ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

Volume 3, Special Issue 3, March 2014

2014 International Conference on Innovations in Engineering and Technology (ICIET'14)

On 21st&22ndMarch, Organized by

K.L.N. College of Engineering, Madurai, Tamil Nadu, India

Cross Domain Opinion Mining in Synonymically Structured Database

Merlin Archana,Karthikeyan.R

M. E Scholar, Dept of Computer Science & Engineering, PSNA College of Engineering & Technology, Dindigul, Anna University, Chennai, India.

Associate Professor, Dept of Computer Science & Engineering, PSNA College of Engineering & Technology, Dindigul, Anna University, Chennai, India.

Abstract — **Opinion** mining aims at classifying sentiment data into polarity categories positive (or) negative.Opinion mining is the field of analyze the people's opinions, sentiments, attitudes and emotions from written language. It has been important for many applications such as opinion summarization, opinion integration and review spam identification. On average, human process six articles per hour against the machine's throughput of 10 per second. However, the opinion information is often unstructured and/or semi-structured data in the internet. Online product reviews are often unstructured, subjective, and hard to digest within a short time period. The main objective of our proposed work is to determine the human opinion from text written in the web page automatically. Sentiment classification aims to automatically predict sentiment polarity of users publishing product based sentiment data. Applying sentiment classifier results in poor performance because each domain using different sentiment word. In order to train a binary classifier from one or more domains we propose a method to overcome the problem of existing cross domain sentiment classification methods. First we create a synonym database for both source and target domains and perform pos tagging. A product based sentiment classification using spectral clustering algorithm to align the domain specific words from different domains into unified clusters for opinion classification is developed. Sentiment sensitivity is achieved with the help of synonym database by measuring the distributed similarity between the words. To investigate the effectiveness of our method, we have compared it with several algorithms and develop a robust and generic cross-domain sentiment classifier.

Keywords— Cross domain Opinion mining, binary classifier, Thesaurus, Sentiment sensitivity.

I. INTRODUCTION

With the rapid growth of e-commerce over the past years, more and more products are sold on the web and more and more people are buying products online. A person wants to buy a product online he or she will typically start by searching for reviews and opinions written by other people on the various offerings. These are important for companies and individuals that want to monitor their reputation and get timely feedback about their products and actions. An increasing no of users are writing reviews. As a result no of product comment grows rapidly. Most popular products can get hundreds of comments in merchant web sites. Many comments are also long, which makes difficult for a customer to read them to make decision on whether to purchase a product or not. If he or she read a few reviews, he or she only gets an unreal judgment about a product. The larger number of reviews also makes it difficult for product manufacturers to keep track of customer opinions on their products and services [1]. It is thus highly difficult to produce a summary of reviews. Automatic aspect level opinion classification [2] is the task of classifying a given review with respect different aspect; because people talk about entities that have many aspect and they have a different opinion about each of the aspects

Opinion-mining systems analyze the piece of text in 3 steps. First, we must identify which part opinion is

expressing. Second, we need to identify who wrote the opinion. Finally, we analyze what is being commented in the review. There are two main techniques for sentiment classification. The symbolic technique uses manually crafted rules and lexicons. Sentiment classification system normally uses unsupervised or supervised learning to classify the sentiment data. Supervised learning algorithms like support vector classifier and naïve Bayesian classifier require labeled data for sentiment classification [3]. However sentiment word is expressed differently in each domain and it is costly to adapt data for each new domain in which we would like to apply the unsupervised learning algorithm [4] to construct an opinion classifier and it is not required labeled data. For example, in the *car* domain the words "Excellent" and "greater mileage" are used to express the positive sentiment, whereas "expensive" and "difficult" often indicate negative sentiment. On the other hand, if we are consider the *laptop* domain the words "longer" and "compact" express positive sentiment, whereas the words "shorter" and "low" express the negative sentiment. A classifier trained on one domain does not perform well on another domain. The cross domain opinion mining problem [5], [6] focuses on the challenge of training a binary classifier from one or more domains to classify the product based sentiment data.

