Cross Domain Sentiment Classification Using Enhanced Sentiment Sensitive Thesaurus (ESST)

[A]P.Sanju, Department of CSE,University College of Engineering Villupuram,Villupuram

Sanjupani_79@yahoo.co.in

[B]T.T.Mirnalinee, Department of CSE, SSN College of Engineering ,Chennai,India
mirna123@gmail.com

***Abstract*— Sentiment classification is classification of reviews into positive or negative depends on the sentiment words expressed in reviews. Generally, sentiments are expressed differently in different domain and annotating label for every domain is expensive and time consuming. In cross domain sentiment classification, a classifier trained in source domain is applied to classify reviews of target domain which produce poor performance due to features mismatch between source domain and target domain. The proposed method develops solution to feature mismatch problem in cross domain sentiment classification by creating enhanced sentiment sensitive thesaurus using wiktionary. The enhanced sentiment sensitive thesaurus aligns different words in expressing the same sentiment not only from different domains of reviews and from wiktionary to increase the classification performance in target domain. Next, feature vector augmentation is performed using enhanced sentiment sensitive thesaurus while training a classifier. The proposed method performs a cross domain sentiment classification on a bench mark dataset Amazon product reviews for different types of products.**

Index Terms—Cross Domain sentiment classification, Enhanced sentiment sensitive thesaurus, Domain adaptation, Data Mining

I. INTRODUCTION

With rapid growth of internet, people write their opinion about the products in online. Such kind of opinions and sentiment information of the customers are overwhelming and growing exponentially which becomes a tedious work for the manufacturer to classify the reviews manually. An automatic sentiment classifier is classification of reviews into positive or negative based on the sentiment words expressed in documents which is necessary to be developed for the manufacturer and the customer in order to analyze the reviews of the customers. The goal of sentiment classification is to discover customer opinion on a product. Sentiment classification has been applied in various tasks such as opinion mining [11], market analysis [14], opinion summarization [12] and contextual advertising [13].

Developing general sentiment classifier is complex because in all domains the sentiment words have different connotation. Sentiment classification problem is more challenging because sentiments can be expressed in a more subtle manner. For example, the sentence, "How could anyone sit through this movie!" [5] contains not even a single sentiment words to express the sentiment but it is obviously negative. These kinds of sentences are available plenty while conveying the reviews of the product purchased.

Existing machine learning approaches [5][9] depends on labeled training data. However, such labeled data are not always available in practical applications and it is well known that sentiment classifier trained in one domain may not produce satisfactory results when it is used in another domain, since sentiments are expressed differently in different domain. Table 1 shows user review sentences from two domains books and electronics. In the books domain, the words "*thrilling*", "*well researched*". and "*interesting"* are used to express positive sentiments and "*disappointed"* is used to express negative sentiments. While in electronics domain "*fast*", *reliable, compact* and "*sharp"* are used to express the positive sentiment, *"blurry"* and "*never recommend*" are negative sentiments. Due to the features mismatch between domain specific words of two domains, classifier trained in one domain may not work well when directly applied to other domain.

In literature, Sinno Jialin Pan et al.[1] proposed spectral feature alignment algorithms to solve feature mismatch problem by aligning domain specific words from different domains into unified cluster

with the help of domain independent words and then unified cluster is used to train a classifier in target domain.

TABLE I
Cross Domain Sentiment Classification Examples: Reviews of Books and Electronics

Label	Books source domain	Electronics target domain
+	This book is excellent and **well researched**	Sandisk products are **fast** and **reliable**
+	This is an **interesting** and **thrilling** story book	**Compact**, easy to operate, looks **sharp**
-	When I read this book, I really **disappointed**	It is **blurry** in dark settings, I would **never recommend** this product to anyone

Danushka Bollegale et al.[3] proposed cross domain sentiment classification by creating sentiment sensitive thesaurus which aligns different words that express the same sentiments. They expanded feature vector using created sentiment sensitive thesaurus while training a binary classifier.

In this paper, the proposed method creates an enhanced sentiment sensitive thesaurus (ESST) which aligns semantically similar features from various domains as well as sentiment features from wiktionary. The idea behind this approach is, it is well known that adjectives are used to express sentiment in all domains, so, it is necessary to collect more adjectives from any lexical knowledge base to solve the feature mismatch problem.

