ISSN XXXX XXXX © 2019 IJESC

Research Article Volume 9 Issue No.3

Corpus Based Dual Sentiment Analysis

Mohan .I[1], Divya Rani .G [2], Pranathi .M .J .A[3]
Assistant Professor[1], Student[2, 3]
Department of Information Technology
Prathyusha Engineering College, India

Abstract:
Sentiment analysis is an important current research area. The importance of sentiment analysis coincides with the growth of social media such as reviews, blogs, and social networks. Sentiment analysis is also called as opinion mining. We propose a system called Dual Sentiment Analysis which is better in performance when compared to Bag-Of-Words (BOW). It is used to analyse and compare the public opinion for the product review. The Dual Sentiment Analysis framework works with polarity classification i.e, positive, negative, neutral (3 classes classification). In this analysis, we use Dual prediction Classification. Dual Prediction classifies test review by considering two sides of a review. To avoid the dependency on external sources, we develop corpus-based method to build a pseudo-antonym dictionary to reverse the review. At the end, result shows the effectiveness of Dual Sentiment Analysis in sentiment classification.

Keywords: Bag-Of-Words, Dual prediction classification, Opinion Mining, Sentiment analysis.

I. INTRODUCTION

To understand what others think, which is been an important part in decision making process. In the past, when an information was needed for decision making, we take a survey for others opinion. Now a days, due to the growth of social media the source for the decision making has been transformed to online, web logs, twitter, social networks, product rating sites, chat rooms etc where the people express their views on almost anything in online. Sentiment Analysis is used for gathering opinions of the public. Sentiment Analysis enables us to know the thought of each and every person. It enables positive and negative evaluations. This book is interesting [Positive] and This book is boring [Negative]. It is a special text mining task. Initially Bag of Words model is used in order to find the sentiment of comments. This is a statistical machine learning. But, the performance of bag of words is limited due to some fundamental deficiencies n handling polarity shift problem. Though it is very simple and efficient, it is not recommendable because it discards some semantic information. Major disadvantage is polarity shift problem. Polarity shift is a type of linguistics circumstances which reverse the polarity of the text. i.e, the sentiment of the text will be reversed from positive to negative and vice versa. For example, positive text," I like Sweets" will be reversed from positive to negative when the negation word, "don't" added to the positive text. This is the major drawback of Bag- Of-Words model under the polarity.

II. RELATED WORKS

[1] Dual Sentiment Analysis: Considering two sides of one Review, Rui Xia, Feng: This was processed to handle the polarity shift problem. They have proposed the Dual Sentiment Analysis model. They used Dual Training and Dual Prediction Algorithm. They developed Corpus-based method to build a Pseudo-Antonym dictionary for review reversion.
[2] Sentiment Classification using Machine Learning Techniques. This focused on topical categorization, attempting to sort documents according to their subject matter (e.g., sports vs. politics). This model used supervised learning approach such as naive Bayes, and support vector machines to sort entire reviews into negative or positive sentiments. In this, sentiment analysis seems to require more understanding than the usual topic-based classification. They concluded results generated by standard machine learning methods are rover to result by human-generated baselines. Yet machine learning method performs well on only traditional topic-based categorization and lack in functionality on sentiment classification.
[3] This system proposes a rule-based multivariate text feature selection method called Feature Relation Network (FRN) that considers semantic information and leverages the syntactic relationships between n-gram features. They used Feature Relation Network (FRN) that has decision tree models, recursive feature elimination, and genetic algorithms. FRN empowers the consideration of heterogeneous n-gram features for improved opinion classification, by joining syntactic data about n-gram relations. FRN selects the features in a more computationally effective way than numerous multivariate and hybrid methods. But Noise and redundancy in the feature space increase the likelihood of over fitting.
[4] Sentiment Analysis of Blogs by Combining Lexical Knowledge with Text Classification. In this paper, they present a unified framework in which one can use background lexical information in terms of word-class associations, and refine this information for specific domains using any available training examples. knowledge-based approaches and learning-based approaches are used. By diverse domains show that this approach performs better than using isolation. The issue of sentiment analysis is further complicated by the fact that bloggers often use jokes and cultural references to illustrate their opinions, making the labeling task unclear for people unfamiliar with the relevant facts or ref
[5] Effects of Adjective Orientation and Gradability on Sentence Subjectivity To compute the subjectivity of a sentence. This system considers two such features semantic orientation, which represents an evaluative characterization of a word's deviation and gradability, which characterizes a word's ability to express a property in varying degrees. supervised classification technique such as semantic orientation, Gradability are used. This technique retains subjective sentences and discards the objective sentences.

