.al. suggested that while being Twitter as a new form of social media and containing useful information, content analysis on Twitter has not yet been well premeditated. They compared the content of Twitter with traditional news

medium i.e. New York Times by unsupervised topic modeling. They used a Twitter-LDA model for discovering topics from a representative sample of the entire Twitter. With the help of text mining technique they compared Twitter topics and the topics from New York Times while considering the topic categories and types. They presented a study based on the relation between the magnitude of prejudiced tweets and retweets and also over topic categories and types. With the help of Twitter-LDA model analyzing short tweets showed its efficiency compared with existing models. They proposed the concept of topic categories and topic types facilitating analysis

of the contemporary differences between Twitter and they found that Twitter is a good source of entity oriented topics having low coverage in traditional news media. While Twitter users having low interests in world news, dynamically spread news of important world events [4].

Hila Becker [5] et.al. suggested that social media sites like Twitter and others contain huge amounts of user contributed messages comprising of real world events. Such events messages sometimes contain interesting and useful information like event time, location, participants, opinions that might interest people in learning about an event. They explained that with effective selection

technique of quality event helps improving applications like event browsing, search etc. they proposed exploring of representative messages

among Twitter messages which corresponds to the same event, which can thereby give high quality, relevant messages capable of providing useful event information. They presented their approach and results on the basis of large-scale dataset of Twitter messages capable of automatic selection of event messages that are relevant and useful. They proposed explanation that sometimes a single event attracts thousands of social media content items, so the proposed methodology should be capable of ranking and filtering event content for applications that aim to communicate that content effectively. They proposed three centrality-based approaches, Centroid as the favored way to select tweets given a cluster of messages relating to an event. Whereas LexRank and Degree tend to select messages that are strongly similar to one another, but sometimes diverge from the main topic of the cluster [5].

Liangjie Hong [6] et.al. reviewed that social network services are now a practicable source of information for the users. Explaining the concept of Twitter they presented that information supposed to be important by the community generally propagates through or via retweets. The characteristics of popular messages can be studied and analyzed for breaking news detection, personalized message recommendation, viral marketing etc. they investigated the problem of prediction of popularity of messages by the number of future retweets and the factors that influence information propagation in Twitter. They put together the specified task into a classification problem and studied features based on the content of the messages, temporal information, metadata of messages and users and the structural properties of the users' social graph on a large scale dataset. With their proposed methodology the prediction of messages is made possible attracting thousands of retweets with good performance [6].

Roja Bandari [7] et.al. presented that News articles are particularly time sensitive comprising of intense competition among news items for wide propagation. This makes the task of foretelling the popularity of news items on the social web as interesting and challenging. They explained that in previous research prediction of online popularity is based on early popularity. They presented a multi-dimensional feature space consequent from properties of an article and estimate the efficiency of these features that can serve as predictors of online popularity. They observed regression and classification algorithms and demonstrated that even with

randomness

in human behavior possibility to predict ranges of popularity on social media can have an overall 84% accuracy. They also studied differences between traditionally prominent sources and those that are immeasurably popular on the social web. They actually predicted the popularity of news items on Twitter with the help of features extracted from the content of news articles. They explained four features that were analyzed - firstly the source of the article, the category, subjectivity in the language and the named entities mentioned. Though in their results these features are not sufficient for predicting the accurate number of tweets that an article will earn but can still be effective for providing a particular range of popularity of the article on Twitter. They have also analyzed that on the basis of number of re-tweets the top news sources on twitter is not conventionally popular news agency. But the most important predictor of popularity is the source of the article. This explains that the readers influenced by the news source that publish the article [7].

