CROSS DOMAIN OPINION MINING USING MAXIMUM ENTROPY BASED CLASSIFIER

V. Manimekalai¹, S. Gomathi @ Rohini²

¹Department of Computer Technology, Dr. N. G. P Arts & Science College, Coimbatore, India
²Department of Computer Application, Sri Ramakrishna college of Arts and Science, Coimbatore, India

mkalai55@gmail.com¹, gomathi.s@srcas.ac.in²

Abstract—The current analysis is play a dominant role in opinion mining is additionally referred to as sentiment analysis because of clear volume of opinion made net resource like discussion type, review sites, blogs, and tweets area unit on the market in digital type. Sentiment analysis is that the field of study that analyzes client opinion, feedback, sentiment analysis, attitudes and feeling from communication. At intervals fraction second, we have a tendency to classify the text in several manner in several seconds. It’s one in all the active analysis areas in linguistic communication process [NLP]. There are a unit range of techniques we want to classify the Opinion reviews. The main problematic in the sentiment analysis is to understand the usage of negation and the taxonomy of positive and negative sentiments recorded by the users in the social group. The proposed approach that extracts and classifies opinion words from one domain called source domain and predicts opinion words of another domain called target domain using a semi-supervised approach, its combines modified maximum entropy and bipartite graph clustering. A comparison of opinion classification of reviews on four different product domains are presented. The results demonstrate that the proposed method perform relatively well in comparison to the other methods. Comparison of SentiWordNet of domain-specific and domain-independent words reveals that on an average 72.6% and 88.4% words, respectively, are correctly classified.

Keywords— Sentiment analysis, Navies Bayes, Support vector machine, Aspect extraction, Positive and negative

1. INTRODUCTION

Sentiment is an emotion or feelings in the customer’s feedback to be analysis in data mining process and the technique to extract and capture the data for analysis in language process and linguistic matter analysis it’s accustomed extract and analyze info from the net order to recognize the subjective opinion of a document or assortment of documents, like web log posts, reviews, news articles and social media feeds like tweets and standing updates Sentiment Classification aims at mining the planet Wide net text of product reviews by customers to classify the reviews into positive or negative opinions [1]. Automated opinion mining from the reviews is beneficial to both consumers and sellers.

The ever increasing use of e-commerce has produced vast consumer generated contents such as reviews. It is becoming imperative for organizations as well as consumersto collect these data for product analysis or rating according touser need.

The sentimental analysis sub task extracting the title and opinion words. How to extract the title and opinion words from the product reviews exactly is a hot research in a long time. The English sentences, need to be split into words. Secondly, the reviews are almost the words-of-mouth. That is said, the syntax of the most reviews doesn’t meet the specifications. For our feature extraction, we want to extract the opinion word and corresponding features from the reviews if possible. Recently, the pattern matching method based on
probability has been applied widely on feature extraction. However, in the former works, only the high frequency patterns in the pattern set would be extracted, and the researches were always, even only, interested in patterns with the high probability for identifying the subjective sentiment. In our work, based on the analysis on a large number of annotated data, [3] we set length, upper and lower limit probability and frequency thresholds to extract patterns with information of POS tags from our training corpus. A prototype system has been created to implement our task. The experiments carried out on multidomain corpus show that our approach achieves satisfactory precision and recall.

Labeled opinions are used to analyze the classifier. Practically, labeled opinions for every domain is not possible, as it is delimited by time and cost, while domain adaptation or transfer learning could be used to circumvent this limitation. In this study, we propose the approach of domain adaptable lexicon which predicts the polarity of lexicon of one domain using a set of labeled lexicon of another domain using a modified entropy algorithm. This algorithm uses enhanced entropy with modified increment quantity instead of traditional entropy algorithm. Dataset of different types of products containing textual reviews has been used for evaluation. Multiple experiments were carried out to analyze the algorithm using accuracy and F-measure [9]. We designed the approach in two phases: (i) preprocessing of dataset and (ii) applying classifier and clustering on dataset.