We model the cross domain opinion classification problem as one of feature expansion, we are adding additional related feature to feature vector that represent the source and target domain reviews to reduce the feature mismatch problem between two different domains. Similar method is used in numerous tasks such as query expansion [7]. For example, in query expansion, a user query containing the word laptop might be expanded to *Laptop* OR *Data processor*, thereby retrieving documents uses either the term *Laptop* or the term *Data processor.*

We create synonymically structured database. It contains collection of thesaurus that aligns different words and it expresses the same sentiment for different domains [8]. Synonym database uses the dictionary based technique. A typical information retrieval task is to select documents from a synonym database in response to a user queries. Sentiment feature were extracted using the cross domain dictionary based technique from synonym database.

Feature of our proposed work can be summarized as follows:

- Product based sentiment classification using Spectral clustering algorithm to align domain-specific words from different domains into unified clusters for opinion classification is developed.
- Sentiment features were extracted using the cross domain dictionary-based technique
- The proposed system uses a distributional approach to construct a sentiment sensitive thesaurus using both labeled and unlabeled data from multiple domains.

II. RELATED WORKS

Opinion mining or Sentiment classification system can be categorized into Single domain [3], [9], [10] and Multiple domain [5], [6] classifiers based upon the domains. On another axis, sentiment classification system can be categorized into word level sentiment classification [10], sentence level sentiment classification [11] and document level sentiment classification [3]. In single domain sentiment classification, a classifier is trained using only labeled data. Kanayama and Nasukawa [9] propose an approach to build a domain oriented lexicon to identify the sentiment words

Cross domain sentiment classification has been recently attention with the advancement in the field of domain adaptation [12], [13] report a number of test on domain adaptation of sentiment classifiers. Blitzer et al. [5] propose a Structural Correspondence Learning (SCL) algorithm reducing the relative error due to adaptation between domains by an average of 30% over the original SCL algorithm and 46% over the supervised baseline, in this work they choosing pivot features using not only common frequency among domains but also mutual information with source labels.

In the existing work, to identify the source domain features are related to which target domain features and it is based on semi supervised adaptation method [8]. In the existing method un-described features for automatically classifying Web news, also validates the self-growth algorithm [14] for the cross-domain word list. Yue Lu et al. [15] propose a method to generate a rated aspect summary which provides a decomposed view of the over-all ratings and calculate the entropy based on the probability distribution of the words. Domain adaptation methods can be broadly classified into supervised and semi supervised based approach. Our proposed method fully based on unsupervised aspect sentiment classification. Recently it has been work on theoretical aspects of domain adaptation [12], [13].

III. PROBLEM SETTING

We define a domain D as a class of entities. For example, different types of products such as cars, laptops are considered as different domains [6]. Given a review written by the customer on a product that relates to a particular domain, the objective of this work is to determine the human opinion from text written in the web page automatically. Here we are training a binary classifier from one or more domains.

We consider a labeled source domain Dsrc and target domain Dtar is represented by L (Dsrc) and L (Dtar) [16]**,** it consist of set of pairs (r, c) where a review, r, is assigned a sentiment label, c. Here, c belongs to +1 or -1, respectively positive sentiment and negative sentiment.

In addition we denote a set of unlabeled source domain Dsrc and target domain Dtar is represented by U (Dsrc) and U (Dtar) [8]. Here our cross domain opinion classifier predicts the sentiment of both labeled and unlabeled data. It enhances the previously existing cross domain opinion classifier.

IV. A MOTIVATING EXAMPLE

Let us consider the reviews shown in Table 1 for two domains: *cars* and *laptops* [6], [8]. Table 1 shows two positive reviews and one negative review from each domain. We have given special importance to the words that express the sentiment of the author in a review using boldface. From Table 1 that the words *like*, excellent and *easy* are used to express a positive sentiment on *cars*, whereas the word *difficult* indicates a negative sentiment [6]**.** On the other hand, in the l*aptops* domain the words *Longer, like, light weight* and *compact* express a positive