To create an enhanced sentiment sensitive thesaurus, first, domain independent features and domain specific features are collected from all domain reviews. Second, co-occurrence matrix is computed between each domain independent features with domain specific features. Third, features are weighted using point wise mutual information. Finally, semantically similar domain specific features from various domains are aligned with the help of domain independent features by computing similarity measure between two domain specific features based on the PMI weighted ratio. At the same time semantically similar sentiment features for each adjective of the given reviews are collected from wiktionary with the help of java wiktionary library and these sentiment features are added to the already created thesaurus.

In rest of the paper is organized as follows: Section II describes related work. Section III describes definitions and Methodology used in proposed work .Section IV describes experimental setup and solution. Section V concludes the proposed work.

II. RELATED WORK

Generally two types of approaches have been used for sentiment classification. First approach is machine learning and second approach is semantic orientation. In machine learning [5] approach classifier is trained using feature vectors and it produce more accurate results. In this approach, documents are represented as feature vector and it is trained by various classification algorithms such as Naive Bayes, Maximum entropy and SVM[9]. The Second approach is semantic orientation [4][10], it does not require prior training, instead it measures semantic orientation of sentences used in documents. Two types of semantic orientation approaches, corpus based and dictionary based techniques were used in existing research work. Corpus based techniques aim to find co-occurrence patterns of the words to determine the sentiments. Dictionary based methods utilize synonyms, antonyms and hierarchies in wordnet or sentiwordnet to determine the sentiments.

In existing methods, reviews are classified at various levels i.e. document level, sentence level and word level. The proposed work is based on document level which classifies the documents as positive or negative, based on overall sentiment words expressed in the documents. Turney [4] used supervised learning techniques with mutual information to predict overall document sentiment by averaging out the semantic orientation of phrases in document. Turney [5] used Naive Bayes, Maximum Entropy and SVM for classifying movie reviews and received best results using SVM. Kennedy and Inkpen [6] used Contextual valence shifter to predict sentiment of the sentences. Chenghua Lin et al.[7] proposed a novel probabilistic modeling framework called joint sentiment topic model which detects sentiments and topic simultaneously from text.

In sentiment classification, many researchers have been concentrated on online lexical resources such as sentiwordnet, wiktionary and Wikipedia which are publically available for research purpose. SentiWordNet[19] is a lexical resource, containing opinion information on terms extracted from the wordnet database where each term is associated with numerical scores indicating positive and negative information. Chien-Liang Liu et al.[17] proposed movie rating and review summarization in mobile environment using Latent semantic analysis. Aurangzeb khan et al.[16] proposed sentiment classification by sentence level semantic orientation using sentiwordnet[18] which classify the reviews into objective or subjective sentences. The semantic score

of the subjective sentences are extracted from sentiwordnet to calculate their polarity as positive or negative.

Blitzer et al.[2] addressed the problem in cross domain sentiment classification using structural correspondence learning algorithm where frequent words in both source and target domain were selected as candidate pivot features and linear predictors are trained to predict the occurrences of those pivot features. In structural corresponding learning-mutual information approach, the mutual information between a feature and the domain label is used to select pivot features instead of co-occurrence frequency. Sinno Jialin Pan et al.[1] proposed a spectral feature alignment algorithm to align domain specific words from different domain into unified clusters with the help of domain independent words as a bridge and bipartite graph is constructed between domain specific and domain independent features and then the clusters can be used to train a sentiment classifier in target domain. Lun yan et al.[8] used self growth algorithm to generate a cross domain sentiment word list which is used in sentiment classification of web news. Danushka Bollegale et al.[3] proposed cross domain sentiment classification by creating sentiment sensitive thesaurus which aligns different words that express the same sentiments. They expanded feature vector using created sentiment sensitive thesaurus while training a binary classifier.

III. METHODOLOGY

A. Definitions

In this section, several definitions are given to clarify basic terminology.

Domain- A domain D denotes a class of entities in the real world. For example, different types of product such as DVD, Kitchen appliances, books and electronics.

Sentiment- Given a specific domain D, sentiment data are the text documents which express the user opinion about the entities of the domain. i.e. positive or negative opinion about any product.

