Text Classification and Analysis with Social Media Platform

Ms. Mrunali D. Chaure, Prof. Jayant P. Mehare
Student of Computer Science and Engineering, Assistant Professor
G. H. Raisoni University, Amravati

Abstract : Text Classification, Text Mining and Sentiment Analysis have received huge attention, especially because of the availability of large volume of textual data on social media, e-commerce websites, blogs and other similar sources. This data is usually unstructured that becomes complex and expensive to extract knowledge from them to make some intelligent decisions. There is a growing need for developing different methodologies and models for efficiently processing the texts and extracting useful information from it. Text analysis, also known as opinion mining, is the process of quantifying the emotional value in a series of words or text, to gain an understanding of the attitudes, opinions and emotions expressed about particular object. One way to extract information is text classification and its analysis that helps in decision making in various field of our lives. One of such a growing field is Social Media. Now a days people are so much active to obtain and share different types of information updates through social media on a 24/7 basis. Many research areas have tried to extract valuable insights from these large volumes of user generated content. This paper performs the study and analysis of Social Media post by using different technique of text analysis and opinion mining along with sentiment analysis. This study will help to identify different techniques, and chose the most suitable one that helps to identify the most efficient to analyses the Social Media post.

IndexTerms: **Decision Making, Natural Language Processing (NLP). Social Networking, Social Media, Text classification, Text Analysis, Unstructured Data.**

I. INTRODUCTION

Text classification is the process of classifying textual documents into predefined categories based on their content. Classifying text is the automated assignment of natural language texts to predefined categories. Text classification is the primary requirement of text retrieval systems, which retrieve texts in response to a user query, to extract some knowledge and text understanding systems, which transform text in some way such as producing summaries, answering questions, making decisions or extracting data.

Text mining has become one of the trendy fields in technology area that has been incorporated in several research fields such as computational linguistics, Information Retrieval (IR) and data mining. Natural Language Processing (NLP) techniques were used to extract knowledge from the textual data that is written by human beings. Text mining reads an unstructured form of data to provide meaningful information patterns in a shortest time period [1]. Now a days, Social networking sites are a great source data generation as most of the people in today's world use these sites in their daily lives to keep connected to each other.

Social networking websites create new ways for engaging people belonging to different communities [2]. Social networks allow users to communicate with people exhibiting different moral and social values. The websites provide a very powerful medium for communication among individuals that leads to mutual learning and sharing of valuable knowledge. On social media it becomes a common practice to not write a sentence with correct grammar and spelling. This practice may lead to different kinds of ambiguities like lexical, syntactic and semantic and due to this type of unclear data, it is hard to find out the actual data order. Therefore, extracting logical patterns with accurate information from such unstructured form of data is a critical task for performing analysis [3].

Social network analysis applications have experienced tremendous advances from past few years due in part to increasing trends towards users interacting with each other on the internet. Social networks are organized as graphs, and the data on social networks takes on the form of massive streams, which can be mined for various purposes. Social Network Text Analysis from the post covers an important era in the social network analytics field. This edited volume, contributed by prominent researchers in this field that helps in presenting a wide selection of topics on social network data mining such as Structural Properties of Social Networks, Algorithms for Structural Discovery and Content Analysis in Social Networks [4].

With the rise of Social Media, people obtain and share different types of information updates on a 24/7 basis. Social media includes social networking sites and blogs where people can easily connect with each other. Social media has been mainly defined as "the many relatively inexpensive and widely accessible electronic tools that facilitate anyone and anytime access information, collaborate on a common effort, or build relationship". Many research areas have tried to gain valuable insights from these large volumes of freely available user generated content. The research areas for e-Commerce, intelligent transportation systems, smart cities, Cyber Crime, etc. are no exception. However, extracting meaningful and actionable knowledge from user generated content is a complex task. As each social media service has its own data collection formats and constraints. The volume of messages posts produced becomes overwhelming for automatic processing and mining [5]. Along with this, social media texts are usually short, informal, with a lot of abbreviations, jargon, slang leads to unstructuredness.

