ISSN(Online):2320-9801
ISSN(Print): 2320- 9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 5, May 2014

Cross-Domain Opinion Mining Using a Thesaurus in Social Media Content

N.Manjunathan,

Assistant Professor, Department of Computer Science and Engineering, Asan Memorial College of Engineering and Technology, Chengalpattu, Chennai, India.

ABSTRACT—Automatic classification of sentiment is important for numerous applications such as opinion mining, opinion summarization, contextual advertising, and market analysis. Typically, sentiment classification has been modeled as the problem of training a binary classifier using reviews annotated for positive or negative sentiment. However, sentiment is expressed differently in different domains, and annotating corpora for every possible domain of interest is costly. Applying a sentiment classifier trained using labeled data for a particular domain to classify sentiment of user reviews on a different domain often results in poor performance because words that occur in the train (source) domain might not appear in the test (target) domain. We propose a method to overcome this problem in cross-domain sentiment classification. First, we create a sentiment sensitive distributional thesaurus using labeled data for the source domains and unlabeled data for both source and target domains. Sentiment sensitivity is achieved in the thesaurus by incorporating document level sentiment labels in the context vectors used as the basis for measuring the distributional similarity between words. Next, we use the created thesaurus to expand feature vectors during train and test times in a binary classifier. The proposed method significantly outperforms numerous baselines and returns results that are comparable with previously proposed cross-domain sentiment classification methods on a benchmark data set containing Amazon user reviews for different types of products. We conduct an extensive empirical analysis of the proposed method on single- and multisource domain adaptation, unsupervised and supervised domain adaptation, and numerous similarity measures for creating the sentiment sensitive thesaurus. Moreover, our comparisons against the SentiWordNet, a lexical resource for word polarity, show that the created sentiment-sensitive thesaurus accurately captures words that express similar sentiments.

INDEX TERMS—Cross-domain sentiment classification, domain adaptation, thesauri creation

I. INTRODUCTION

User express their opinions about products or services they consume in blog posts, shopping sites, or review sites. Reviews on a wide variety of commodities are available on the Web such as, books (amazon.com), hotels (tripadvisor.com), movies (imdb.com), automobiles (caranddriver.com), and restaurants (yelp.com). It is useful for both the consumers as well as for the producers to know what general public think about a particular product or service. Automatic document level sentiment classification [1], [2] is the task of classifying a given review with respect to the sentiment expressed by the author of the review. For example, a sentiment classifier might classify a user review about a movie as positive or negative depending on the sentiment expressed in the review. Sentiment classification has been applied in numerous tasks such as opinion mining [3], opinion summarization [4], contextual advertising [5], and market analysis [6]. or negative sentiments and then create a summary for each sentiment type for a particular product. A contextual advert placement system might decide to display an advert for a particular product if a positive sentiment is expressed

ISSN(Online):2320-9801
ISSN(Print): 2320- 9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 5, May 2014

in a blog post.

Supervised learning algorithms that require labeled data have been successfully used to build sentiment classifiers for a given domain [1]. However, sentiment is expressed differently in different domains, and it is costly to annotate data for each new domain in which we would like to apply a sentiment classifier. For example, in the electronics domain the words “durable” and “light” are used to express positive sentiment, whereas “expensive” and “short battery life” often indicate negative sentiment. On the other hand, if we consider the books domain the words “exciting” and “thriller” express positive sentiment, whereas the words “boring” and “lengthy” usually express negative sentiment. A classifier trained on one domain might not perform well on a different domain because it fails to learn the sentiment of the unseen words.

The cross-domain sentiment classification problem [7], [8] focuses on the challenge of training a classifier from one or more domains (source domains) and applying the trained classifier on a different domain (target domain). A cross-domain sentiment classification system must overcome two main challenges. First, we must identify which source domain features are related to which target domain features. Second, we require a learning framework to incorporate the information regarding the relatedness of source and target domain features. In this paper, we propose a cross-domain sentiment classification method that overcomes both those challenges.