2. RELATED WORK

A. Product Rating Using Sentiment Analysis

Customer feedbacks are the milestones for the success functionality for the companies. A producer will get the correct result of his product from the customer feedback. He can make necessary changes to his product according to the feedback. But most users always fail to give their feedbacks. To avoid the difficulty of providing feedback, this paper focuses on the technique of providing automatic feedback on the basis of data collected from Twitter. These data streams are filtered and analyzed and feedback is obtained through opinion mining. Here we mainly analyze the data for mobile phones. The experiments have shown 80% accuracy in the sentimental analysis. Our framework is able to provide fast, valuable feedbacks to companies. The man gives his thoughts and interests in the social networking sites, about the various products he uses, this can be of very interesting for the industries to rate their products and to get the correct feedback from the people in different strata's of the society. We extracted the Twitter data using the Twitter 4J API [2]. Since Twitter 4j is a third party API it helps to retrieve the most recently posted tweets in a short amount of time. By using the unigram approach the frequency of the words are calculated and the overall product rating is being calculated. By using the reversedistance weighing method the feature detection process is implemented.

B. Comparative Study on Machine Learning Techniques in Sentimental Analysis

Sentimental Analysis is reference to the task of Natural Language Processing to determine whether a text contains subjective information and what information it expresses, i.e., whether the attitude behind the text is positive, negative or neutral. This paper focuses on the several machine learning techniques which are used in analyzing the sentiments and in opinion mining. Sentimental analysis with the blend of machine learning could be useful in predicting the product reviews and consumer attitude towards to newly launched product [7]. This paper presents a detail survey of various machine learning techniques and then compared with their accuracy, advantages and limitations of each technique. On comparing we get 85% of accuracy by using supervised machine learning techniques which is higher than that of unsupervised learning techniques.
C. Product Opinion Mining Using Sentiment Analysis on Smartphone Reviews

The author provides you with sentimental analysis of various smartphone opinions on smartphones dividing them Positive, Negative and Neutral Behavior. This is basically being obtained by studying the various posts being posted by varied number of users considering their areas of interest categorizing the smartphones. Analysis of plenty of words coupled in a sentence represent various sentiments of users and the various experiences and impact that product has given them. This analysis compiles a structural modeling approach and Bayesian Interface system to identify the polarity of the opinion which subsequently classifies positive and negative opinions.

The overall accuracy of the classifier thus trained using Naïve Bayes Classification technique was around 40% which was quite un-satisfactory to deal with and to rely on. So in order to have a reliable trained model which is able to classify the data according to our needs we have opted another approach which is SVM. The same data set which when parsed under this approach produces good result with an accuracy of around 90%.

D. Feature Extraction from Online User Reviews

Online reviews have a huge commercial value if analyzed correctly. Most common way of analyzing reviews is subjecting them to sentimental analysis. However, this does not provide fine grain actionable information due to the lack of specificity. This study suggests a mechanism to extract product features that have been discussed within the reviews there by allowing a more meaningful opinion analysis. This can be challenging as different features are expressed in different ways in reviews. The proposed method intends to use double propagation and rule-based mining to overcome the extraction problem. This feature extraction module is developed as a part of a larger online review analysis platform for electronic products which is currently being implemented. Finally, this paper presents the results observed in the experiments carried out on multiple datasets.

E. Sentiment analysis from product reviews using SentiWordNet as lexical resource

In the current social, technological and economic context, customers make their decisions based mostly on the opinion of other consumers. On the other side, companies need quick feedback from their customers in order to adapt to their needs in real time. The effective connection between these two aspects relies on opinion mining tools, which automatically process consumers’ reviews and opinions about products or services. This paper presents a semantic approach for a sentiment analysis application, which is based on using the SentiWordNet lexical resource. The experimental validation proved a 61% average rate of success of the application.

