Tracing Dynamics of Opinion Behaviors with a Cross-Domain Sentiment Classification Algorithm based on Sentiment Related Index

Anu Ganesan, Arockia Babi Reebha
PG Scholar, Head of the Department
Computer Science and Engineering
Pavendar Bharathidasan College of Engineering and Technology
Mathur, Thiruchirappalli, India

ABSTRACT

Sentiment analysis has become one of the most active research areas in natural language processing. Its application is spread across business services to Research areas. Opinion target extraction means collecting the opinion from the user. It allows users to shared small elements of content such as short sentences, individual images, or video links that may be the major reason for their popularity. In propose a novel cross-domain sentiment classification algorithm based on sentiment related index (SRI), which is termed SentiRelated, to analyze the sentiment polarity for short texts. SentiRelated utilizes SRI to expand feature vectors based on unlabeled data from the target domain. In this way, some significant sentiment indicators for the target domain are added to feature vectors. At last, validate SentiRelated algorithm on two typical cross domain twitter datasets. Based on the Unigram, Bi-gram and associative techniques of sentiment analysis the target domain was analyzed and review was categorized. Finally the output compared with the SVM algorithm in R data studio to identify the accuracy of the classification algorithm. The experimental results show that, compared with existing algorithms, SentiRelated algorithm can improve the performance of cross-domain sentiment classification for short texts.

Keywords - Content Analysis, Sentiment Classification, Opinion Mining, SentiRelated index, Unigram, Bigram, SVM algorithm

I. INTRODUCTION

With the rapid development of social media, sentiment analysis, also known as opinion mining, has become one of the most active research areas in natural language processing. Twitter is an web micro blogging tool that disseminates more than 400 million messages per day, together with huge amounts of information about almost all industries, from entertainment to sports, health to business and also it also consists of reviews from consumers about the products. It's straightforward to use both for sharing information and for collecting it. All the above characteristics make twitter a best place to collect real time and latest data to analyze and do any sought of research for real life situation.

Sentiment Polarity Analysis gives an idea about the recent development of opinion minion that was based on the content of the user's reviews. Content based sequential opinion influence framework to model opinion influence and employ it for opinion prediction. To predict the sentiment of the opinion propose two model with different prediction strategies, CSIM_S and CSIM_W. CSIM_S use the sentiment category label as the target and CSIM_W obtains sentiment based on the sentiment scores of all fine-grained opinion words.

Opinion target extraction collecting the opinion from the user based on their review comments. It validates Cross Domain Sentiment Classification algorithm on three typical datasets. One is considered as target that was analyzed by using the two trained datasets. The experimental results show the comparison of support vector machine algorithms with Cross Domain Sentiment Classification that define the accuracy of content based sentiment classification for reviews.

This paper has following contributions :

1. The traditional sentiment analysis performed based on the user's reviews not based on the content of the user's review.
2. Extract the customer reviews of Music, Automobile and Office domains from

twitter. The Office dataset considered as target dataset, that was sentiment analyze by using the Music and Automobile domains.

3. The sentiment analysis done by using Unigram, Bigram and also based on Domain independent features. Based on the threshold value the review was analyzed and the polarity was identified.
4. The analyzed output file was uploaded in R studio and it compare with support vector machine algorithm to produce a chart that represent the accuracy of sentiment analysis.

II. RELATED WORK

In today's world Social media has become a vital part to make the business decision. Through social media the manufacturing companies can able to identify and Predict the Market based on the customer's review content that was posted on the social network. While prediction, companies only focus on the sentiment based on few polarity index. With the aim of accurate prediction, it should focus the content as well.