We model the cross-domain sentiment classification problem as one of feature expansion, where we append additional related features to feature vectors that represent source and target domain reviews to reduce the mismatch of features between the two domains. Methods that use related features have been successfully used in numerous tasks such as query expansion [9] in information retrieval [10], and document classification [11]. For example, in query expansion, a user query containing the word car might be expanded to car OR automobile, thereby retrieving documents that contain either the term car or the term automobile. However, to the best of our knowledge, feature expansion techniques have not previously been applied to the task of cross-domain sentiment classification.The proposed method can learn from a large amount of unlabeled data to leverage a robust cross-domain sentiment classifier.

In our proposed method, we use the automatically created thesaurus to expand feature vectors in a binary classifier at train and test times by introducing related lexical elements from the thesaurus. We use L1 regularized logistic regression as the classification algorithm. However, the proposed method is agnostic to the properties of the classifier and can be used to expand feature vectors for any binary classifier. As shown later in the experiments, L1 regularization enables us to select a small subset of features for the classifier.

Our contributions in this work can be summarized as follows:

. We propose a fully automatic method to create a thesaurus that is sensitive to the sentiment of words expressed in different domains. We utilize both labeled and unlabeled data available for the source domains and unlabeled data from the target domain.

. We propose a method to use the created thesaurus to expand feature vectors at train and test times in a binary classifier.
. We compare the sentiment classification accuracy of our proposed method against numerous baselines and previously proposed cross-domain sentiment classification methods for both single.

II. PROBLEM SETTING

We define a domain D as a class of entities in the world or a semantic concept. For example, different types of products such as books, DVDs, or automobiles are considered as different domains. Given a review written by a user on a product that belongs to a particular domain, the objective is to predict the sentiment expressed by

ISSN(Online):2320-9801
ISSN(Print): 2320- 9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 5, May 2014

the author in the review about the product. We limit ourselves to binary sentiment classification of entire reviews.

We denote a source domain by D_{src} and a target domain by D_{tar}. The set of labeled instances from the source domain, LðD$_{src}$Þ, contains pairs ð**t**; cÞ where a review, **t**, is assigned a sentiment label, c. Here, c 2 f1; _1g, and the sentiment labels þ1 and _1, respectively, denote positive and negative sentiments. In addition to positive and negative sentiment reviews, there can also be neutral and mixed reviews in practical applications. If a review discusses both positive and negative aspects of a particular product, then such a review is considered as a mixed sentiment review. On the other hand, if a review does not contain neither positive nor negative sentiment regarding a parti-cular product then it is considered as neutral. Although this paper only focuses on positive and negative sentiment reviews, it is not hard to extend the proposed method to address multicategory sentiment classification problems.

III. SENTIMENT SENSITIVE THESAURUS

As we saw in our example in Section 3, a fundamental problem when applying a sentiment classifier trained on a particular domain to classify reviews on a different domain is that words (hence features) that appear in the reviews in the target domain do not always appear in the trained model. To overcome this feature mismatch problem, we construct a sentiment sensitive thesaurus that captures the relatedness of words as used in different domains. Next, we describe the procedure to construct our sentiment sensitive thesaurus.

Given a labeled or an unlabeled review, we first split the review into individual sentences and conduct part-of-speech (POS) tagging and lemmatization using the RASP system [12]. Lemmatization is the process of normalizing the inflected forms of a word to its lemma. Lemmatization reduces the feature sparseness and has shown to be effective in text classifica-tion tasks [13].We then apply a simple word filter based on POS tags to filter out function words, retaining only nouns, verbs, adjectives, and adverbs. In particular, adjectives have been identified as good indicators of sentiment in previous work [14], [15]. Following the previous work in cross-domain sentiment classification, we model a review as a bag of words. We then select unigrams and bigrams from each sentence. For the remainder of this paper, we will refer both unigrams and bigrams collectively as lexical elements. In previous work on sentiment classification it has been shown that the use of both unigrams and bigrams are useful to train a sentiment classifier [7]. We note that it is possible to create lexical elements from both source domain labeled reviews (LðD$_{src}$Þ) as well as unlabeled reviews from source and target domains (UðD$_{src}$Þ and UðD$_{tar}$Þ).

