SENTIMENT ANALYSIS OF ONLINE COMMENTS WITH SPECIAL FOCUS ON MIXED CODE –A REVIEW

Reseena Mol N.A[1] *Dr.S.Veni*[2]

ABSTRACT

Extensive amount of data, which is in unstructured or semi structured form, is generated in social media by every day. Sentiment Analysis is the method of digging out subjective information from online feedbacks. And the main challenge with sentiment analysis is handling mixed code comments. This paper reviews the problems in handling mixed-code online comments and different techniques used for sentiment analysis of mixed code online comments. Beneficiaries of this paper are researchers, teachers, and students who have keen interest in this topic.

Key words: Lexicon based, Machine learning, Mixed code, Natural Language Processing, Neural Network, Sentiment Extraction.

1. INTRODUCTION

Sentiment Analysis is the combined usage of computational language processing and natural language processing techniques to mine hidden opinion from text subjective to the context. As people feel free to express their opinions through social media, now it is possible to extract their attitude or sentiment towards an item or for an event. As in general, two challenges were identified while analyzing the sentiment of online comments. First one is excessive amount of data are generated daily and the processing of these large amount of data is very complex. Second one is most commonly people were using the use of mixed codes for expressing their opinions.

[1]Research Scholar, Department of Computer Science
Karpagam Academy of Higher Education, Coimbatore.

[2]Professor, Department of CS, CA & IT
Karpagam Academy of Higher Education, Coimbatore.

In this paper trying to review on the hitches of mixed code sentiment analysis. People use mixed code comments because of Attitude change, as they want to communicate with minimum number of words or lack of vocabulary, as they can't stick on to a single language. Sentiment Analysis of monolingual data work very efficiently with the help of Conventional sentiment analysis algorithms, but shows several problems while implementing for multiple language code mixed data [3]. As Indian languages doesn't have a rich semantics and lexical tools, it's a complex process to detect the subjectivity from mixed code post.

This paper is organized in different sections. Section II the details of sentiment analysis and its techniques are explained.Literatre Review is presented in Section III and conclusion are given in Section IV.

2. SENTIMENT ANALYSIS METHODS

Different Sentiment Analysis methods are:

A. Lexicon based Sentiment Analysis

To find out the sentiment score of sentences, predefined rules are used for identifying positive, negative or neutral scores.

B. Machine learning based Sentiment Analysis

Most commonly used is this one and popular algorithms based on machine learning methods, are Naïve Bayes, Support Vector Machine, Decision Tree and k-nearest Neighbor.

C. Deep Learning based Sentiment Analysis

It is new technique which depends on number of layers in a Neural Network. This method demands more research in this area for better result.

3. LITERATURE REVIEW

According to Gazi Imtiyaz Ahmad, Jimmy Singla, Nikita [2], Indian people have a tendency of using social media to express their opinions regarding politics, movies, sports etc. As they were not so much expert over English language they are using either the regional languages like Tamil, Malayalam, Bengali etc. or the mixed codes like English-Hindi, English-Malayalam etc. to post their comments. As Indian languages doesn't have a rich semantics and lexical tools, it's a complex process to detect the subjectivity from mixed code post.This paper also explained three methods for sentiment analysis: Lexicon based, Machine learning based and Deep learning based sentiment analysis.

As stated by Shalini K, Aravind Ravikumar, Vineetha R C, Aravinda Reddy D, Anand Kumar M and Soman K.P [3], with fast growing usage of social media, there is a proportionate growth in posting of mixed code comments in facebook,twitter etc. This is mainly due to attitude change of today's user, as they want to communicate with minimum number of words and lack of vocabulary, as they can't stick on to a single language. Proposed method includes a hidden layer within convolutional neural network for performing sentiment analysis. Precision of the output obtained by using this method when applied for code mixed. Bengali data was 0.732, but the precision fall down to 0.513, when applied with Telugu movie, since Telugu language is morphologically more rich and have more synonyms.

As reported by Elisa Claire Alemán Carreón, Hirofumi Nonaka, Toru Hiraoka, Minoru Kumano, Masaharu Hirot, Takao Ito [4] they conducted a study on emotional response of Chinese customers about Japanese hotels, their needs and demands,they extracted keywords from reviews from the Chinese portal site Ctrip using entropy calculations from a manually classified sample of data to and used these in machine learning experiments. Initially, they conducted experiments with different kernels for SVM, but finally decided to use the linear kernel for the benefit of weight factor obtainability. Using the weight vectors of classifiers, as well as frequency of the words in data set, they concluded that Chinese customers have a preference for big and clean rooms, big thermal baths or bathhouses,expect good cost performance regardless of price.