Cross Domain sentiment Classification- Given a set of labeled reviews Ds= $\{(x_i, y_i)\}$ from source domain where x_i represents features and y_i represent sentiment label $y_i \in \{+1, -1\}$. To predict the label in unlabeled target domain Dt=$\{x_j\}$ where x_j represent features in target domain. Classifier is trained by labeled reviews of source domain and it is applied to classify the reviews of unlabeled target domain.

B. Enhanced Sentiment Sensitive Thesaurus using Wiktionary

In cross domain sentiment classification, classifier trained from one domain (source) is applied to another (target) domain. Obviously it produces poor performance because trained features are mismatched with target domain features. This feature mismatch problem in cross domain sentiment classification is solved by creating enhanced sentiment sensitive thesaurus which collect more semantically similar features from all reviews of source and target domain as well as from Wiktionary. Our contribution is extracting more sentiment features from wiktionary with the help of java wikitionary library tool (JWKTL) and these features are appended to enhanced sentiment sensitive thesaurus (ESST) to improve the classification performance in target domain.

Wiktionary is online dictionary which has glosses and synset for each word. That is, adjectives in this resource are clearly defined with list of synonyms. Unlike Wikipedia, it focuses on lexical instead of encyclopedic knowledge. Wiktionary contains many types of information also found in linguistic knowledge base (LKB) i.e. wordnet, like definitions, synonyms, and hyponyms, and also additional types of information,e.g. abbreviations, compounds or contractions, which are usually not found in LKBs. Another difference to LKBs is that each language-specific edition of Wiktionary contains not only entries for words in that particular language, but also for words in other languages. Wiktionary has about 3.7 million word entries in 171 language editions.

The Enhanced sentiment sensitive Thesaurus is constructed using the knowledge from the Wiktionary and labeled/unlabled data from source domain and unlabeled data from target domain. To interact with wiktionary, JWKTL(java wiktionary library) is needed which is released by Ubiquitous Knowledge Processing Lab.

Given a labeled or unlabeled review, first split the reviews into set of sentences, and then perform parts of speech (POS) tagging and Lemmatization is performed using RASP[15] system. Lemmatization is used to reduce the features by converting singular and plurals into base form. By using a simple word filter based on the POS tagging to filter out unwanted words retaining nouns, verbs, adjectives, adverbs alone. Table 2 shows the example of unigrams, bigrams and sentiment elements extracted from the reviews. First, model the review as bag of words and then extract unigrams, bigrams from each sentence. Bigrams are necessary in sentiment classification, since semantic orientations of sentences are identified by bigrams. Next, from each source domain labeled reviews, sentiment features are created by appending the label of the review to each feature. The notation *p to indicate positive features and *N to indicate negative features.

To construct ESST, first Domain independent features are extracted from all domains which occur frequently in all domains. Domain independent features are useful to bridge the gap between source and target domains. Domain independent features are extracted from each domain by computing mutual information between domain and features. If the feature has high mutual information to the domain which is considered as domain specific feature. If the feature has less mutual information to the domain which is considered as domain independent features. Later on the co-occurrence matrix is computed between domain Independent features and domain specific features. Semantic meanings of the sentences are found based on co-occurrences of the terms used in the documents. After the computation of co-occurrence between domain independent features and domain specific features, value of the features in the co-occurrence matrix are weighted using point wise mutual information equation 1. Fig 1 shows the construction of cross domain sentiment classifier using enhanced sentiment sensitive Thesaurus