Next, from each source domain labeled review we create sentiment elements by appending the label of the review to each lexical element we generate from that review. For example, consider the sentence selected from a positive review on a book shown in Table 2. In Table 2, we use the notation "*P" to indicate positive sentiment elements and "*N" to indicate negative sentiment elements. The example sentence shown in Table 2 is selected from a positively labeled review, and generates positive sentiment elements as show in Table 2. Sentiment elements, extracted only using labeled reviews in the source domain, encode the sentiment information for lexical elements extracted from source and target domains.

We represent a lexical or sentiment element u by a feature vector **u**, where each lexical or sentiment element w that co-occurs with u in a review sentence contributes a feature to **u**. Moreover, the value of the feature w in vector **u** is denoted by fð**u**; wÞ. The vector **u** can be seen as a compact representation of the distribution of an element u over the set of elements that cooccur with u in the reviews. The distributional hypothesis states that words that have similar distributions are semantically similar [16].Fig. 1. Constructing feature vectors for two lexical elements u_1 and u_2 from a positive labeled source domain review L_{src}, two unlabeled reviews from source (U_{src}) and target (U_{tar}) domains. Vector $\mathbf{u}_1$ contains the sentiment element v_1_P and the lexical elements v_1, v_2. Vector $\mathbf{u}_2$ contains lexical elements v_1 and v_2. The relatedness, _ðu_1; u_2Þ, between u_1 and u_2 is given by (2).

ISSN(Online):2320-9801
ISSN(Print): 2320- 9798

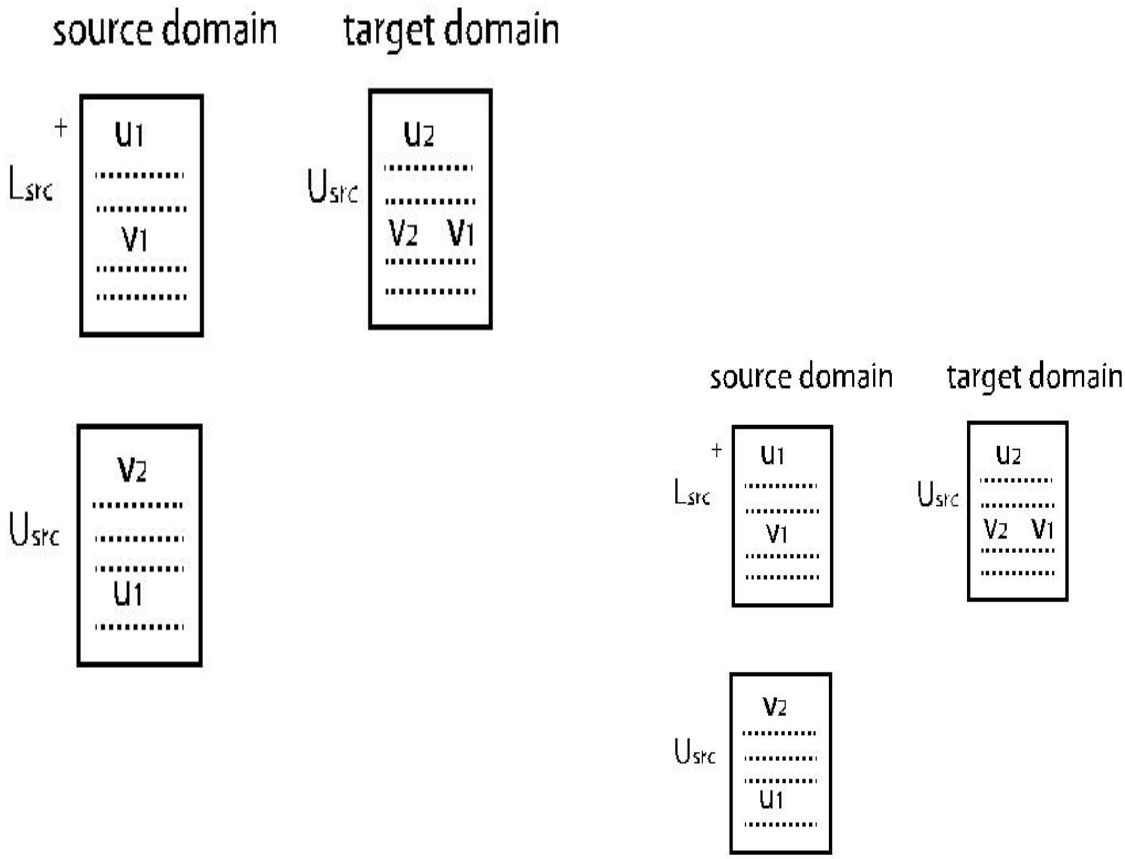
International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 5, May 2014