The research paper [1] H. Sankar, V. Subramaniyaswamy, "INVESTIGATING SENTIMENT ANALYSIS USING MACHINE LEARNING APPROACH", International Conference on Intelligent Sustainable Systems (ICISS) (2017) uses the feature selection technique that treats the document as a sentence of the document or group of words. Each sentence is represented in the form of feature vectors which contains occurrence or non-occurrence of a feature in a binary form which represents the number of occurrence. [2] Lu Ma, Dan Zhang, Jian-wu Yang, Xiong Luo ., "SENTIMENT ORIENTATION ANALYSIS OF SHORT TEXT BASED ON BACKGROUND AND DOMAIN SENTIMENT LEXICON EXPANSION", 2016 5th International Conference on Computer Science and Network Technology (ICCSNT) aims at identifying the opinions, emotions, and evaluation the natural language expresses. Extract the sentiment orientation of the evaluation object by analyzing the short text. Judging the sentiment orientation of short text is equivalent to classify short text as positive, negative and neutral. The method contains two parts: the sentiment orientation analysis of short text and the background of the evaluation object.

[3] Shokoufeh Salem Minab, Mehrdad Jalali, Mohammad Hossein Moattar, "ONLINE ANALYSIS OF SENTIMENT ON TWITTER", 2015 International Congress on Technology, Communication and Knowledge (ICTCK). It extracts and filter the most related characteristics then classify events that become matrix in the previous phase by using the decision vector machine. [4] Shulong Tan, Yang Li, Huan Sun, Ziyu Guan, Xifeng Yan, Jiajun Bu, Chun Chen Xiaofei He, "INTERPRETING THE PUBLIC SENTIMENT VARIATIONS ON TWITTER", IEEE Transactions on Knowledge and Data Engineering (Volume: 26 , Issue: 5 , May 2014). It analyze public sentiment variations on Twitter and mine possible reasons behind such variations. To track public sentiment, combine two state-of-the-art sentiment analysis tools to obtain sentiment information towards interested targets in each tweet.

III. EXISTING SYSTEM

To assess the performance of state-of-the-art sentiment analysis systems in tweet sentiment classification across three distinctive Twitter data sets. Existing system used the sentence extraction approach that is used to extract the opinion for summarization. It aims to summarize mixed opinions about a topic and generate a list of contrastive pairs of sentences with different sentiment polarities. Formally framed the problem as an optimization problem. To solve the optimization problem using sentence extraction. To extract comparable sentences from each set of opinions and generate a comparative summary containing a set of contrastive sentence pairs. To measure the content similarity of two sentences in the same group of opinions either both are positive or both are negative. This similarity function allows us to assess which sentences are good representatives of each group. And then, measure the contractiveness of a positive sentence and a negative sentence. Since a good pair of contrastive sentences are generally also similar in content but opposite in sentiment polarity, also call this measure a cross group similarity measure. Problem of these approach take these sentences with different polarities as input and further generate a contrastive opinion summary.

It has following Drawbacks:

1) Sentence extraction difficult to find the optimal opinions.
2) Redundancy occur.
3) Finding opinion target extraction and opinion summarization is very hard.

IV. PROPOSED SYSTEM

In propose a novel cross-domain sentiment classification algorithm and content based sentiment analysis algorithm based on term frequency, to analyze the sentiment polarity for short texts. It expand feature vectors based on unlabeled data from the target domain. In this way, some important sentiment indicators for the target domain are appended to feature vectors. At last, validation of algorithm on one target dataset by using two typical datasets. The project, mainly focus on positive and negative sentiment reviews. The first strategy is to identify the reviews as positive or negative by using the positive and negative words used in the review comments. Then expand features based on the co-occurrence frequency between a candidate of additional related feature and a domain-independent feature. Compared with point wise mutual information, sentiment related index considers the distributions of word occurrences instead of the co-occurrence frequency between different words, thus surmounting the challenge caused by infrequent features and words. Then calculate weightage and ranking in each opinions using content analysis algorithm. Once that was calculated then compare the result with the existing approach in R studio.

The overall process has been divided into five core module as follows.

1) Dataset Collection
2) Preprocessing and Analysis the Sentiment Polarity
3) Measure Association of Co-occurrence using PMI
4) Calculate the Ranking for Opinion
5) Comparison Result

A. Dataset Collection

The proposed system, aim to tackle the problem of sentiment polarity categorization, which is one of the fundamental problems of sentiment analysis. Data used for the sentiment analysis are online product reviews of different domains that was collected from twitter.com.