As per the above mentions application of social media platform now a day's social media has been the important part of one's life and plays a vital role in transforming people's life style. Since the emergence of these Social networking sites like Twitter, Facebook and Instagram as key tools for news, journalists and their organizations have performed a high-wire act. These sites have become a day to day routine for the people. However, one does not need to look very far to experience the darker side of social media use. News reports of cyber

bullying, violence, criminal activity, and suicide fueled by social media is shocking and troubling. Social networks are an inherent part of today's Internet and used by more than a billion people worldwide. They allow people to share ideas and interact with other people, from old friends to strangers. This interaction reveals a lot of information, often including personal information visible to anyone who wants to view it. Children are also growing around and surrounded by mobile device and uses interactive social networking sites. So it becomes necessary to avoid this drawback and improve the detection of violence posting on the social media [6].

This paper mainly studies the text analysis techniques that can be helpful for the detection and prevention of positive, negative and violence type of post on social media sites. In further sections of this paper, Section II, shows different levels of opinion mining that helps in suggesting positive, negative and Neutral kinds of opinions. In section III, different techniques used by different authors for this text classification and analysis related to social media post. Section IV gives the brief introduction about text analytics and briefly explains why text analysis is essential in Social Media platform along with its proposed architecture diagram. Their studies are studied and briefly describe in Literature review section. Finally, in section V the paper is concluded.

II. LEVELS OF OPINION MINING

There are mainly two approaches that can help to perform the task of Opinion mining from the textual data easily available from different sources that is Feature based opinion mining and Sentiment classification. In case of sentiment classification, the sentiment or opinion of the user is classified as positive or negative or neutral and then it will summarized to generate an opinion. But in case of feature based opinion mining, the features are selected from reviews and the sentiment is extracted after feature selection. There are mainly three levels of opinion mining; Document level, Sentence level and Entity or Aspect Level these are briefly as follows [7]:

1. Document Level:

As the name suggest the task at document-level analysis is to classify whether a whole document is expressing a positive or negative opinion. This level of opinion mining assumes that each document gives opinion about a single product or single topic.

2. Sentence Level:

In this type of analysis, the task is to identify each sentence and judge that weather it gives a positive, negative or neutral sentiment. Here Neutral is usually means this sentence does not give any opinion. This level of analysis is related to subjective classification, which distinguishes objective sentences that expresses factual information from subjective sentences.

3. Entity and Aspect Level:

Aspect level analysis performs fine-grain analysis. Aspect level is earlier called feature level. Instead of looking at language constructs that is documents, sentences, paragraphs, etc., aspect level directly look for opinion. It is based on idea that an opinion consists of a sentiment i.e. positive or negative and a target of opinion.

III. LITERATURE REVIEW

There are lots of studies done by different authors that are paying attention on text analysis for most useful Social Media platforms to detect the uses cases as well as prevent the wrong impact happening with them. In this section, techniques and methods used by different researchers for Text analysis, opinion mining and cyber hate detection analysis are describe briefly.

Social networking websites, such as Facebook [8] are rich in texts that enable user to create various text contents in the form of comments, wall posts, social media, and blogs. Due to ubiquitous use of social networks in recent years, an enormous amount of data is available via the Web. Application of text mining techniques on social networking websites can reveal significant results related to person-to-person interaction behaviors. Moreover, text mining techniques in conjunction with social networks can be used for finding general opinion about any specific subject, human thinking patterns, and group identification in large-scale systems.

Author D. Correa, A. Sureka in year 2011 design a system for the radicalization of different types on the web is given. Link Based for making decision on detection of radical information this feature Identify the structure between documents. Hyperlink between web sites is included in this feature. It also includes relationship between users on web site. Content Based this feature provide content and structure of documents lexical, syntactic, graph based , structural features , link based feature is include in content based. This feature mainly used in link based bootstrapping algorithm, content based feature used in text classification techniques.