F. Extracting Product Features from Online Reviews for Sentimental Analysis

For elaborately understanding what product features the reviews focuses on, we propose an approach based on patterns to extraction features (titles). Through setting length, upper and lower limit probability and frequency thresholds, we extract patterns of POS tags and features from the training corpus. To enhance adaptability of the pattern set, we merge some fundamental patterns into a new fuzzy pattern. Then a pattern matching algorithm is applied to extract the titles and opinion words from the reviews. We conducted a platform to extract features from product reviews automatically, the result of our experiments shows that our approach is effective.
3. METHODOLOGY

Classification of opinions can be done using a modified maximum-entropy algorithm. The increment quantity is modified according to the importance of the measure of words as specified. The maximum entropy classifier is closely related to the Naïve Bayes classifier, except that it uses a search-based optimization to find weights for the features that maximize the likelihood of the training data. It can handle mixture of Boolean, integer, and real-valued features. It is also used when the conditional independence of the features cannot be assumed, in problems like text classification where features are words and are not independent.

A. Overview of the methodology

For our experiments, two different datasets were used. Each dataset consisted of labeled positive and negative text review documents. All the results from above sections reveal that the proposed approach gives better accuracy than baseline methods. Word is an important entity as it indicates sentiment or opinion of an object. The proposed framework is based on modified entropy classifier. Opinionated words are extracted based on the term frequency and inverse document frequency. Increment quantity is modified as granularity refined from document to word level which shows drastic difference between traditional maximum-entropy and modified entropy. Bipartite graph clustering is applied on classified data which has enhanced the results.

B. Modules

- Data set pre-processing
  - Stop Word Removal
  - POS tagging applying NLP
- Feature extraction
- Maximum entropy classifier
- Iterative Clustering

Dataset pre-processing

Data Preprocessing may be a technique is employed to alter the data into a clean data set. In alternative words, once the information is collected from completely different sources it collects in raw format that is not potential for the analysis.

The data area unit pre-processed exploitation Json program. The often used Stop words like, and, are, this, etc. area unit removed. Stemming words, Special characters, numbers, White areas etc. area unit removed.

Stop word Removal

Stop words are words which are filtered out in the preprocessing step. These words are, for example, pronouns, articles, etc. It is important to avoid having these words within the classifier model, because they can lead to a less accurate classification.

POS tagging Applying NLP

Natural language processing could be a platform that works out the quality structure of sentences, as an example, that teams of words along (as "phrases") and that words are the subjective word or objective word of a verb. Stanford language process (or NLP) could be a part of text mining that performs a special quite linguistic analysis that essentially helps a machine “read” text. The processes facilitate to extract the part of speech like verb, Noun, Subjective words, adverb etc. the method of distribution one in every of the part of speech to the given word is named elements Of Speech tagging. It’s ordinarily mentioned to as
POS tagging. It specifies nouns, verbs, adverbs, adjectives, pronouns, conjunction and their sub-categories during this tagging.

**Feature Extraction**

The Feature extraction was developed based on numerical/statistical analysis technique. This technique used to rank the features and differentiate classes. The feature extraction was used to reduce the dimensionality of dataset.

**Maximum entropy classifier**

Maximum Entropy is widely used for Natural Language Processing tasks, such as part-of-speech tagging, and language modeling and text segmentation. Maximum Entropy combines contextual features in a principled way and allows unrestricted use of them. Propose that the Maximum Entropy does not assume that the features are conditionally independent of each other. Based on the Principle of Maximum Entropy and from all the models that fit our training data, selects the one which has the largest entropy. The Max Entropy classifier can be used to solve a large variety of text classification problems.

**Iterative Clustering**

The clustering algorithm is applied on classified word lists and documents until it reaches convergence. All extracted words from source domain are tagged, and weight is calculated for each word using mutual information available for words. Target words are extracted and compared with the source. If they match then they will be categorized as domain-independent, otherwise domain-specific. Domain-independent words are from both source and target domains; whereas, domain-specific are from target domain only. Using domain-independent words weight is assigned to domain-specific words and classified accordingly. Each file form target domain is assigned score on which basis it is classified as positive or negative. Each word has weight assigned to it. Summation of weights of words in each sentence gives score to sentence. Then addition of all sentence score is nothing but score of file. On the basis of this, file is classified.