candidates for sentiment elements. However, we emphasize the fact that the relatedness values between the lexical elements listed in the sentiment-sensitive thesaurus are computed using co-occurrences with both lexical and sentiment elements, and, therefore, the expansion candi-dates selected for the lexical elements in the target domain reviews are sensitive to sentiment labels assigned to reviews in the source domain.

To construct the sentiment sensitive thesaurus, we must compute pairwise relatedness values using (2) for numer-ous lexical elements. Moreover, to compute the pointwise mutual information values in feature vectors, we must store the cooccurrence information between numerous lexical and sentiment elements. By using a sparse matrix format and approximate vector similarity computation techniques [21], we can efficiently create a thesaurus from a large set of reviews. In particular, by using approximate vector simi-larity computation techniques we can avoid computing relatedness values between lexical elements that are likely to have very small relatedness scores thus are unlikely to become neighbors of a given base entry.



ISSN(Online):2320-9801
ISSN(Print): 2320- 9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 5, May 2014

sentence	Excellent and broad survey of the development of civilization.
POS tags	Excellent/JJ and/CC broad/JJ survey/NN1 of/IO the/AT development/NN1 of/IO civilization/NN1
lexical elements (unigrams)	excellent, broad, survey, development, civilization
lexical elements (bigrams)	excellent+broad, broad+survey, survey+development, development+civilization
sentiment elements	excellent*P, broad*P, survey*P, development*P, civilization*P, excellent+broad*P, broad+survey*P, survey+development*P, development+civilization*P

TABLE 2
Generating Lexical and Sentiment Elements from a Positive Review Sentence

IV. FEATURE EXPANSION

A fundamental problem in cross-domain sentiment classi-fication is that features that appear in the source domains do not always appear in the target domain. Therefore, even if we train a classifier using labeled data from the source domains, the trained model cannot be readily used to classify test instances in the target domain. To overcome this problem, we propose a feature expansion method where we augment a feature vector with additional related features selected from the sentiment-sensitive thesaurus created in Section 4. In this section, we describe our feature expansion method.

First, following the bag-of-words model, we model a review d using the set $fw_1; \ldots; w_N g$, where the elements w_i are either unigrams or bigrams that appear in the review d. We then represent a review d by a real-valued term-frequency vector **d** 2 IR^N , where the value of the jth element d_j is set to the total

The values of the first N dimensions that correspond to
unigrams and bigrams w_i that occur in the review d are set to d_i, their frequency in d. The subsequent k dimensions that correspond to the top ranked base entries for the review d, are weighted according to their ranking score. Specifically, we set the value of the rth ranked base entry v^r_d to 1=r. Alternatively, one could use the ranking score, $score\eth v^r_d; d\Þ$, itself as the value of the appended base entries. However, both relatedness scores as well as normalized term-frequencies can be small in practice, which leads to very small absolute ranking scores. On the other hand, the expanded features must have lower feature values compared to that of the original features in particular feature vector. We have set the feature values for the original features to their frequency in a review. Because Amazon product reviews are short, most features occur only once in a review. By using the inverse rank as the feature value for expanded features, we only take into account the relative ranking of base entries and at the same time assign feature values lower than that for the original features.