According to Zia-ul-rehmna, Syed Anwar Hasnain, [5],they have performed study on sentiment analysis of different languages like English, European languages like German, Italian, Spanish etc. and Asian languages like Chinese, Arabic, Bengali etc. During the last five years these analysis have been done mainly through three techniques,Machine learning based techniques, Lexicon based methods and Hybrid techniques. From the works analyzed, authors concluded that the most commonly used one is Lexicon based but Machine learning techniques gives more accuracy. They have suggested the idea that sentiment analysis will be very useful to take act upon is based on the comments which have been posted in social media like facebook, twitter etc. based on an event.

As stated by Pravalika A, Vishvesh Oza, Meghana, N.P and Sowmya Kamath, S [6], Authors narrated Sentiment Analysis as the process of detecting the opinion from comments by applying computational language processing and natural language processing. Nowadays most commonly people uses bilingual code mixed text for expressing their opinions. And the problem is that traditional sentiment analysis algorithms are well played with monolingual text. Here the authors takes this issue and proposed an evaluated two methods¬--Lexicon based and Machine learning approach to extract the sentiment of mixed code text from social media. In the first approach, by taking regularly using Hindi and English abbreviated words and slangs for representing the opinions about films a domain for lexicons were formed. Sentiment of the sentence is extracted by sentiment combination rules which were derived as per the lexicon rules based on the sentiments of each words. In

the second sentiment extraction approach, a prototype is required to recognize the sentiments and that training model is made based on the mixed data included in facebook comments which in turn helps to study the semantics and regularly happening rules. Naïve bayes, Support Vector Machine, Decision Tree, Reandom Tree Multilayer Perception Tree were applied for sentiment classification after the development of feature vector. And the methods were evaluated with real world mixed code data and reached at the conclusion that lexicon based achieves better accuracy i.e. by 86%, but machine learning approach attaines an accuracy of only 72%.

As per Yonas Woldemariam [7], compared two categories of sentiment analysis methods, lexicon-based and machine learning approaches to determine which one will be suitable to identify the sentiments from forum discussion posts. With Lexicon based algorithm, author used apache-hadoop framework and used Stanford coreNLP library with the Recursive Neural Tensor Network (RNTN) model. In Lexicon based method, polarity is assigned to each word after tokenizing it from the sentence calculated polarity sum value of a sentence to make a conclusion about the sentence, as it belongs to a positive, negative or neutral one. Sentiment Treebank having 215, 154 phrases which is characterized using Amazon Turk is utilized for training purpose in the later method.. As a result of experimental evaluation RNTN have more precision than lexicon-based by 9.88% accuracy on all types of comments, but with Lexicon based provides better performance on classifying positive comments and F1-score values of the Lexicon-based is greater by 0.16 from the RNTN.

As per Rupal Bhargava, Yashvardhan Sharma and Shubham Sharma [8], they had proposed to develop a system for mining sentiments from mixed code data, where the base language is English and combined commonly with four different regional languages of India (Tamil, Telugu, Hindi and Bengali). As it is a complex task, it is partitioned into two segments, i.e. identifying the Language and Mining of Sentiments. Performance analysis of Language identification algorithm and Sentiment analysis algorithm were done. Obtained results are compared with already set benchmarks on machine translated sentences in English, and found to be around 8% better in terms of precision.

According to Shashank Sharma,PYKL Srinivas,Rakesh Chandra Balabantaray [9], main challenge with feedback in social media is mixed code, i.e., the mixture of two languages. Word of one language will be represented phonetically with the help of other language. They considered English –Hindi Combination here. They identified the normal way of writing mixed code text, which includes Phonetic Typing, Abbreviation, Word play, Intentionally misspelt words and Slang words. Separated Hindi words from the mixed text and then transliterated to Romanized English language. With Lexicon based approach, identified the sentiments of sentence into a positive or negative. The proposed model acquired 85% accuracy.

As stated by Prof.Dinkar Sitaram, Ms.Savitha Murthy, Debraj Ray, Devnash Sharma,Kashyap Dhar [10], sentiment analysis of mixed code text is very complex as it doesn't have any clear grammatical structure. Proposed system performs sentiment analysis at different levels i.e. at expression level and also at sub expression levels. Most commonly repeating grammatical transitions are studied by training a classifier on the mixed language training data. RNTN (Recursive Neural Tensor Network) is used as classifier in this system. Here experiments are restricted to mixed code of two languages Hindi and English and done on short text that comes on social media like facebook, twitter etc. As stated by Xing Fang and Justin Zhan [11] , explained Sentiment analysis or opinion mining as the field where analysis of public 's attitudes or feelings on a context is being done. Here the authors identified sentimental polarity categorization as the ultimate problem of sentiment analysis. They conducted the study on Online product

reviews from Amazon.com. They have proposed method for polarity categorization and also the F1 scores of sentence level categorization and review level categorization have performed.