Phil Long [8] et.al. recognized that the future of education may emphasize new technologies like ubiquitous computing devices, flexible classroom designs, and also innovative visual displays. But big data and analytics is the major factor shaping the future of higher education. Basing decisions on data and evidence is a bit obvious and also research indicates that data driven decision making helps enhancing organizational output and productivity. They explained that Higher education as a field is capable of gathering an astonishing array of data about its "customers," and

has conventionally been inefficient in its use of data often operating with substantial delays in analyzing readily evident data and feedback. They proposed that analytics in education should be transformative and the use of analytics can generate privacy, profiling, information sharing, and data stewardship in higher education. The basic proposal by them included that in education the value of analytics and big data can be found in guiding reform activities in higher education, and can assist educators in improving teaching and learning [8].

Mattias Rost [9] et.al explained that online services are capable of providing opportunities for understanding human behavior with the help of large aggregate data sets that their operation collects. But the data sets collected by them are not able to model or mirror the world events unproblematic ally. They have considered the data from Foursquare for analyzing social media as a communicative and not as representational system. They have

highlighted four features of Four square's use: firstly the relationship between attendance and check-ins, second event check-ins, third commercial incentives to check-in and humorous check-ins. They proposed that social network data should be viewed and analyzed as communicative data. This means data that is produced as a side effect of communication between users, rather than as a representation of some underling activity. Looking at the data overall they observed that how the number of check-ins does not have clear correspondence with visitor numbers and the role of promotions as a motivation for check-ins. Also the evidence of humor on the other hand showed itself to be fairly complicated and hardly seemed to affect any large scale analysis [9].

X. Chen [10] et.al. Explained that students while being on social media share their experiences and opinions feelings about the learning process. Such type of data is useful for analyzing the student behavior which may directly lead to enhance student learning. With the help of qualitative analysis and large scale data mining techniques student's problems can be encountered and analyzed from various tweets and posts. The algorithm can then be trained to detect student's problems. Through this methodology student's experiences can be analyzed. Their methodology basically provides learning analytics, educational data mining, learning technologies. In their research social media data can be analyzed for educational purpose overcoming manual qualitative analysis and large scale computational analysis. With this analysis college students experiences can be encountered which thereby can help increasing the quality of education [10].

III. CONCLUSION

Here in this paper a complete survey of all the technique that are used for the opinion mining especially in social media are discussed and analyzed their limitations and advantages of the technique. Since Opinion mining enables various users to take decision on the basis of various review or opinions of the users and hence get a final conclusion, here also various opinions given on social media web is collected and a final decision is taken based on these opinion.

REFERENCES

- [1] Laxmi Choudhary and Bhawani Shankar Burdak "Role of Ranking Algorithms for Information Retrieval", 2010
- [2] Nikhil Sanyog Choudhary, Himanshu Yadav and Anurag Jain "Message Filtering Techniques in Social Networks over Web Environment- A Survey", IJETAE, 2014.
- [3] Liaoruo Wang, Tiancheng Lou, Jie Tang and John E. Hopcroft "Detecting Community Kernels in Large Social Networks", 2011



-
- [4] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan and Xiaoming Li, "Comparing Twitter and Traditional Media using Topic Models", 2011
 - [5] Hila Becker, Mor Naaman and Luis Gravano "Selecting Quality Twitter Content for Events", Association for the Advancement of Artificial Intelligence, 2011.
 - [6] Liangjie Hong, Ovidiu Dan and Brian D. Davison "Predicting Popular Messages in Twitter", ACM, 2011.
 - [7] Roja Bandari, Sitaram Asur and Bernardo Huberman, "The Pulse of News in Social Media: Forecasting Popularity", 2011.
 - [8] Phil Long and George Siemens "Penetrating the Fog: Analytics in Learning and Education", Educause Review, 2011.
 - [9] Mattias Rost, Louise Barkhuus, Henriette Cramer and Barry Brown "Representation and Communication: Challenges in Interpreting Large Social Media Datasets", ACM, 2013.
 - [10] Xin Chen, Mihaela Vorvoreanu and Krishna Madhavan, "Mining Social Media Data for Understanding Students' Learning Experiences", IEEE Transactions, 2014.