International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 4, Issue 4, ISSN No. 2455-2143, Pages 206-210
Published Online August 2019 in IJEAST (http://www.ijeast.com)

REVIEW PAPER ON SENTIMENT ANALYSIS OF AMBEDKAR.ORG USING TEXT MINING

Amrapali P. Tribhuvan
Department of Computer Science and Information Technology
Dr. Babasaheb Ambedkar Marathwada University, Aurangabad

Abstract—**India is a country of various religions, castes and cultures. It is based on secularism and democracy. Judiciary, legislature and executive are the major pillars of democracy. Media plays the role of watchdog and it is known as 'Fourth estate'. The responsibility of media is to empower and enrich the democracy through promoting pluralism and diversity. The media should be the voice of people and they must ensure that every section of society is represented in the media properly. Despite such concepts and theories on the role of media, the media in India failed to represent properly the Dalit community of in our country and to provide them opportunities in the field of media. A Dalit- Bahujan Media, Dr. Babasaheb Ambedkar and his People, ambedkar.org, Indian Dalit web portal is taken for the study. Text mining and sentiment analysis have received huge attention recently, especially because of the availability of vast data in form of text available on social media, e-commerce websites, blogs and other similar sources. The main objective of this paper is to understand the presence of Dalits in this new media and analyze the feedback about of Dalit website. The method adopted is sentiment analysis, which aims to extract emotions and opinions from text.**

Keywords—**Sentiment analysis, Opinion Mining, text mining, Part-of-Speech(POS), Opinion Word Extraction, Polarity Identification, Summary Generation, Dalit, web portal**

I. INTRODUCTION

The general proposition that the social organization of the Indo-Aryans was based on the theory of Chaturvarnya and that Chaturvarnya means division of society into four classes—Brahmins (priests), Kshatriyas (soldiers),Vaishyas (traders) and Shudras (menials) does not convey any idea of the real nature of the problem of the Shudras nor of its magnitude. Chaturvarnya would have been a very innocent principle if it meant no more than mere division of society into four classes. Unfortunately, more than this is involved in the theory of Chaturvarnya. Besides dividing society into four orders, the theory goes further and makes the principle of graded inequality. the basis for determining the terms of associated life as between the four Varnas. Again, the system of graded inequality is not merely notional. It is legal and penal. Under the system of Chaturvarnya, the Shudra is not only placed at the bottom of the gradation but he is subjected to innumerable ignominies and disabilities so as to prevent him from rising above the condition fixed for him by law. Indeed until the fifth Varna of the Untouchables came into being, the Shudras were in the eyes of the Hindus the lowest of the low. This shows the nature of what might be called the problem of the Shudras[1]

The Word "Dalit" comes from the Sanskrit root dal and means broken, ground –down, downtrodden or oppressed. Those previously known as untouchables, depressed classes etc. "Dalit" refers to one's caste rather than class; it applies to member of those menial castes which have born the stigma of "untouchability" because of the extreme impurity and pollution connected with their traditional occupations. Dalits are "outcastes" falling outside the traditional four-fold caste consisting of the hereditary Brahmin, Kshatriya, Viashya and Shudra classes; they are considered impure and polluting and are therefore physically and socially excluded and isolated from the rest of society. [2]

Dalits started blogs and websites and expressed their emotions and thoughts to the public. Dr. Babasaheb Ambedkar is an icon of most Dalit websites in India and he gained popularity recently among the public. There is several Dalit blogs function in the country for the strengthen of Dalits such as upliftthem.blogspot.com, Major Dalit websites in India are Dalitnetwork.org, roundtableindia.co.in, Ambedkar.org, Dalitfreedomnetwork.com, fdrambedkar.org, navsarjan.org, Dalitindia.com, dsnuk.org (Dalit Solidarity Network-UK), Dalitstudies.org.in, idsn.org (International Dalit Solidarity Network), Dalitfoundation.org.

Only few scholars analysed Dalit cyberspace in India. Dalit websites in India have different features. Dalits use the cyberspace to enrich and empower their identity and subjectivity.