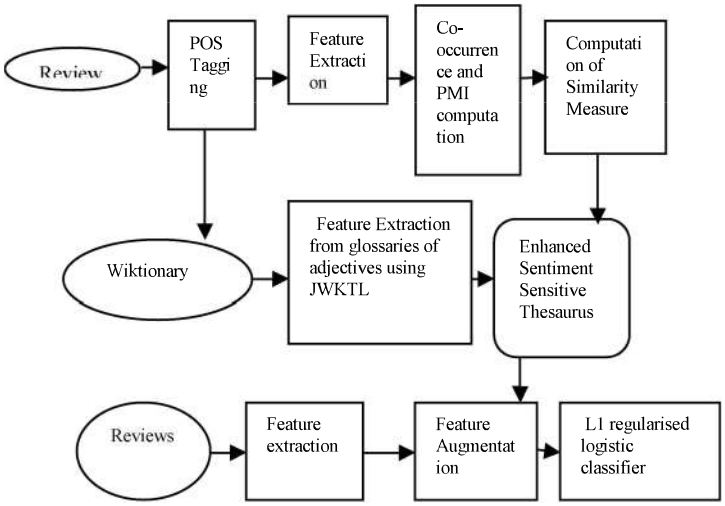

Fig .1.Construction of Cross domain sentiment classifier using ESST

TABLE II
EXAMPLE OF UNIGRAM ,BIGRAM AND SENTIMENT FEATURES FROM POSITIVE REVIEW SENTENCE

Sentence	An excellent workbook full of delicious recipes
POS tags	An_AT1 excellent_JJ workbook_NN1 full_JJ of_IO delicious_JJ recipes_NN2
unigrams	Excellent, workbook, full, delicious, recipes
bigrams	Excellent+workbook, workbook full, full+delicious, delicious+recipes
Sentiment elements	Excellent*p, workbook*p, full*pl, delicious*p, recipes*p Excellent+workbook*p,workbook+full*p,full+delicious*p,delicious+recipes*p

For each domain independent feature s, domain specific features v that co-occurs with domain independent feature contributes feature vector **s.** The value of the feature v in vector **s** is denoted by M (**s,**v).

$$M(\boldsymbol{s},v) = log\left(\frac{\frac{C(s,v)}{N}}{\frac{\sum_{i=1}^{n} C(i,v)}{N} * \frac{\sum_{j=1}^{m} C(s,j)}{N}}\right) \quad (1)$$

M (*s,v*) is the point wise mutual information between a domain independent feature *s* and domain specific feature *v*. In equation 1, C(s,v) represents number of review sentences in which both domain independent feature s and domain specific feature v co-occur, C(s,j) represents number of times domain independent feature s occur in review sentences. C(i,v) represents number of times domain specific feature v occur in the review sentences. n and m denotes the total no of domain independent features and domain specific features and $N = \sum_{i=1}^{n}\sum_{j=1}^{m} C(i,j)$.

After the computation of PMI values, domain specific features from various domains are aligned with the help of DI features. Semantically similar domain specific features from various domains are aligned by finding similarity score between each domain specific feature with every other domain specific features of PMI weighted matrix and the same procedure is followed to align domain independent features. For example, similarity score between two domain independent features s, t (both s and t have feature vectors **s**, **t**) is computed by following formula

$$T(s,t) = \frac{\sum_{v \in \{x \mid M(s,x)>0\}} M(t,v)}{\sum_{v \in \{x \mid M(t,x)>0\}} M(t,v)} \quad (2)$$

The similarity score T(s, t) is the proportion of PMI weighted features of the feature t that are shared with feature s**.** The distributional hypothesis states that words that have similar distributions are semantically similar[19]. i.e. Two words are semantically similar if two words occur with same distributional contexts.

Enhanced Sentiment sensitive thesaurus aligns semantically semantically similar domain specific features from various domains by computing similarity measure equation (2) in PMI weighted matrix. For every domain specific feature s, ESST list up many domain specific features based on descending order of the similarity score.

Moreover, ESST collect more semantically similar features from wiktionary using java wiktionary library (JWKTL) which is very useful tool to extract more semantic relatedness features

from Wiktionary. To enhance sentiment sensitive thesaurus, first, seed adjectives are extracted from the given reviews.. Next, the extracted adjectives are given as input to wiktionary using JWKTL. The corresponding glossaries of each adjective are extracted from the wiktionary. Finally, unigram and bigrams from the extracted glossaries are collected for each adjective and then these are appended to the ESST.

C. Augmentation of Features

After creating ESST thesaurus, then simply augment domain specific features of various domains with the original features of source domain by finding suitable domain specific features from the created ESST Thesaurus which creates a new feature vector representation for cross-domain sentiment classification. This new representation of feature vector is used to train a sentiment classifier to predict the label of target domain. The new feature vector representation of each review is defined as

$$y_i = [x_i, (DS(x_i)]$$

Where *xi* is the original features of source domain, *DS (xi)* is the augmented features from ESST, *yi is* the new representation of feature vector which is generated while training a classifier.