Note that the score of a base entry depends on a review d. Therefore, we select different base entries as additional features for expanding different reviews. Furthermore, we do not expand each w_i individually when expanding a vector **d** for a review. Instead, we consider all unigrams and bigrams in d when selecting the base entries for expansion. One can visualize the feature expansion process as a lower dimen-sional latent mapping of features onto the space spanned by the base entries in the sentiment-sensitive thesaurus. By adjusting the value of k, the number of base entries used for expanding a review, one can change the size of this latent

ISSN(Online):2320-9801
ISSN(Print): 2320- 9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 5, May 2014

Number of Reviews in the Benchmark Data Set

Domain	positive	negative	unlabeled
kitchen	1000	1000	16746
DVDs	1000	1000	34377
electronics	1000	1000	13116
books	1000	1000	5947

TABLE 3

Using the extended vectors $\mathbf{d}^0$ to represent reviews, we train a binary classifier from the source domain labeled reviews to predict positive and negative sentiment in reviews. We differentiate the appended base entries v^r_d from w_i that existed in the original vector **d** (prior to expansion) by assigning different feature identifiers to the appended base entries. For example, a unigram excellent in a feature vector is differentiated from the base entry excellent by assigning the feature id, "BASE ¼ excellent" to the latter. This enables us to learn different weights for base entries depending on whether they are useful for expanding a feature vector. Once a binary classifier is trained, we can use it to predict the sentiment of a target domain review. We use the above-mentioned feature expansion method coupled with the sentiment-sensitive thesaurus to expand feature vectors at test time for the target domain as well.

V. EXPERIMENTS

5.1 Data Set

We use the cross-domain sentiment classification data set[1] prepared by Blitzer et al. [7] to compare the proposed method against previous work on cross-domain sentiment classification. This data set consists of Amazon product reviews for four different product types: books, DVDs, electronics, and kitchen appliances. Each review is assigned with a rating (0-5 stars), a reviewer name and location, a product name, a review title and date, and the review text. Reviews with rating >3 are labeled as positive, whereas those with rating <3 are labeled as negative. The overall structure of this benchmark data set is shown in Table 3. For each domain, there are 1,000 positive and 1,000 negative examples, the same balanced composition as the polarity data set constructed by Pang et al. [1]. The data set also contains some unlabeled reviews for the four domains. This benchmark data set has been used in much previous work on cross-domain sentiment classification and by evaluating on it we can directly compare the proposed method against existing approaches.

Following previous work, we randomly select 800 positive and 800 negative labeled reviews from each domain as training instances (total number of training instances are 1;600 _ 4 ¼ 6;400), and the remainder is used for testing (total number of test instances are 400 _ 4 ¼ 1;600). In our experiments, we select each domain in turn as the target domain, with one or more other domains as sources. Note that when we combine more than one source

We use classification accuracy on target domain as the evaluation metric. It is the fraction of the correctly classified target domain reviews from the total number of reviews in the target domain, and is defined as follows:

The above-mentioned procedure creates four thesauri (each thesaurus is created by excluding labeled training data for a particular target domain). For example, from the three domains DVDs, electronics, and books, we generate 53,586 lexical elements and 62,744 sentiment elements to create a thesaurus that is used to adapt a classifier trained on those three domains to the kitchen domain. Similar numbers of features are generated for the other domains as well. To avoid generating sparse and probably noisy features, we require that each feature occur in at least two different review sentences.

ISSN(Online):2320-9801
ISSN(Print): 2320- 9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 5, May 2014

5.2 Cross-Domain Sentiment Classification

To evaluate the benefit of using a sentiment sensitive thesaurus for cross-domain sentiment classification, we compare the proposed method against three baseline methods in Table 4. Next, we describe the methods compared in Table 4.

. No adapt. This baseline simulates the effect of not performing any feature expansion. We simply train a binary classifier using unigrams and bigrams as features from the labeled reviews in the source domains and apply the trained classifier on a target

. Nonsentiment sensitive thesaurus (NSST). To evaluate the benefit of using sentiment features on our proposed method, we create a thesaurus only using lexical elements. Lexical elements can be derived from both labeled and unlabeled reviews whereas, sentiment elements can be derived only from labeled reviews. We did not use rating information in the source domain labeled data in this baseline. A thesaurus is created using those features and subsequently used for feature expansion. A binary classifier is trained using the expanded features.