Studies of digital culture in India have ignored the subaltern presence in cyberspace. Very few studies indicate some realities of Dalit internet interventions.

International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 4, Issue 4, ISSN No. 2455-2143, Pages 206-210
Published Online August 2019 in IJEAST (http://www.ijeast.com)

Dalit websites placed Dr. Babasaheb Ambedkar as an icon and there are several websites in the name of Dr. Babasheb Ambedkar like ambedkar.org, drambedkar.org etc. they present the ideas of Ambedkar on history and politics.

II. MOTIVATION OF THE STUDY

Today people of all ages and from all over the world use web for collecting opinions. An important part of our information-gathering behaviour has always been find out what people thinks, with the growing availability and popularity of opinion-rich resources such as online review sites and personal blogs, new opportunities and challenges arises as people now can and do, actively use information technologies to seek out and understand the opinions of others. The demand for information on opinion and sentiment, "What other people think" has always been an important piece of information for most us during the decision-making process. [3] The uses are hunger for and reliance upon online advice and recommendations. There are many Dalit web portals which allows used to express their opinion such as ambedkar.org, jaibhim.co.in, utharakalam.com, navayan.com, Dalitica.org, Dalitindia.com, ambedkar.blogspot.in, samatha.in, ambedkaree.com, Dalitawaz.com etc.

Web content mining aims to extract useful information from contents of the web page. It involves scanning of all the contents on a web page to find its relevance with the search query. Opinion mining is kind of web content mining.

Sentiment analysis (or opinion mining) is defined as the task of finding the opinion of others about specific entities. Opinion mining is recent discipline of information retrieval and of computational linguistics which is document is about but the opinion it expresses. Opinions are very important for anyone who is going to make decision. Sentiment analysis or opinion mining system also has an important potential role as enabling technologies for other system.

III. LITERATURE REVIEW

Sentiment analysis or opinion mining extracts the subjective information from the source materials such as reviews using techniques such as natural language processing, and text analytics. Opinion plays essential part in our information-gathering behaviour before taking a decision. Online review sites, and personal blogs facilitate gathering of sentiments of products or object using information technologies. The main objective of Opinion mining is to determine the polarity of comments (positive, negative or neutral) by extracting features and components of the object that have been commented on in each document [4, 5].

Jeonghee Yi et al. (2003) proposed a Sentiment Analyser to extract opinions about a subject from online data documents. Sentiment analyser uses natural language processing techniques. The Sentiment analyser finds out all the references on the subject and sentiment polarity of each reference is determined. The sentiment analysis conducted by the researchers utilized the sentiment lexicon and sentiment pattern database for extraction and association purposes. Online product review articles for digital camera and music were analysed using the system with good results.[6]

Lun-Wei Ku et al.(2005) present techniques for automatic opinion summarization based on topic detection. The system selects representative words from a document set to identify the main concepts in the document set. A term is considered to represent a topic if it appears frequently across documents or in each document. The authors use many weighting mechanism to detect the representative words (topic words) at sentence, paragraph or document level. The identified topic then further used for opinion summarization. [7]

Chopra Rohit (2006) opines in his article "Global primordial ties: virtual identity politics in Online Hindutva and Online Dalit discourse" that there is deep ideological difference between Hindu Websites and Dalit Websites. The differences in the field of technology and culture in India which occurred in 1990s came with new dimensions in the representation of collective identity which is called as 'Global primordially'. The Hindu movements in the country spread their ideologies through internet which do other oppressed and marginalized communities. His "Technology and Nationalism in India: Cultural Negotiations from Colonialism to Cyberspace" gives detailed information about techno cultural Hindu nationalism. The Dalit websites in India stand against Savarna ideology and Hindu nationalism.[8]

Bo Pang et al.(2008) used machine learning techniques to investigate the effectiveness of classification of documents by overall sentiment. Experiments demonstrated that the machine learning techniques are better than human produced baseline for sentiment analysis on movie review data. The experimental setup consists of movie-review corpus with randomly selected 700 positive sentiment and 700 negative sentiment reviews. [9]

Opinion mining (often referred as Sentiment Analysis) refers to identification and classification of the viewpoint or opinion expressed in the text span; using information retrieval and computational linguistics. The opinion expressed on the topic is given significance rather than the topic itself [10].