TABLE I
POSITIVE (+) AND NEGATIVE (-) REVIEWS IN TWO DIFFERENT DOMAINS: **CARS** AND **LAPTOPS**

	Cars	***Laptops***
+	I would **like** to start by expressing my wishes towards nano red It was **excellent**	Lap top battery life is **longer** and also available with touch screen
+	It gives **greater** mileage and **easy** to drive	My brother **like** this laptop. It is **light weight** and **compact**
–	**Difficult** to travel long distance	Processing speed is **low**

sentiment, whereas the word *low* expresses a negative sentiment. Although words such as excellent, like would express a positive sentiment in both domains, and difficult a negative sentiment, it is unlikely that we would face with words such as greater mileage for laptops or battery life longer in reviews on cars [6]**.** Therefore, a model that is trained only using reviews on cars might not have any weights learned for battery life longer and light weight, which makes it difficult to accurately classify reviews on laptops using this model. One solution to this problem of feature mismatch we have to construct the synonym database for sentiment classification using multiple domains. Synonym database contains collection of thesaurus. It aligns different words from different domains and produce same sentiment for different domains [8], [16].

V. PROPOSED WORK

A. Data Set

We use the car (www.cars.com) and laptop (www.newegg.com) data set from shopping web sites and compare the proposed method against previous work on cross domain opinion mining. This data set consists of reviews for two different product types: cars and laptop [3]. Each review is assigned with a star ratings (0-5 stars), product name, user name, review date and time, comments and feedback. Comments with rating >3 are positive, whereas those with rating >3 are negative [8]. The data set also contains some unlabeled data for three different domains.

B. Data Preparation and Review Analysis

The data preparation step performs data preprocessing and cleaning on the datasets of car and laptop for the subsequent analysis. Most commonly used preprocessing steps is removing non-textual contents and markup tags and also removing information that are not required for opinion classification, such as review dates and reviewers' names. The review analysis step analyzes the features of reviews so that we have to retrieving the interesting information like product features or opinions, this step is important for opinion classification.

C. Sentiment Sensitive Thesaurus

In this section we construct a synonym database and create sentiment sensitive thesaurus using training data for data set [8]. Given a labeled or an unlabeled review, we first spilt the review into individual sentence. The main aim of sentence splitting is to identify the grammar and sentence structured used by the user. After splitting the entire comments, carry out part-of-speech (POS) tagging on each review sentence [17]. Sentiment sensitivity is achieved in the thesaurus by measuring the distributional similarity between words. Here we describe a method to construct a sentiment sensitive thesaurus for feature expansion [7].

D. Spectral Clustering Algorithm

Spectral clustering has become one of the most popular modern clustering algorithms. The goal of spectral clustering is to cluster data that is connected but not necessarily compact or clustered within convex boundaries. The clusters can be used to reduce the mismatch between domain-specific words of the two domains. Spectral clustering algorithms [6] form the clusters by unique features from different domains.