. Proposed (SST: sentiment sensitive thesaurus). This is the proposed method described in this paper. We use the sentiment sensitive thesaurus created using the procedure described in Section 4 and use the thesaurus for feature expansion in a binary classifier.

. In-domain. In this method, we train a binary classifier using the labeled data from the target domain. This method provides an upper bound for the cross-domain sentiment analysis. This upper baseline demonstrates the classification accuracy we can hope to obtain if we had labeled data for the target domain. Note that this is not a cross-domain classification setting.

5.3 Effect of Relatedness Measures

The choice of the relatedness measure is an important decision in a thesauri-based approach.

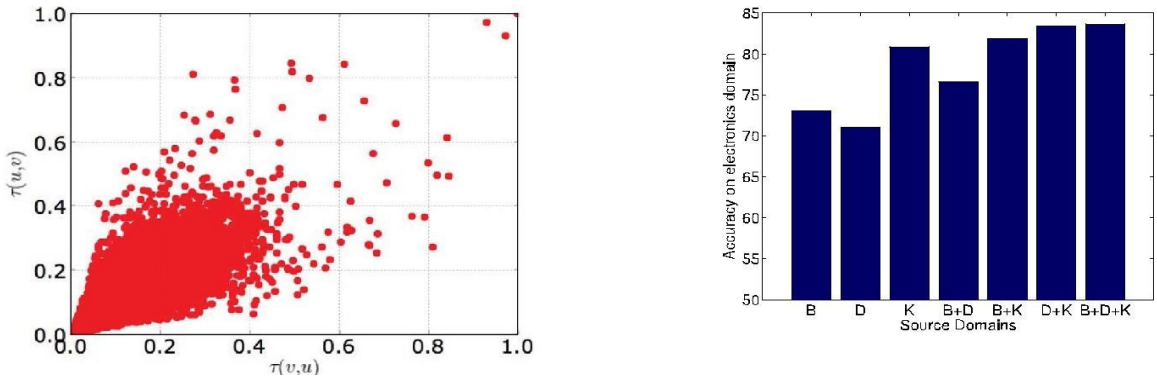


Fig. 2. Correlation between relatedness scores.

Different relatedness measures will list different lexical elements as neighbors for a particular lexical element. Therefore, the set of expansion candidates will be directly influenced by the relatedness measure used to create the thesaurus. To study the effect of From Table 5, we see that the Proposed relatedness measure reports the highest overall classification accuracy followed by the Reversed baseline, Lin's Similarity Measure, and the Cosine Similarity in that order. However, it must be noted that the differences in performance among those

ISSN(Online):2320-9801
ISSN(Print): 2320- 9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 5, May 2014

relatedness measures are not statistically significant. This result implies that a wide-range of relatedness measures can be used to create a sentiment sensitive thesaurus to be used with the feature expansion method proposed in the paper. Further investigations into the insensitivity of the proposed method to the relatedness measures revealed three important reasons that we will discuss next.

First, recall that the proposed feature expansion method (see Section 5) does not use the absolute value of relatedness scores, but only uses the relative rank among the expansion candidates. Therefore, two relatedness measures that produce different absolute scores can obtain similar performance if the relative rankings among expansion candidates are similar.From Fig. 2, we see that _ðu; vÞ is highly correlated to _ðv; uÞ. In fact the Pearson correlation coefficient for Fig. 2 is as high as 0.8839 with a tight confidence interval of [0.8835, 0.8844]. This experimental result indicates that, although by definition (2) is asymmetric, its level of asymmetry is very small in practice. Both the Proposed method and its Reversed baseline (8) reporting similar accuracy values in Table 5 further supports this finding.

5.4 Feature Analysis

To analyze the features learned by the proposed method we train the proposed method using kitchen, DVDs, and electronics as source domains. The proposed feature expansion method produces 137,635 unique features for 4,773 reviews. However, the L1 regularization produces a sparse model that contains only 1,668 features by selecting the most discriminative features from the training in-stances. For three example features, Table 6 shows their model weights and top three expansions. Correct related features are found as expansion candidates by the proposed method. For example, excellent is expanded by bigram invaluable þ resource, and worst is expanded by the bigram absolute þ junk.