Studies of digital culture in India have ignored the subaltern presence in cyberspace. Very few studies indicate some realities of Dalit internet interventions.. Nayar Pramod K (2011) identified the major characteristics of Dalit cyberspace by examining some websites in his article "The Digital Dalit: Subalternity and Cyberspace". Central to the Dalit websites is the construction of a different or alternative history of India. Focusing on the historically sanctioned and severely oppressive caste system, websites offer a sociological account of the Hindu social order that enabled the upper castes to subjugate the lower. He says that the website dalit freedom network created a transnational frame to read India's dark history and it compared the caste system to slavery.[11]

International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 4, Issue 4, ISSN No. 2455-2143, Pages 206-210
Published Online August 2019 in IJEAST (http://www.ijeast.com)

Zhu et al. (2011) proposed aspect based opinion polling from free form textual customers reviews. The aspect related terms used for aspect identification was learnt using a multi-aspect bootstrapping method. A proposed aspect-based segmentation model, segments the multi aspect sentence into single aspect units which was used for opinion polling. Using a opinion polling algorithm, they tested on real Chinese restaurant reviews achieving 75.5 percent accuracy in aspect-based opinion polling tasks. This method is easy to implement and are applicable to other domains like product or movie reviews.[12] One way to extract information is text mining and sentiment analysis, that include: data acquisition, data pre-processing and normalization, feature extraction and representation, labelling, and finally the application of various Natural Language Processing (NLP) and machine learning algorithms.

IV. Purpose of the Study

In the context of traditional Hindu society, Dalit status has often been associated with occupations regarded as ritually impure, such as any involving leather work, butchering or removal of rubbish, animal carcasses, and waste. Dalits worked as manual labourers cleaning streets, latrines, and sewers. Engaging in these activities was considered to be polluting to the individual, and this pollution considered contagious. As a result, Dalits were commonly segregated, and banned from full participation in Hindu social life. Elaborate precautions were sometimes observed to prevent incidental contact between Dalits and other castes.[13]

The information revolution and information technology have played big role in democratizing knowledge and information. Educated Dalits gained golden chance to be a part of main stream. Internet provided them opportunities to express their views and ideas. There is no untouchability and caste based censorship or gate keeping in the internet.

Studies of digital cultures in India have ignored the subaltern presence in cyberspace. We need to see cyber-culture as an extension of techno-culture and therefore factor in the uneven, even prejudicial, movement of technology, accounting for the role of institutional, commercial, state and social actors and their power relations in determining India's Information Technology (IT) policy. [13]

Central to Dalit websites is the construction of a different or alternative, history of India focusing on historically sanctioned and severely oppressive caste-system. Websites offer a sociological amount of the Hindu social order that enable the upper-caste to subjugate the "lower'. In the case of such particular website, what we get is transnational frame in which to read slavery , "Dalit constitute the largest number of people categorized as victims of modern day slavery".

Dalit from around the world can now meet and discuss issues and concern that affects them. Textual information in world can be broadly classified into two main categories, fact and opinions. Facts are objective statements about entities and events in the world. Opinions are subjective statements that reflect people's sentiments or perceptions about the entities and events.

Dalits are a mixed population, consisting of numerous social groups from all over South Asia; they speak a variety of languages and practice a multitude of religions. The term Dalit has been interchangeably used with term scheduled castes, and these terms include all historically discriminated communities of India outcaste and Untouchables. While discrimination based on caste has been prohibited and untouchability under the constitution of India, discrimination and prejudice against Dalits in South Asia remains.