After that they applied sentiment classifier. Major task of this sentiment classifier is to contemplate resulted subjectivity with enhanced results.

[6] Hidden Sentiment Association in Chinese Web Opinion Mining to categorize review documents into positive or negative in which as thumbs up represented positivity of document and thumbs down represents negativity of document an unsupervised learning method was stated. This categorize review documents into positive or negative. This technique has limitation as it relies on external search engine.

III. PROPOSED SYSTEM

We develop a simple and efficient sentiment analysis called dual sentiment analysis (DSA).

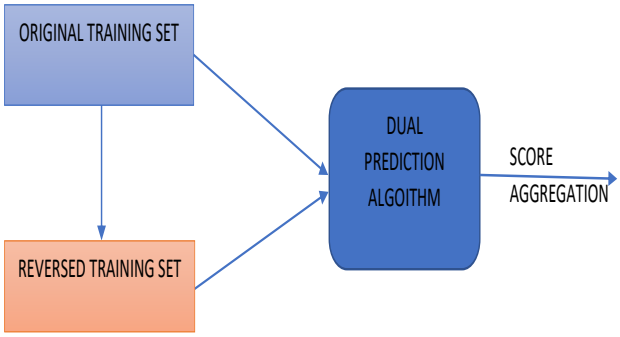

Figure.1. architecture of sentiment analysis

In our system, we use unsupervised learning approach in which we develop a dictionary-based method for the natural language processing. We use dual prediction algorithm to overcome the polarity shift problem, which can lead to misclassification. In this algorithm, by removal of negative words in the reversed review the positive classification will be high. Therefore, the combination of both original and reversed review to produce the accumulated score. For example, the REVIEW TEXT is I don't like this mobile then,

	CLASS	SCORE
ORIGINAL REVIEW BEFORE INVERT DICTIONARY	Negative positive	0
REVERSED REVIEW AFTER INVERT DICTIONARY	negative	-1

We develop two dictionaries, positive and negative dictionary to analyse and predict the sentiment of the review. With the help of these two dictionaries, the basic sentiment score is calculated. Some expressions have more positive or more negative than others to define its strength we have two new dictionary one is incrementor and another is decrement or dictionary to improve the sentiment score. For example , "good" has more strength than "barely good" but less than "very good". We extend our framework from 2-class polarity classification (positive, negative) to 3-class polarity classification (positive, negative, neutral).

REVIEW	CLASS	SCORE	Score after dictionaries
This book is very interesting	positive	1	2
This book is barely good	positive	1	0

The above table explains about the impact of increment or and decrement or dictionaries, the review "This book is very interesting" comes under the positive class but the word "very" is a strong word so the score was incremented. Similarly, the review "This book is barely good" comes under the positive class but the word "barely" is decrementor word so the score was decremented. Finally, we develop a corpus-based method to construct a pseudo-antonym dictionary, which removes DSA's dependency on an external antonym dictionary for review reversion. Apparently, this can reduce some prediction errors caused by polarity shift.

IV. STEPS IN SENTIMENT ANALYSIS

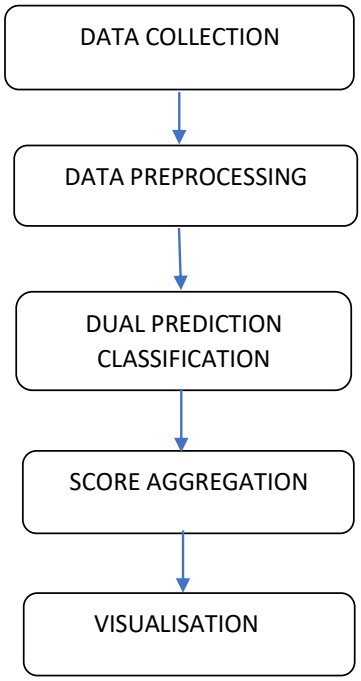

Figure.2. modules of sentiment analysis.