6.6 Comparison against Previous Work

We compare our proposed method against two previously proposed cross-domain sentiment analysis methods. Next, we briefly describe those methods. They are described in detail in Section 8.

. SCL-MI. This is the structural correspondence learning (SCL) method proposed by Blitzer et al. [25]. This method utilizes both labeled and un-labeled data in the benchmark data set. It selects pivots using the mutual information between a feature (unigrams or bigrams) and the domain label. Next, linear classifiers are learned to predict the existence of those pivots. The learned weight vectors are arranged as rows in a matrix and singular value decomposition (SVD) is performed to reduce the dimensionality of this matrix. Finally, this lower dimensional matrix is used to project features to train a binary sentiment classifier.

. Spectral feature alignment (SFA). This is the SFA method proposed by Pan et al. [8]. Features are classified as to domain-specific or domain-indepen-dent using the mutual information between a feature and a domain label. Both unigrams and bigrams are considered as features to represent a review. Next, a bipartite graph is constructed between domain-specific and domain-independent features. An edge is formed between a domain-specific and a domain-independent feature in the graph if those two features cooccur in some feature vector.

ISSN(Online):2320-9801
ISSN(Print): 2320- 9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 5, May 2014

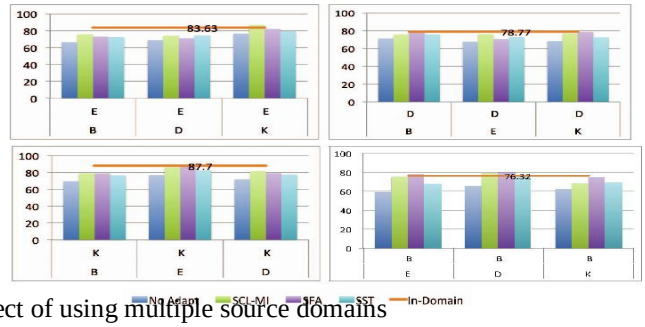


Fig. 3. Effect of using multiple source domains

VI. RELATED WORK

Sentiment classification systems can be broadly categorized into single-domain [1], [2], [27], [28], [29], [30] and cross-domain [7], [8] classifiers based upon the domains from which they are trained on and subsequently applied to. On another axis, sentiment classifiers can be categorized depending on whether they classify sentiment at word level [31], [32], sentence level [33], or document level [1], [2]. Our method performs cross-domain sentiment classification at document level.

In single-domain sentiment classification, a classifier is trained using labeled data annotated from the domain in which it will be applied. Turney [2] measures the cooccurrences between a word and a set of manually selected positive words (e.g., good, nice, excellent, and so on) and negative words (e.g., bad, nasty, poor, and so on) using pointwise mutual information to compute the sentiment of a word. Kanayama and Nasukawa [29] proposed an approach to build a domain-oriented sentiment lexicon to identify the words that express a particular sentiment in a given domain. By construction, a domain specific lexicon considers sentiment orientation of words in a particular domain. Therefore, their method cannot be readily applied to classify sentiment in a different domain.

Compared to single-domain sentiment classification, which has been studied extensively in previous work [3], cross-domain sentiment classification has only recently received attention with the advancement in the field of domain adaptation [34], [35], [36]. Aue and Gammon [37] report a number of empirical tests on domain adaptation of sentiment classifiers. They use an ensemble of nine classifiers to train a sentiment classifier. However, most of these tests

Blitzer et al. [7] propose the SCL algorithm to train a cross-domain sentiment classifier. SCL is motivated by the alternating structural optimization (ASO), a multitask learning algorithm, proposed by Ando and Zhang [38]. Given labeled data from a source domain and unlabeled data from both source and target domains, SCL chooses a set of pivot features which occur frequently in both source and target domains. Next, linear predictors are trained to predict the occurrences of those pivot features. Positive training instances for a particular pivot feature are auto-matically generated by removing the corresponding pivot feature in feature vectors. Feature vectors that do not contain a particular pivot feature are considered as negative training instances for the task of learning a predictor for that pivot feature. It is noteworthy that this approach does not require any manually labeled feature vectors for learning the pivot feature predictors. For each pivot feature, a linear weight vector is computed and the set of weight vectors for all the pivot features under consideration are arranged in a matrix. Next, SVD is performed on this weight matrix to construct a lower dimensional feature space. Each feature vector is then mapped to a lower dimensional representa-tion by multiplying with the computed matrix. Finally, each original