Dalits realized the advantage of internet and they started blogs and websites to express their thoughts and views. Dalits challenged the main stream narratives of caste and history and presented counter narratives. But the scholars and researchers tried to neglect the efforts of Dalits and they were not ready to conduct deep studies on Dalit cyberspace. Savarna intellectuals were thinking about the vicious role of Dalit websites.

A comparison of local Dalit cyberspace with national level will prove that the local internet activism is very low. It will also provide information about the nature and feature of Dalit engagements and discourses through examining the contents of websites. Such study can help other oppressed sections of the society to understand the advantages of new media and modern technology.

V. Proposed System

The universe of this study is contents of a website. It analyses all contents about Dalit issues and the contents presented in Dalit perspective. This study will also check the response from reader's feedback. For sentiment analysis different machine learning techniques can be used. Supervised learning, semi-supervised learning and unsupervised learning algorithm can be used. This mining classifies an evaluative text or feedback contents as being positive or negative sentiments or opinion.

A. Part-of-Speech(POS):

Part-of-Speech (POS) tagging is the process of assigning a part-of-speech like noun, verb, pronoun, adverb, adjective or other lexical class marker to each world in a sentence. The input to tagging algorithm is a string of words of a natural language sentence and finite list of part-of-speech tags. The output is a single best POS tag for each word. Stanford Tagger is an application that analyze sentence for POS tagging. After POS tagging on a sentence, data structure form' of phrase-structure tree [14].

B. Opinion Word Extraction:

If a sentence contains one or more product features and one or more opinion words, then the sentence is called an opinion sentence. Opinion words are extracted in the following manner

- For each sentence in database, if sentence is opinion sentence then extract all the adjective words as opinion words.
- For each feature consider each nearby adjective as its opinion word.

After opinion word extraction in feature pruning, redundancies are reduced from extracted opinion word. [15]

C. Polarity Identification:

In this step orientation of an opinion sentence, i.e., positive or negative is predicted. In general, the dominant orientation of the opinion words is used in the sentence to determine the orientation of the sentence. That is, if positive/negative opinion prevails, the opinion sentence is regarded as a positive/negative one. [16-17]

For finding out opinion word polarity following steps are followed.

- If opinion words are present in database then assign polarity stored in database.
- If opinion words are not present in database then find its synonym. If synonym is present then assign polarity of that synonym to the opinion word and store it in database.
- If opinion words are not present in database and its synonym is not present, then find its antonym. If antonym is present then assign opposite polarity of that antonym to the opinion word and store it in database.

For finding out opinion word polarity following steps are followed.

- If odd number of negation words is present in the opinion sentence then assign opposite polarity of opinion word to that sentence.
- If even number of negation words is present in the opinion sentence then assign polarity of opinion word to that sentence.
- If negation word is not present in the opinion sentence then assign polarity of opinion word to that sentence.

D. Summary Generation:

To generate the final review summary following steps are used:

- For each discovered feature, related opinion sentences are put into positive and negative categories according to the opinion sentences' orientations. A count is computed to show how many reviews give positive /negative opinions to the feature.
- All features are ranked according to the frequency of their appearances in the reviews. Feature phrases appear before single word features as phrases normally are more interesting to users. Other types of rankings are also possible.[18]

Tiwari et. al (2017), uses content based approach for the online audits, film ratings etc. using sentiment analysis. These reviews were gathered by supervised machine learning strategies. [19]. Arun et. al (2017), they retrieved the data and then transformed it into text files as input dataset. Then sentiment analysis was done after removing the stop words followed by determining the polarity of the words and classifying the text as positive and negative. [20]

VI. CONCLUSION

This study provides primary information about Dalit web portals in India to understand the presence of Dalits in this new media and analyze the feedback about of Dalit website. This study has identified major features of ambedkar.org website. It examined ideas and concepts provided by Dalits. The method adopted is sentiment analysis, which aims to extract emotions and opinions from text. A basic goal is to classify text as expressing either positive or negative emotion.