A. DATA COLLECTION

The Data we are analyzing in our process is restaurant reviews. So, the respective domain related sentiment words that can be present in review must be analyzed and added to the dictionaries. The data has been collected from the Kaggle website of the Japanese restaurant. Data was scraped from table log. Table log is a crowd-sourced restaurant-rating services. This site has largest restaurant-review site which has 5.9 million reviews of 800,000 lists of restaurants Japan. You can't only find reviews of restaurants but also you can refer to information of each restaurants and eating places in Japan.

B. DATA PREPROCESSING

Data pre-processing is an important step in the data mining process. This technique transforms the raw data into understandable format. It is a method of resolving issues such as incomplete and inconsistent data. We use NLTK library (natural language toolkit) for pre-processing task. There are some steps in natural language processing (NLP) such as:

i. TOKENIZATION

Tokenization is the process of breaking up sequence of string into words, phrases, symbols, or other meaningful elements which is called tokens. In this process of tokenization, some characters like punctuation marks are discarded.

ii. POS TAGGER

Part-of-Speech Tagger (POS tagger) assigns parts of speech to each word (tokens) such as noun, verb, adjective, etc.

C. DUAL PREDICTION CLASSIFICATION

In dual prediction classification, the original reviews get reversed with the help of the pseudo-antonym dictionary build using the corpus-based method. This method is called as

dictionary-based method, which is an unsupervised learning method. We develop two dictionaries positive and negative to categorize the sentiments of the reviews. Then, the pseudo-antonym dictionary or invert dictionary is used to find the invert words, which changes the sentiment of the review. Additionally, we have two dictionaries incrementors and decrementors for the improvement of the sentiment score.

D. SCORE AGGREGATION

Score aggregation is process of calculating the accumulated score of the dataset. This score is calculated using the classified sentiment and its calculation. Score aggregation identifies the recommendation percentage (%) based on the reviews in the dataset.

V. FUTURE ENHANCEMENTS

This paper can be further enhanced by classifying the sentiment of emoticons. Emoticons are the expression of opinion or sentiment. These are the complex task to identify the emoticon which is still under the research and can be enhanced in future.

VI. CONCLUSION

Sentiment analysis is in research for years. In this paper we have found the sentiment of the comments whether it is positive, negative or neutral. We propose a corpus based pseudo-antonym dictionary or invert dictionary, to address the polarity shift problem in sentiment classification. Since it is an unsupervised learning method (Dictionary based method), there is no need of prior training to the system. Limitation are classifying the sentiment of emoticons which will be enhanced in the future.

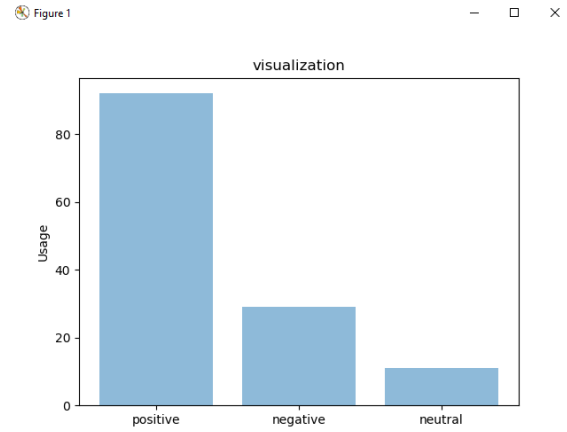

Figure.3. Polarity shift classification on restaurant reviews.

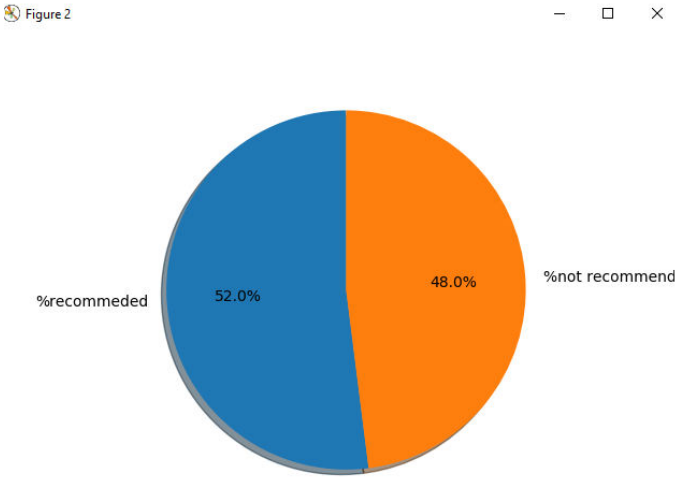

Figure.4. Overall recommendation percentage of the restaurant.