Algorithm 1 Construction of cross domain sentiment classifier using enhanced sentiment sensitive Thesaurus

Input: labeled source Domain data D_{sr}= { X_{sr},Y_{sr}} and unlabeled source domain D_{sr}={X_i} and unlabeled target Domain D_{tr} = {X_{tr}}and Wiktionary dump file

Output: predict the label of target domain.

1. Extract Domain Independent features and domain specific features from the given Reviews.
2. Create co-occurrence matrix between domain independent features with domain specific features.
3. Compute Point Wise Mutual Information for each features using equation (1).
4. Create Enhanced sentiment thesaurus by aligning domain specific features by computing similarity measure using equation 2 between domain specific features based on PMI Weighted ratio. Similarly align domain independent features by computing similarity measure between DI features.
5. Glossaries of each adjectives are extracted from wiktionary using java wiktionary library (JWKTL) by giving seed adjectives from reviews. Unigram and bigram are generated from glossaries which are appended to the ESST.
6. Find a new representation of feature vector by Feature augmentation while training a classifier.
7. Test the classifier in target domain.

IV. Experiments

A. Data Sets

Amazon product reviews are bench mark data set which consist of four different product types: Books, DVDs, electronics and kitchen appliances are chosen for the proposed work. Each review is assigned with rating (0-5 stars).Review with rating >3 are labeled as positive and review with rating<3 are labeled as negative. The data sets structure is shown in table3.For each domain, there are 1000 positive reviews with 1000 negative reviews. Each domain also has some unlabeled reviews.

TABLE III
AMAZON PRODUCT DATA SETS

Domain	Positive	Negative	Unlabeled
Kitchen	1000	1000	16746
DVDs	1000	1000	34377
Electronics	1000	1000	13116
Books	1000	1000	5947

For experiments, among four domains, one domain is taken as source domain and another domain is selected as target domain. To create enhanced sentiment sensitive thesaurus, 800 positive reviews and 800 negative reviews are selected from Source domain and some unlabeled reviews are selected from target domain. The enhanced sentiment sensitive thesaurus is automatically created by giving labeled/unlabeled reviews of source domain and unlabeled reviews of target domain .After creating ESST thesaurus, feature augmentation is performed to build a cross domain sentiment classifier.

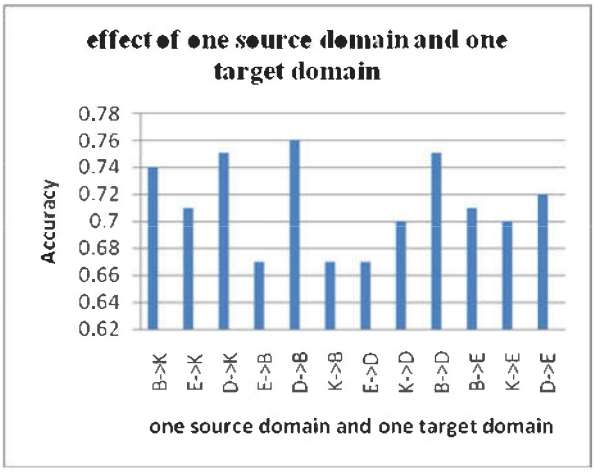

Fig. 2. Classification accuracy of one source and one target domain using ESST

Fig 2 shows the effect of ESST thesaurus for one source domain / one target domain which shows the classification accuracy of 12 different combinations of one source domain and one target domain using created enhanced sentiment sensitive thesaurus.

To study the effect of ESST thesaurus for one source domain and one target domain, 12 different combinations of one source domain and one target domain is analyzed. From the analysis, D→ B , DVD domain is selected as source domain and books domain is selected as target domain which produce high accuracy 0.76 percentage. Next D→ K and B→.D produce 0.75 percentage accuracy. Here B denotes books domain and K denotes kitchen domain. While training 800 positive and 800 negative reviews are selected to train sentiment classifier and remaining 200 reviews are used for testing. This enhanced sentiment sensitive thesaurus highly effective to classify the reviews of unseen target domain features because it collect more semantically similar elements from wiktionary.