Author [9] works on an approach to identify a ranked list of radically influential users in Web forum. The radical measure variety of collocation-based association measures, and designed an algorithm based on Page Rank to rank the radically influential users. Among the proposed association measures, the contingency coefficient measure is found as the most promising measure, when embedded in the customized PageRank algorithm along with the radicalness measure. The experimental results on a standard data set are promising that outperforms the existing User Rank algorithm. It is also found that the collocation-based association measures deal with such ranking problem more effectively than textual and temporal similarity based measures.

Author Shabnoor Siddiqui, Tajinder Singh in year 2016 works on Social Media its Impact with Positive and Negative Aspects. Social media has become the routine for each and every person; peoples are seen addicted with these technology every day. With different fields its impact is different on people. Social media has increased the quality and rate of collaboration for students. Business uses social media to enhance an organization's performance in various ways such as to accomplish business objectives, increasing annual sales of the organization. Young people are seen in contact with these media daily .Social media has various merits but it also has some demerits which affect people negatively. False information can lead the education system to failure, in an organization wrong advertisement will affect the productivity.

Author Mane Priyanka , Rathid Sonali , Sanap Deepali & Shirude Bhavana in year 2016 works on influential users dominate the mind of naive users using their radical thoughts. Influential users compel the naive users to do wrong things. This system identify radical influential user from the web forum and rank this user according to their comment on forum. According to the user rank this influential users are

removed from the forum. In this system start to implement the first module of our system which is forum crawling and preprocessing. This system is useful for removing the radical influential users from the web forum.

The authors, presented survey on sentiment analysis and opinion mining. In this survey they have explained opinion oriented information right of entry, challenges, opinion categorization and their summarization. Many researchers used machine learning methods for emotion examination. The techniques explained by the authors consists of guidance of classifier on the datasets and that uses the skilled model for classification of new documents. There are some another techniques that uses optional methods such as dictionary of word lexicons.

The authors [10] make uses of Social networks, collect and analyze raw data that people posting as real time messages about their opinions on a variety of topics in daily life. Data available in social media is obviously only one type of information that can be of interest when trying to detect a possible terrorist or radical group, there are several cases for example in which the social media has been used by radical thinkers to act as influencers and encourage fanatics with the same radical views to take violent action. With this, Author propose a framework for opinion mining and extremist content detection in online social media data which is the public text post on Facebook, he most popular social networking site. With this framework, machines can learn how to automatically extract the set of messages from Facebook public pages, using API graph calls, filter out non-opinion messages. Determine their sentiment regarding the issue of interest directions (i.e. positive, negative) and detect violent or extremist content. This model is helpful for law enforcement and cybercrime analysts with analyzing and monitoring social media, in the search of digital trace of violence or radicalism that can be exploited in further digital forensic investigation.

The authors [11] both proposed to use supervised sequential labeling methods for topic and opinion extraction. Results showed that the supervised learning methods can achieve state-of-the-art performance on lexicon extraction. However, the limitations occurs as these methods need to manually annotate a lot of training data in each domain. Recently, Qiu *et al.* proposed a rule-based semi-supervised learning methods for lexicon extraction. However, their method requires to manually define some *general* syntactic rules among sentiment and topic words. In addition, it still requires some annotated words in the target domain to be identified.

This project paper [12] mainly focuses on detection of terrorism content in social media. The content can be in form of text to spread the terrorist sentiments or threats/messages. The content used here can also be of the form of terror group images or their pictures of arms and attacks. Since if such content is published on social media it can spread vast unrest and agitation in public of any state or nation. So this paper, tries to automatically detect such content in real time and prevent from being uploaded on internet or disabled/removed from social media site if anywhere by chance.

The authors [13], work for detection and prevention of the existence of malicious actors such as bots on Twitter, vandals on Wikipedia, fake accounts on Facebook, trolls on Twitter and Slashdot, and spammers who seem to be omnipresent. They presents methods to identify malicious actors in 4 settings: Twitter, Facebook, Slashdot, and Wikipedia as: (i) network based techniques where the structure of the social network is used, (ii) text based methods where the linguistic content of posts is examined, (iii) behavior-based methods which study actions of users, and (iv) real-time processes which enable defenders of social media to keep a step ahead of malicious actors. The paper tries to identify commonly used features for classifying actors into malicious vs. benign and give a brief explanation of different algorithms both specific to social platforms and general algorithms that are platform independent.