4. CONCLUSION

Sentiment analysis can be used to extract the attributes of expressions like the polarity i.e. either a positive, negative or neutral statement, subject, i.e. the item that is being discussed and the opinion owner i.e. the person or entity that expresses the opinion so, extracting sentiment of social media comments will be very useful and there is a possibility to take actions based on the comments in social media upon an event. This paper presents a brief overview of sentiment analysis of mixed code, its problems and different techniques for sentiment analysis. From the works analyzed, identifies the fact most commonly used one is Lexicon based, but Machine learning techniques gives more accuracy. Deep Learning methods are not that much popular in detecting sentiment.

References

[1] Bo Pang1 and Lillian Lee2," Opinion mining and sentiment analysis", Foundations and Trends in Information Retrieval Vol. 2, No 1-2 (2008) 1–135 , 2008

[2] Gazi Imtiyaz Ahmad,Jimmy Singla,Nikita Former Assistant Professor,School of CSE, "Review On Sentiment Analysis Of Indian Languages With A Special Focus On Code Mixed Indian Languages",2019 International Conference on Automation, Computational and Technology Management (ICACTM) Amity University,352,978-1-5386-8010-0/19/$31.00 ©2019 IEEE.

[3] Shalini K, Aravind Ravikumar, Vineetha R C, Aravinda Reddy D, Anand Kumar M and Soman K.P, "Sentiment Analysis Of Indian Languages Using Convolutional Neural Networks ",2018 International Conference on Computer Communication and Informatics (ICCCI -2018), Jan. 04 - 06, 2018, Coimbatore, INDIA.

[4] Elisa Claire Alemán Carreón, Hirofumi Nonaka, Toru Hiraoka, Minoru Kumano, Masaharu Hirota, Takao Ito , "Emotional contribution analysis of online reviews", ICAROB2018,vol.23,pages-359-362,ISSN-2188-7829.

[5] Zia-ul-rehmna,Syed Anwar Hasnain," A Comprehensive Study On Sentiment Analysis",Journal of ISOSS 2017 Vol.3(1),145-156.

[6] Pravalika A, Vishvesh Oza, Meghana N P and Sowmya Kamath S,Department of Information Technology,National Institute of Technology Karnataka, Surathkal, Mangalore ,"Domain-Specific Sentiment Analysis Approaches For Code-Mixed Social Network Data"
8th ICCCNT 2017,July 3 - 5, 2017, IIT Delhi,

[7] Yonas Woldemariam,Department of Computing,Science Umea University,Umea, Sweden,yonasd@cs.umu.se," Sentiment Analysis In A Cross-Media Analysis Framework",10.1109/ICBDA.2016.7509790,IEEE,Hangzhou, China.

[8] Rupal Bhargava, Yashvardhan Sharma and Shubham Sharma,Department of Computer Science & information System,Birla Institute of Technology & Science, Pilani Campus Pilani, "Sentiment Analysis For Mixed Script Indic Sentences",2016 Intl. Conference on Advances in Computing, Communications and Informatics (ICACCI), Sept. 21-24, 2016, Jaipur, India

[9] Shashank Sharma,PYKL Srinivas,Rakesh Chandra Balabantaray," Text Normalization Of Code Mix And Sentiment Analysis ",2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 978-1-4799-8792-4/15.

[10] Prof.Dinkar Sitaram,Ms.Savitha Murthy,Debraj Ray ,Devnash Sharma,Kashyap Dhar.," Sentiment Analysis Of Mixed Language Employing Hindi English Code Switching",Proceedings of 2015 Intl. Conference on Machine Learning and Cybernetics ,Guhangzhou 12-15 July 2015.s

[11] Xing Fang and Justin Zhan,," Sentiment Analysis Using Product Review Data",Journal of Big data (2015),DOI: 10.1186/s40537-015-0015-2.

[12] Se, Shriya, R. Vinayakumar, M. Anand Kumar, and K. P. Soman. "Predicting the sentimental reviews in Tamil movie using machine learning algorithms." Indian Journal of Science and Technology 9, no. 45 (2016).

[13] Jain, Anuja P., and Padma Dandannavar. "Application of machine learning techniques to sentiment analysis." In Applied and Theoretical Computing and Communication Technology (iCATccT), 2016 2nd International Conference on, pp. 628-632. IEEE, 2016.

[14] Seshadri, Shriya, Anand Kumar Madasamy, Soman Kotti Padannayil, and M. Anand Kumar. "Analyzing sentiment in indian languages micro text using recurrent neural network." IIOAB J 7 (2016): 313- 318.