V. CONCLUSION

The proposed work performs a cross domain sentiment classification by creating enhanced sentiment sensitive thesaurus. The Enhanced sentiment sensitive thesaurus is created automatically based on the reviews of source domain and target domain. Enhanced sentiment sensitive thesaurus aligns more semantically similar sentiment features from source and target domain as well as additional sentiment features from wiktionary. Next, feature vector augmentation is performed during training of the reviews using ESST.L1 regularized logistic classification algorithm is used to train classifier which produces sparse output and it is useful to predict label of unseen features of target domain.

REFERENCES

[1]. S. J. Pan, X. Ni, J.-T. Sun, Q. Yang, and Z. Chen, "Cross-domain Sentiment classification via spectral feature alignment," in Carolina, USA,ACM 978-1-60558-799-8/10/04.

[2].J. Blitzer, R. McDonald, and F. Pereira, "Domain adaptation with structural correspondence learning," in *EMNLP 2006*, 2006.

[3].Danushka Bollegala, David Weir and John carroll, "Cross Domain sentiment classification using a sentiment sensitive thesaurus," in IEEE Transactions on knowledge and data engineering, vol. 25, no. 8, August 2013.

[4]. P. D. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews," in ACL'02 proceedings of the 40th annual meeting on association for Computational linguistics, pages 417-424.

[5].B.Pang .L. Lee, and S. Vaithyanathan, "Thumbs up?Sentiment classification using machine learning techniques," proceedings of the conference on empirical methods in Natural language processing (*EMNLP), Philadelphia, July 2002* pp. 79–86.

[6].A.Kennedy and D.Inkpen, "Sentiment classification of movie reviews using contextual valence shifters," computational Intelligence, vol 22, pp.110-125, 2006.

[7].Chenghua Lin, Yulan he, Richard Everson and Stefan Ruger, " Weakly supervised joint sentiment- topic detection from text," vol 24, Journal of latex class files, January 2011.

[8]. Lun Yan and Yan Zhang, "News sentiment Analysis based on cross domain sentiment word lists and content classifiers," ADMA 2012,LNAI 7713,pp 577-588,springer-verlog Berlin Heidelberg 2012.

[9]. T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in Proceedings 10"' European Conference on Machine Learning (ECML), Springer Verlag, 1998, pp.137–142.

[10]. V. Hatzivassiloglou and K. R. McKeown, "Predicting the semantic orientation of adjectives," in *ACL 1997*, 1997, pp. 174–181.

[11] B. Pang and L. Lee, "Opinion mining and sentiment analysis, "*Foundations and Trends in Information Retrieval*, vol. 2, no. 1-2, pp.1–135, 2008.

[12] Y. Lu, C. Zhai, and N. Sundaresan, "Rated aspect summarization of short comments," in the proceedings of the 18th conference on world wide web, ACM, New Yark pp.131-140

[13] T.-K. Fan and C.H. Chang, "Sentiment-oriented contextual advertising, "*Knowledge and Information Systems*, vol. 23, no. 3, pp.321–344, 2010.

[14] M. Hu and B. Liu, "Mining and summarizing customer reviews, " in proceeding of the tenth ACM SIG KDD international conference on knowledge discovery and data mining , New Yark, USA 2004, pp.168-177.

[15]. T. Briscoe, J. Carroll, and R. Watson, "The second release of the RASP system," in *COLING/ACL 2006 Interactive Presentation Sessions*, 2006.

[16]. Aurangzeb khan,Baharum Baharudin, " Sentiment classification by sentence level semantic orientation using sentiwordnet from online reviews and Blogs," in Int.J Comp Sci.Emerging Tech,vol 2, no.4 august 2011.

[17]. Chien-Liang Liu, Wen- hoar Hsaio,Chia-hoang lee,Gen-chi and Emery jou, " Movie Rating and review summarization in Mobile Environment" in IEEE 2012, vol 42.pp 397-407.

[18] A. Esuli and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining," in *LREC 2006*, 2006, pp. 417–422.

[19] Z. Harris, "Distributional structure," *Word*, vol. 0, pp. 146–162, 1954.