This paper [14], works on most dangerous methods to hurt people's feelings by identifying the social networks, which becomes an essential way of communication now a days. This is termed as cyber-aggression, or in some cases called cyber-bullying. In this paper auther researches to classify situations of cyber-aggression on social networks, specifically for Spanish-language users of Mexico. They applied Random Forest, Variable Importance Measures (VIMs), and OneR to support the classification of offensive comments Particularly in three cases of cyber-aggression as: racism, violence based on sexual orientation, and violence against women. The study shows that to get accurate classification of cyber-aggression, comments can help to take measures to diminish this phenomenon. It is observed that the exploration of other machine-learning techniques and the continuous update of the offensive data set may not be ruled out for future. It is necessary to perform other kind of cyber-aggression case, e.g. those suffered by children. Authors also suggest that, building an automatic labeling system for offensive Comments made by social networks users, helps in minimizing human error. Authors also seek to identify the victims of cyber-aggression to provide them with psychological attention according to the case of harassment that they suffer.

IV. TEXT ANALYTICS

Text analytics is the process of text mining that involves the process of converting unstructured text data into meaningful data for analysis. This generally helps to measure product review before purchase, customer opinions, feedbacks that helps in search facility out of multiple options available in the market. Social Media comments for any particular object or any live post. There are many machine learning techniques used as linguistic, statistical and machine learning technique that helps in analyzing text. The process of text analytics involves retrieval of information from unstructured data and the process of structuring the input text to derive patterns and trends that further evaluates and interprets the output data. Text analytics determines the keywords, topics, category, semantics, tags from the millions of text data available in an organization in different files and formats. There are different software's available for performing this text analytics and further action for analysis by other tools such as business intelligence, big data analytics or predictive analytics tools [16].

There are different applications of Text analytics as for Sentiment analysis, and entity modeling that supports facts based on decision making. Search and access for unstructured data, big data analytics, Social Media Monitoring, Cyber hate detection, business intelligence, record management and access, scientific discovery, different security applications, etc.

Why Text Mining and Social Media Analytics in Essential?

In the beginning of social media or microblogging, public relation organizations would screen clients' posts on business websites trying to distinguish and oversee displeased clients. With their expansion this isn't sufficient and people seems it very simple along with it is rich with opportunity. Number of social media users is 2.46bn worldwide in 2017 [17] with Facebook, YouTube, and twitter are most trafficked sites on the Internet. In any case, even these insights fail to give a full record of the impact that social networking or microblogging sites are having. Users spend over 135 minutes per day online [18] through social media or microblogging sites. Facebook alone has an overall market infiltration rate 26.3%; in North America it is 72.4% [19]. These rates are developing rapidly, with Facebook has 2.07bn users which was only 1.59bn in the end of 2015 [20]. YouTube's mining of its videos demonstrates 100 million individuals like, dislike, comment or share those videos every week [21]. Within two years this figure doubled. Facebook now incorporates social activities in its online promotions, for example by enabling user to investigate whether their partners have favored or chosen on items being advertised while they were watching YouTube videos. Essentially, hashtags on twitter have given clients another speedy approach that helps to express their likes, dislikes or comments; and these offers opportunities for companies to study about their sentiments. In today's global marketplace, the social media data become one of the rich sources for companies to understand the market value and customer diversity [22]. Figure 1. Below shows the architecture diagram for proposed text analysis model with social media platform.