B. Pre-processing and Analysis the Sentiment Polarity

The first task to be done is the pre-processing task which is one of the important tasks in sentiment analysis work. It will clean the dataset by reducing its complexity in order to prepare the data for the classification task. Firstly, the dataset was tokenize to split up the words into terms of tokens, then the stemming will reduce the tokens into a single type, normally a root word.

C. Measure Association of Co-occurrence using PMI

The first strategy is to expand features based on the co-occurrence frequency between a candidate of additional related feature (candidate in short for the rest of this paper) and a domain-independent feature. In information theory, Pointwise mutual information (PMI) is commonly used to measure the association between two different elements based upon the co-occurrence frequency of elements. To overcome the challenge of data sparsity for product reviews, here propose a novel strategy for expanding features based on Sentiment Related Index (SRI). Similar to pointwise mutual information, sentiment related index is used to measure the association between different lexical elements (unigrams and bigrams) in a specific domain.

D. Calculate the Ranking for Opinion

Calculate ranking based on twitter product review. PMI remove the cross domain co-occurrence words using word net. Then calculate weightage and ranking in each opinions using SentiRelated algorithm, SVM algorithm and enhanced enhanced SentiRelated algorithm.

E. Comparison Result

To compare the performance of proposed SentiRelated algorithm and enhanced SentiRelated algorithm with existing state-of-the-art algorithms classification tasks on two cross domain datasets. It shows the accuracy of the result of the proposed and existing approach.

V. IMPLEMENTATION and RESULTS

The input is the customer's review from the twitter that was in natural language entered by the user. Below statistics was compared with the existing and content based approach accuracy.

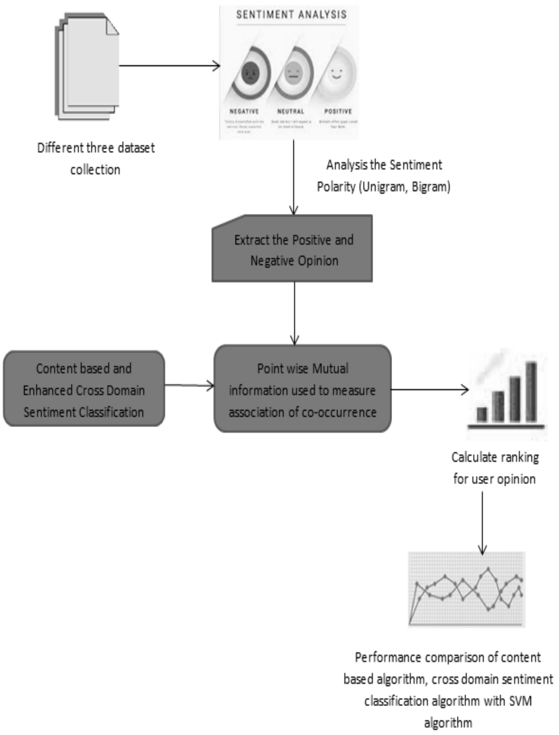
SENTIMENT ANALYSIS
NEGATIVE
NEUTRAL
POSITIVE
Different three dataset collection
Analysis the Sentiment Polarity (Unigram, Bigram)
Extract the Positive and Negative Opinion
Content based and Enhanced Cross Domain Sentiment Classification
Point wise Mutual information used to measure association of co-occurrence
Calculate ranking for user opinion
Performance comparison of content based algorithm, cross domain sentiment classification algorithm with SVM algorithm