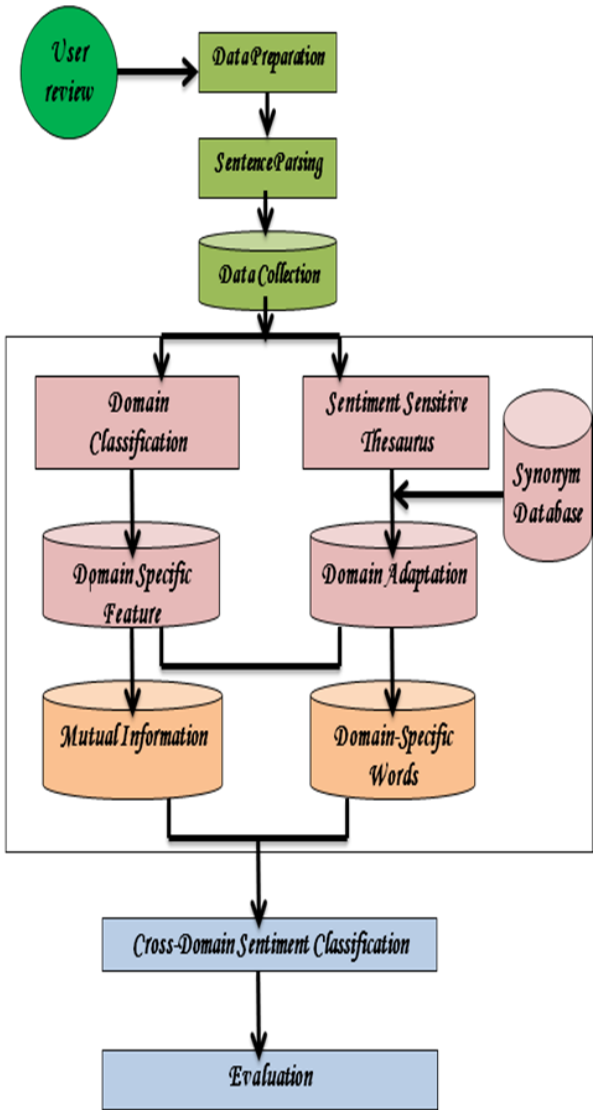

Fig.1 Proposed system for opinion classification

Finally we compare the mutual information with those clusters and classifying the comments for car and laptop domain. The classification of sentiment data expressed in terms of polarity. A stated opinion can be classified in either of these two categories: positive sentiment and negative sentiment. Spectral clustering algorithm classifies the sentiment data into above two terms.

Given a set of points $V = \{v1, v2, \dots, vn\}$ and their corresponding weighted graph *G*, the goal is to cluster the points into *k* clusters, where *k* is an input parameter.

1. Form the affinity matrix $A \in R$ ^n $\times n$, where$Aij = mij$, if$i \neq j$; $Aij = 0$.
2. Define the diagonal matrix D, and construct the matrix $L = D^{-\frac{1}{2}}AD^{-1/2}$
3. Find the Largest eigenvectors of L and form the matrix $X = Lij = \frac{Lij}{(\varepsilon jLij)^{1}}$
4. Apply the k-means algorithm on L to clusters n points into k clusters

The above spectral clustering algorithm [6], [18] is used for cross domain sentiment classification.

E. Experimental Evaluation

The main metric measured here is improving the accuracy with spectral clustering algorithm while classifying sentiment data. Here we take car and laptop data set. Compared to other algorithm spectral clustering provides high accuracy. And in using clustering it can be reduce the cap between two different domains. Here we use the Intel Pentium-IV 2.4 GHz to process and it uses 1 GB DDR-2 RAM to load and store requirements needed for the process execution

To identify domain-independent and domain-specific features us extracting two consecutive words from reviews and estimate the semantic orientation of the extracted phrases

We represent a lexical element feature vector u, and sentiment element w that co-occurs with u. Here we compute the point wise mutual information g (u, w) to measure the similarity distribution of words because we are implementing this for car and laptop domain and these data sets contains different words to express the sentiment While performing POS tagging, we are Calculate the Point wise mutual information Semantic orientation g (u, w)

$$g(u, w)=\log\left(\frac{\frac{c(u,w)}{N}}{\frac{\sum_{i=1}^{n} c(i,w)}{N} \times \frac{\sum_{j=1}^{m} c(u,j)}{N}}\right) \quad \dots (5.1)$$

Where c (u, w) is the number of sentence in each review
Where g (u, w) is the PMI (i.e., Point wise Mutual Information)

In cross domain opinion mining system we have more than one source domain. Selecting the one source domain to adapt given target domain for sentiment classification is the main problem. This problem is encountered with the help of synonym database. Comments with rating >3 are positive comment, whereas those with rating >3 are negative comments.

Car and Laptop Reviews are classified with the help of spectral clustering algorithm
Accuracy measures are calculated by

$$Accuracy = \frac{No\ of\ correctly\ classified\ reviews}{Total\ no\ of\ reviews\ in\ domains} \quad \dots (5.2)$$

Accuracy is calculated by ratio of number of correctly classified reviews to total number of reviews in domains.

VI. RESULTS AND DISCUSSION

Data sets of car and laptop reviews are classified. The results for each step explained in the proposed method.