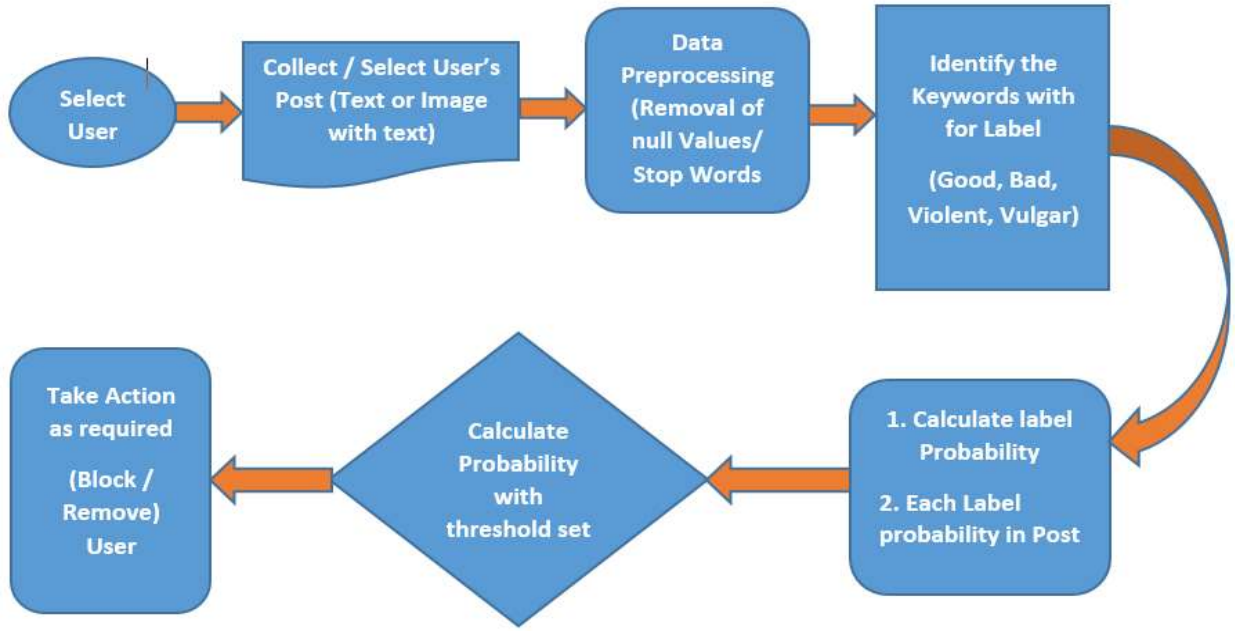

Figure 1. Architecture Diagram for Text Analysis

V. CONCLUSION

As we have seen the use of Social media platforms by all age people is growing on increasing, people are seen addicted with these technology every day. So it becomes necessary to develop some techniques and methods for analyzing customer's post as they are present in thousands. There are different terms and methods like opinion mining, sentiment analysis, word alignment model, etc. associated with this concept. There are different levels of opinion mining and according to our requirement opinions are generated. With different fields its impact is different on people. Social media has increased the quality and rate of collaboration for students. Business uses social media to enhance an organization's performance in various ways such as to accomplish business objectives, increasing annual sales of the organization. This technique helps to stop violence data on the social media. Here literature study is performed on the different techniques used by different researchers in the field and then proposed system that Determining the category of the post and that helps to find out the average percent value of malicious post on the social media.