VII. REFERENCES

[1]. Rui Xia, Feng Xu, Chengqing Zong, Qianmu Li, Yong Qi, and Tao Li "Dual Sentiment Analysis: Considering Two Sides of One Review," IEEE Trans. Know.Data Eng,vol. 27, no. 8, Aug. 2015.

[2]. Sentiment Classification using Machine Learning Techniques" Bo Pang and Lillian Lee Department of Computer Science Cornell University Ithaca,NY 14853 USA Shivakumar Vaithyanathan IBM Almaden Research Center 650 Harry Rd. San Jose, CA 95120 USA

[3].A. Abbasi, S. France, Z. Zhang, and H. Chen, "Selecting attributes for sentiment classification using feature relation networks," IEEE Trans. Knowl. Data Eng., vol. 23, no. 3, pp. 447–462, Mar. 2011.

[4]."Sentiment Analysis of Blogs by Combining Lexical Knowledge with Text Classification" Prem Melville IBM T.J. Watson Research Ctr. P.O. Box 218 Yorktown Heights, NY.

[5]."Effects of Adjective Orientation and Gradability on Sentence Subjectivity" Vasileios Hatzivassiloglou Department of Computer Science Columbia University New York, NY 10027, Janyce M. Wiebe Department of Computer Science New Mexico State University Las Cruces, NM 88003.

[6]. "Hidden Sentiment Association in Chinese Web Opinion Mining" Qi Su, Xinying Xu1, Honglei Guo, Zhili Guo, Xian Wu, Xiaoxun Zhang, Bin Swen1 and Zhong Su Peking University IBM China Research Lab2 Beijing, 100871, China Beijing, 100094, China.

[7]." A Survey of Dual Sentiment Analysis Considering Two Sides of One Review" Ashwini Damodar Kanherkar, Sarika M. Chavan.

[8]. E. Agirre, and D. Martinez, "Exploring automatic word sense disambiguation with decision lists and the web," in Proc. COLING Workshop Semantic Annotation Intell. Content, 2000, pp. 11–19.

[9]. J. Cano, J. Perez-Cortes, J. Arlandis, and R. Llobet, "Training set expansion in handwritten character recognition," in Proc. Struct.,Syntactic, Statistical Pattern Recognit., 2002, pp. 548–556.

[10]. Y. Choi and C. Cardie, "Learning with compositional semantics asstructural inference for subsentential sentiment analysis," in Proc.Conf. Empirical Methods Natural Language Process., 2008, pp. 793–801.

[11]. I. Councill, R. MaDonald, and L. Velikovich, "What's great and what's not: Learning to classify the scope of negation for improved sentiment analysis," in Proc. Workshop Negation Speculation Natural Lang. Process., 2010, pp. 51–59.

[12]. S. Das and M. Chen, "Yahoo! for Amazon: Extracting market sentiment from stock message boards," in Proc. Asia Pacific Finance Assoc. Annu.Conf., 2001.

[13]. M. Hu and B. Liu, "Mining opinion features in customer reviews," in Proc. AAAI Conf. Artif. Intell., 2004, pp. 755–760.

[14]. D. Ikeda, H. Takamura, L. Ratinov, and M. Okumura, "Learning to shift the polarity of words for sentiment

classification," in Proc. Int. Joint Conf. Natural Language Process., 2008.

[15]. W. Jin, H. H. Ho, and R. K. Srihari, "Opinion Miner: A novel machine learning system for web opinion mining and extraction," in Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2009, pp. 1195–1204.

[16]. S. Kim and E. Hovy, "Determining the sentiment of opinions," in Proc. Int. Conf. Comput. Linguistic, 2004, pp. 1367–1373.

[17]. A. Kennedy and D. Inkpen, "Sentiment classification of movie reviews using contextual valence shifters," Comput. Intell., vol. 22, pp. 110–125, 2006.

[18].I.Mohan, M.Moorthi, "SENTIMENT CLASSIFICATION ON SOCIAL NETWORK DATA" International Journal of Pharmacy &Technology Volume-9 Issue-2 Pages 29775-29785 2017

[19]. I. Mohan,"Knowledge Discovery using Big Data" Journal of Current Computer Science and Technology Volume 5 Issue 05 2015