4. **EXPERIMENTAL RESULTS**

**Dataset Used**

Amazon Dataset: Electronics, books, cloths

**Precision**

Precision is a method of describing random errors and it is computed using the following equation

\[
\text{Precision } P = \frac{tp}{(tp + fp)}
\]

Where \(tp\) is true positive and \(fp\) is false positive

**Recall**

Recall is fraction of relevant data that are retrieved and it is computed using the equation.

\[
\text{Recall } R = \frac{tp}{(tp + fn)}
\]

Where \(fn\) is false negative
Accuracy

Helps to know how accurate the decision tree is formed. It is calculated with the help of

\[
\text{Accuracy} \ A = \frac{tp + tn}{tp + tn + fp + fn}
\]

Where \( tp \) is true positive and \( fp \) is false positive

![Maximum Entropy Based Classifier](image)

5. CONCLUSION

This mechanism to extract features from online user reviews. This aims to increase the value of sentimental or opinion analysis of such data, as it provides a more meaningful understanding into the customer perception. Its importance will further be enhanced, as internet is becoming the main source of information for user perception. This research also shows the importance of considering subjective features as well as objective features for opinion analysis, which would be useful in future research. Results during the experiments show that this approach is highly effective in feature extraction with the additional benefit of being domain independent. Future work of this research will mainly focus on expanding the proposed mechanism into a state where it can extract implicit features (instances where the feature word is not mentioned within the review) as well as extracting explicit features. This will further enhance the feature extraction and the subsequent opinion analysis of reviews.

6. REFERENCES

[1]. B. Liang, Q. Liu, J. Xu, Q. Zhou, P. Zhang. "Aspect-Based Sentiment Analysis Based on Multi-Attention CNN," Journal of Computer Research and Development. Chinese, vol. 54, No. 8, pp. 1724-1735, 2017.

[2]. J. Yuan, S. Mcdonough, Q. You, and J. Luo. "Sentrubite: image sentiment analysis from a mid-level perspective." In Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining, p. 10. ACM, 2013.
[3]. P. Aliandu, "Sentiment Analysis on Indonesian Tweet", in The Proceedings of the 7th ICTS, Bali, 2013, pp. 203 - 208.

[4]. Farhan Hassan Khan, Saba Bashir, Usman Qamar, TOM: Twitter opinion mining framework using hybrid classification scheme Decision Support Systems 57 (2014) 245257.

[5]. C. Hung and H.-K. Lin, "Using objective words in sentiwordnet to improve sentiment classification for word of mouth," IEEE Intelligents Systems, p. 1, 2013.

[6]. A. Pak and P. Paroubek, “Twitter as a corpus for sentiment analysis and opinion mining,” In LREc, vol. 10, May. 2010, pp. 1320-1326.

[7]. Varsha Sahayak, Vijaya Shete, Apashabi Pathan, “Sentiment Analysis on Twitter Data”, International Journal of Innovative Research in Advanced Engineering (IJIRAE) Issue 1, Volume 2 (January 2015) Page 1-178.

[8]. M.F. Caropreso, S. Matwin, F. Sebastiani, “A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization,” Text databases and document management: Theory and practice, 2001, pp. 78-102.

[9]. E. Cambria, B. Schuller, Y. Xia and C. Havasi, “New avenues in opinion mining and sentiment analysis”. IEEE Intelligent Systems, vol. 28, no. 2, pp. 15-21, 2013.

[10]. R. Mukras., J. Carroll. A comparison of machine learning techniques applied to sentiment classification, 2004.

[11]. Z ZAIN, "High Speed And Lowpower Gdi Based Full Adder”, Journal of VLSI Circuits And Systems, 1 (01), 5-9, 2019.

[12]. Pb Agus Ristono*,”Design Of Reliable And Efficient Manchester Carry Chain Adder Based 8-Bit Alu For High Speed Applications”, Journal Of VLSI Circuits And Systems, 1 (01), 1-4,2019.