We select the two source domains cars (C), and laptops (L). When we using two source domains we take 400 positive and 400 negative labeled reviews from each source domain and also we take unlabeled reviews from each domain. Here we classify car and laptop domain individually and combining both. While classifying the dataset individually it achieves low accuracy compared to

multiple domains. Below graphical representation shows the accuracy for sentiment classification.

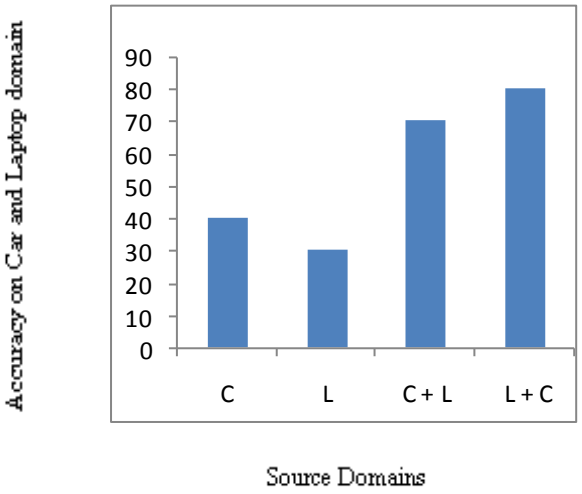

Fig.2 Multiple domain sentiment classification

Fig. 2 shows the effect of using multiple sources for sentiment classification. We see that the car domain is the best source domain when adapting to the laptop domain. A more interesting observation is that the accuracy that we obtain when we use two source domains is always greater than single domain. From the above graphical representation the classification accuracy is improved. This is because of using multiple domains.

After implementing the single and multiple domain sentiment classification, we are planning to compare the proposed method results with existing cross domain sentiment classification methods such as Structural Correspondence Learning algorithm, Spectral Feature Alignment algorithm and no adaptation method. Here we draw a table to measure and compare the accuracy between the different domains so that we have to improve the accuracy.

TABLE II
COMPARISON WITH EXISTING WORK

Method	**Laptop**	**Car**
No adapt	0.7261	0.7053
SCL	0.8206	0.7893
SST	0.8518	0.8386
Proposed(SC)	0.8940	0.8821

From the table, the SCL methods classify the sentiment data by measuring the mutual information between different domains and produce classification accuracy. The SFA method is using domain features to classify the sentiment. Our method is fully based on the thesauri database to classify the sentiment. Here the automatic created thesaurus to expand the feature vectors and classifying the data.

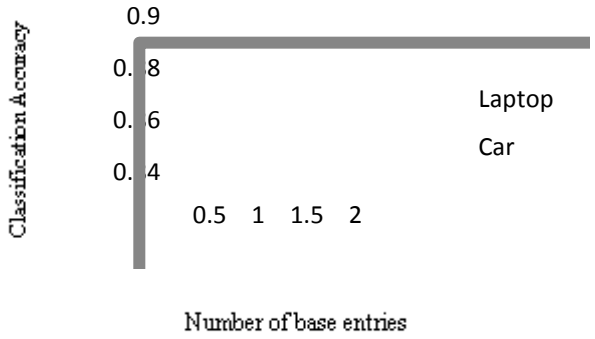

Fig. 3 performance of proposed method with size of thesaurus

Fig. 3 shows opinion classification accuracy for laptop and car dataset. Here we see that when we increase the size of thesauri database initially, the classification accuracy also increased with the help of feature expansion that is adding additional behavior to the feature vector from that we reduce the mismatch between two domains. However, when we further increase the size of thesauri database the accuracy drops and saturates. From the above graphical representation our proposed work produces high accuracy.

From above all graphical representation, our aim is to increase the sentiment classification accuracy and construct a generic and robust sentiment classifier.

VII. CONCLUSION AND FUTURE WORK

Typically, opinion mining system has been modeled as the problem of training a binary classifier using reviews for positive or negative sentiment. Sentiment analysis is the process of extracting knowledge from the people's opinion. Each domain using different sentiment word and annotating corpora for every possible domain of interest is costly. Hence Spectral clustering algorithm to align domain-specific words from different domains into unified clusters for opinion classification is developed. This suggested work is presented to create a sentiment sensitive distributional thesaurus to identify the product based sentiment data, and we are planning to improve the accuracy and performance of sentiment classification over multiple domains and compare the performance and accuracy with previously existing work.