Fig 1 : Architecture of the Proposed system

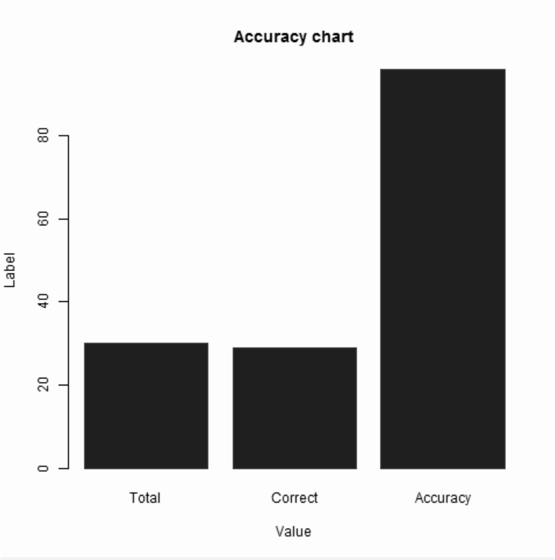

Fig 2 : Result Comparison of Existing and Proposed approach

VI. CONCLUSION and FUTURE WORK

Proposed approach can be used to extract both facts and opinions from social media content. It uses two different cross domain datasets to analyze the sentiment of another domain. Most of the time the rating of a specific reviews plays an important role for sentiment analysis. Because in tweets there is a chance to include the duplicate review. Because of that duplicate review there is a chance to incorrect analysis for a specific product. To avoid this along with the content of a specific review the additional attributes like Rating and helpful to be used for finding the term frequency. From that the duplicate reviews can be easily filtered out from the analysis. Hence the accuracy of the sentiment polarity will be very high.

VII. REFERENCES

[1] Daniele Cenni, Paolo Nesi, Gianni Pantaleo, Imad Zaza., “Twitter vigilance: A multi-user platform for cross-domain Twitter data analytics, NLP and sentiment analysis”, In Proceedings of the IEEE International Conference.

[2] H. Sankar, V. Subramaniyaswamy, "INVESTIGATING SENTIMENT ANALYSIS USING MACHINE LEARNING APPROACH", International Conference on Intelligent Sustainable Systems (ICISS) (2017).

[3] Lavika Goel, Anurag Prakash, “Sentiment Analysis of Online Communities Using Swarm Intelligence Algorithms”, 2016 8th International Conference on Computational Intelligence and Communication Networks (CICN).

[4] Lu Ma, Dan Zhang, Jian-wu Yang, Xiong Luo ., “SENTIMENT ORIENTATION ANALYSIS OF SHORT TEXT BASED ON BACKGROUND AND DOMAIN SENTIMENT LEXICON EXPANSION”, 2016 5th International Conference on Computer Science and Network Technology (ICCSNT).

[5] Shokoufeh Salem Minab, Mehrdad Jalali, Mohammad Hossein Moattar, “ONLINE ANALYSIS OF SENTIMENT ON TWITTER”, 2015 International Congress on Technology, Communication and Knowledge (ICTCK).

[6] Shulong Tan, Yang Li, Huan Sun, Ziyu Guan, Xifeng Yan, Jiajun Bu, Chun Chen Xiaofei He, "INTERPRETING THE PUBLIC SENTIMENT VARIATIONS ON TWITTER", IEEE Transactions on Knowledge and Data Engineering (Volume: 26 , Issue: 5 , May 2014).

[7] Desheng Dash Wu, Lijuan Zheng, David L. Olson, “A DECISION SUPPORT APPROACH FOR ONLINE STOCK FORUM SENTIMENT ANALYSIS”, IEEE Transactions on Systems, Man, and Cybernetics: Systems (Volume: 44 , Issue: 8 , Aug. 2014).

[8] Oussalah M, Bhat F, Challis K, Schnier T. A software architecture for Twitter collection, search and geolocation services, In Knowledge-Based Systems. Vol. 37, pp.105-120, 2013.

[9] Alexandre Trilla, Francesc Alias, "SENTENCE-BASED SENTIMENT ANALYSIS FOR EXPRESSIVE TEXT-TO-SPEECH, IEEE Transactions on Audio, Speech, and Language Processing (Volume: 21 , Issue: 2 , Feb. 2013).

[10] Ruhi U., Social Media Analytics as a Business Intelligence Practice: Current Landscape & Future Prospects, In Journal of Internet Social Networking & Virtual Communities, 2014.