VII. REFERENCE

[1] Dr. Ambedkar B. R. (1946) “Who were the Shudras?” vol. 1 Messer, Charles Scribner’s sons Publication, New York

[2] Oxford Sanskit English Dictionary, World (1964)

[3] Bo Pang; Lee Lillian; Vaithyanathan Shivakumar (2002) “Thumbs up? Sentiment classification using machine learning techniques”, In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), (Pg 79–86)

[4] Turney P (2002) "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews", ACL‟02

[5] Dave D; Lawrence A; Pennock.D; (2003) "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews", Proceedings of International World Wide Web Conference (W;W‟03).

[6] Jeonghee Yi, J ; Nasukawa T. ; Bunescu R; Niblack W.(2003) “Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques”, In: Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM-2003). Melbourne, Florida.

[7] Lun-Wei Ku;Li-Ying Li;Tung-Ho Wu; Chen Hsin-Hsi (2005) Major topic detection and its application to opinion summarization. In Proceedings of the SIGIR, (Pg.627– 628)

[8] Chopra Rohit (2006) “Global primordial ties: virtual identity politics in Online Hindutva and Online Dalit discourse”, New Media Society 8(2) (Pg.187-206)

[9] Pang Bo (2008) “Opinion Mining and Sentiment Analysis, foundations and Trends in information Retrival” Vol 2 No 1-2 (2008), 1-135

[10] Bing Liu (2008) Exploring User Opinions in Recommender Systems, Proceeding of the second KDD workshop on Large Scale Recommender Systems and the Netflix Prize Competition", Aug 24, 2008, Las Vegas, Nevada, USA.

[11] Nayar Pramod K (2011) The Digital Dalit: Subalternity and Cyberspace , www.academia.edu/1482588, The Shri Lanka Jopurnal of Humanities, XXXVII (1 &2), (Pg.69-74)

[12] Zhu, Jingbo Wang; Huizhen Zhu; Muhua Tsou; Benjamin K. ; Matthew Ma, (2011) "Aspect-Based Opinion Polling from Customer Reviews", IEEE Transactions on Affective Computing, Volume: 2,Issue:1 (Pg. 37-41)

[13] Balasubrahmaniam J. (2011) Dalits and a Lack of Diversity in the Newsroom', Economic and Political Weekly

[14] Tribhuvan Padmapani; Tribhuvan Amrapali (2011) Opinion Mining of Customer Reviews-ICKE (International Conference for Knowledge Engineering)-2011 Volume 1(Pg. 46-50)

[15] Tribhuvan Padmapani; Tribhuvan Amrapali (2013), Feature Based Opining Mining and Summarization of Product Reviews- IJMR (International Journal of Multidisciplinary Research) Vol. 1, issue 12 (VIII) March 2013, ISSN: 2277-9302. (Pg.21-23)

[16] Tribhuvan Padmapani; Tribhuvan Amrapali (2014) A Peer Review of Feature Based Opining Mining and Summarization"- IJCSIT(International Journal of Computer Science and Information Technologies) Vol. 5(1), ISSN: 0975-9646, (Pg. 247-250)

[17] Tribhuvan A.P. ; Jadhav M.E. (2015) Sentiment Analysis of Ambedkar.org, a Dalit Web Portal in India. Advances in Computational Research, ISSN: 0975-3273 & E-ISSN: 0975-9085, Volume 7, Issue 1, (Pg.203-205)

[18] Tribhuvan Amrapali Prakash. (2015) Study of content analysis within and across Dalit Media , Galaxy Link Vol-III, Issue –II, ISSN: 2319-8508 (Pg. 6-12)

[19] Tiwari, P.; Mishra, B. K.; Kumar, S.; Kumar, V. (2017). Implementation of n-gram methodology for rotten tomatoes review dataset sentiment analysis. International Journal of Knowledge Discovery in Bioinformatics (IJKDB), 7 (1), (Pg.30-41)

[20] Arun, K.; Srinagesh, A.; Ramesh, M. (2017). Twitter Sentiment Analysis on Demonetization tweets in India Using R language. International Journal of Computer Engineering in Research Trends, 4 (6), (Pg.252-258)