REFERENCES

[1] Said A. Salloum, Mostafa Al-Emran, Azza Abdel Monem, Khaled Shaalan, "A Survey of Text Mining in Social Media: Facebook and Twitter Perspectives", *Advances in Science, Technology and Engineering Systems Journal*, Vol. 2, No. 1, 127-133 (2017).
[2] RIZWANAIFA, CRISTINE K. KING, "A Survey on Text Mining in Social Networks", *The Knowledge Engineering Review*, Vol. 00:0, 1–14 ©c 2004, Cambridge University Press
DOI: 10.1017/S000000000000000.
[3] Sorensen, L. 2009. "User managed trust in social networking comparing facebook, myspace and linkdin", *In Proceedings of 1st International Conference on Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic System Technology*, (Wireless VITAE 09), Denmark, 427-431.
[4] Puja Munjal, Aditi Gupta, Mahima Abrol, Hema Banati and Sandeep Kumar, "Social Media Based Opinion Mining Using Lexical Sentiment Analysis", *International Conference on Paradigm Shift in World Economies: Opportunities and Challenges* - ISBN :978-1-63535-729-5 (2017).
[5] João Filipe Figueiredo Pereira, "Social Media Text Processing and Semantic Analysis for Smart Cities", FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO, arXiv:1709.03406v1 [cs.SI] 11 Sep 2017.
[6] Ahmed Imran KABIR, Ridoan KARIM, Shah NEWAZ, Muhammad Istiaque HOSSAIN, "The Power of Social Media Analytics: Text Analytics Based on Sentiment Analysis and Word Clouds on R", Research Gate, Informatica Economică vol. 22, no. 1/2018. DOI: 10.12948/issn14531305/22.1.2018.03.
[7] Vibhuti Patel, Mital Panchal, "A survey on Opinion Mining Methods from Online Reviews", International Journal of Scientific Research in Science, Engineering and technology, In December, 2015.
[8] Aggarwal, C. 2011. Text mining in social networks. In Social Network Data Analytics. 2nd edn. Springer, 353-374. Baumer.
[9] T. Anwar and M. Abulaish, "Ranking Radically Influential Web Forum Users," in *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 6, pp. 1289-1298, June 2015. doi: 10.1109/TIFS.2015.2407313.
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7050292&isnumber=7084215
[10] Bo Pang, Lillian Lee, "Opinion Mining and Sentiment Analysis", *Foundations* and Trends in Information Retrieval Vol. 2, Nos. 1–2 (2008).
[11] E. Mouhssine and C. Khalid, "Social Big Data Mining Framework for Extremist Content Detection in Social Networks," *2018 International Symposium on Advanced Electrical and Communication Technologies (ISAECT)*, Rabat, Morocco, 2018, pp. 1-5. doi: 10.1109/ISAECT.2018.8618726.
[12] Fangtao Li, Sinno Jialin Pan, Ou Jin, Qiang Yang and Xiaoyan Zhu, "Cross-Domain Co-Extraction of Sentiment and Topic Lexicons", *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 410–419, Jeju, Republic of Korea, 8-14 July 2012.
[13] Terrorism detection from social media, available [online]:*https://www.leadingindia.ai/downloads/projects/SMA/sma_6.pdf*
[14] Srijan Kumar, Francesca Spezzano and V.S. Subrahmanian, "Identifying Malicious Actors on Social Media", available [online]: https://cs.stanford.edu/~srijan/badactorstutorial/
[15] Guadalupe Obdulia Gutiérrez-Esparza, Maite Vallejo-Allende and José Hernández-Torruco, "Classification of Cyber-Aggression Cases Applying Machine Learning", *Applied Science*, May 2019, 9(9), 1828; https://doi.org/10.3390/app9091828.
[16] From Research page: What is Text Analytics? [Online] available: https://www.predictiveanalyticstoday.com/text-analytics/
[17] K. Gordon. Number of social mediausers worldwide from 2010 to 2021 (in billions) [Online]. Available: https://www.statista.com/statistics/278414/number-of-worldwide-socialnetwork-users/
[18] Unknown. Daily time spent on social networking by internet users worldwide from 2012 to 2017 (in minutes) [Online]. Available: https://www.statista.com/statistics/433871/daily-social-media-usageworldwide/
[19] Unknown. Percentage of global population using Facebook as of June 2017, by region [Online]. Available: https://www.statista.com/statistics/241552/share-of-global-populationusing-facebook-by-region/
[20] Unknown. Number of monthly active Facebook users worldwide as of 3rd quarter 2017 (in millions) [Online]. Available: https://www.statista.com/statistics/264810/number-of-monthly-activefacebook-users-worldwide/
[21] K. Vance, W. Howe, and R. P. Dellavalle, "Social internet sites as a source of public health information," Dermatologic clinics, vol. 27, pp. 133-136, 2009.
[22] J. S. Brown and P. Duguid, The Social Life of Information: Updated, with a New Preface: Harvard Business Review Press, 2017.