ACKNOWLEDGEMENT

We would like to thank all those who gave us their support to complete this paper.

REFERENCES

[1] Xiaowen Ding, Bing Liu, Philip S. Yu, "A Holistic Lexicon-Based Approach to Opinion Mining", *WSDM'08*, February 11-12, Palo Alto, California, USA, 2008.

[2] Samuel Brody , Noemie Elhadad " An Unsupervised Aspect-Sentiment Model for Online Reviews Online Reviews" Human Language Technologies The Annual Conference of the North American Chapter of the ACL, pages 804–812, 2010.

[3] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs Up? Sentiment Classification Using Machine Learning Techniques", Proc. ACL-02 Conf. Empirical Methods in Natural Language Processing (EMNLP '02), pp. 79-86, 2002.
[4] R. G. ad R. Li and R. Chellappa. "Domain adaptation for object recognition: An unsupervised approach". In ICCV, 2011.
[5] J. Blitzer, M. Dredze, and F. Pereira, "Biographies, Bollywood, Boom-Boxes and Blenders: Domain Adaptation for Sentiment Classification", Proc. 45th Ann. Meeting of the Assoc. Computational Linguistics (ACL '07), pp. 440-447, 2007.
[6] S.J. Pan, X. Ni, J.-T. Sun, Q. Yang, and Z. Chen, "Cross-Domain Sentiment Classification via Spectral Feature Alignment", Proc. 19th Int'l Conf. World Wide Web (WWW '10), 2010.
[7] H. Fang, "A Re-Examination of Query Expansion using Lexical Resources," Proc. Ann. Meeting of the Assoc. ComputationalLinguistics (ACL '08), pp. 139-147, 2008.
[8] Bollegala, David Weir and John Carroll, "Cross-domain Sentiment Classification Using a Sentiment Sensitive Thesaurus", IEEE Transaction on Data and Knowledge Engineering, Vol. 25, No. 8, 2013.
[9] H. Kanayama and T. Nasukawa, "Fully Automatic Lexicon Expansion for Domain-Oriented Sentiment Analysis", Proc. Joint Conf. Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP '06), pp. 355-363, 2006.
[10] H. Takamura, T. Inui, and M. Okumura, "Extracting Semantic Orientation of Phrases from Dictionary", Proc. Conf. North Am. Ch. Assoc. for Computational Linguistics (NAACL '07), pp. 292-299, 2007.
[11] H. Yu and V. Hatzivassiloglou, "Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences," Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP '03), pp. 129-136, 2003.
[12] J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Wortman, "Learning Bounds for Domain Adaptation", Proc. Advances in Neural Information Processing Systems Conf. (NIPS '08), 2008.
[13] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J.W. Vaughan, "A Theory of Learning from Different Domains", Machine Learning, vol. 79, pp. 151-175, 2009.
[14] Lun Yan, Yan Zhang and G. Karypis, "News Sentiment Analysis Based on Cross Domain Sentiment Word List", ADMA @ springer-verlag berlinheidelberg, 2012.
[15] Yue Lu, Cheng Xiang Zhai, Neel Sundaresan, "Rated Aspect Summarization of short comments", Madrid, Spain ACM, 2009.
[16] Danushka Bollegala, David Weir and John Carroll, "Using Multiple Sources to Construct a Sentiment Sensitive Thesaurus for Cross-Domain Sentiment Classification", 49th Annual Meeting of the Association for Computational Linguistics, pages 132–141, 2011.
[17] Rui Xia and Chengqing Zong, "A POS-based Ensemble Model for Cross-domain Sentiment Classification", Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 614–622, 2011.
[18] Sajib Dasgupta, Vincent Ng, "Which Clustering Do You Want?Inducing Your Ideal Clustering with Minimal Feedback", Journal of Artificial Intelligence Research 39 , pp. 581–632, 2010.