ISSN(Online):2320-9801
ISSN(Print): 2320- 9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 5, May 2014

feature vector is augmented with its lower dimensional representation to form a new (extended) feature vector. A binary classifier is trained using labeled reviews (positive and negative sentiment labels) using this new set of feature vectors. In the SCL-MI approach, a variant of the SCL approach, mutual information between a feature and the source label is used to select pivot features instead of the cooccurrence frequency. However, in practice it is hard to construct a reasonable number of auxiliary tasks from data, which might limit the transfer ability of SCL for cross-domain sentiment classification. Moreover, the heur-istically selected pivot features might not guarantee the best performance on target domains. In contrast, our method uses all features when creating the thesaurus and selects a subset of features during training using L1 regularization. Moreover, we do not require SVD, cubic in time complexity, which can be computationally costly for large data sets.

VII. CONCLUSION

We proposed a cross-domain sentiment classifier using an automatically extracted sentiment sensitive thesaurus. To overcome the feature mismatch problem in cross-domain sentiment classification, we use labeled data from multiple source domains and unlabeled data from source and target domains to compute the relatedness of features and construct a sentiment sensitive thesaurus. We then use the created thesaurus to expand feature vectors during train and test times for a binary classifier. A relevant subset of the features is selected using L1 regularization. The proposed method significantly outperforms several baselines and reports results that are comparable with previously pro-posed cross-domain sentiment classification methods on a benchmark data set. Moreover, our comparisons against the SentiWordNet show that the created sentiment-sensitive thesaurus accurately groups words that express similar sentiments. In future, we plan to generalize the proposed method to solve other types of domain adaptation tasks.

REFERENCES

[1] 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),
79-86, 2002.

[2] P.D. Turney, "Thumbs Up or Thumbs Down? Semantic Orienta-tion Applied to Unsupervised Classification of Reviews," Proc. 40th Ann. Meeting on Assoc. for Computational Linguistics (ACL '02),
a. 417-424, 2002.

[3] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis,"

[4] Foundations and Trends in Information Retrieval, vol. 2, nos. 1/2,
a. 1-135, 2008.

[5] Y. Lu, C. Zhai, and N. Sundaresan, "Rated Aspect Summarization of Short Comments," Proc. 18th Int'l Conf. World Wide Web (WWW '09), pp. 131-140, 2009.

[6] T.-K. Fan and C.-H. Chang, "Sentiment-Oriented Contextual Advertising," Knowledge and Information Systems, vol. 23, no. 3, 321-344, 2010.

[7] R.K. Ando and T. Zhang, "A Framework for Learning Predictive Structures from Multiple Tasks and Unlabeled Data," J. Machine Learning Research, vol. 6, pp. 1817-1853, 2005.

[8] K. Yoshida, Y. Tsuruoka, Y. Miyao, and J. Tsujii, "Ambiguous Part-of-Speech Tagging for Improving Accuracy and Domain Portability of Syntactic Parsers," Proc. 20th Int'l Joint Conf. Artificial Intelligence (IJCAI '07), pp. 1783-1788, 2007.

[9] H. Guo, H. Zhu, Z. Guo, X. Zhang, X. Wu, and Z. Su, "Domain Adaptation with Latent Semantic Association for Named Entity Recognition," Proc. Ann. Conf. North Am. Ch. Assoc. for Computa-tional Linguistics (NAACL '09), pp. 281-289, 2009.

[10] M. Dredze, J. Blitzer, P.P. Talukdar, K. Ganchev, J.V. Graca, and F. Pereira, "Frustratingly Hard Domain Adaptation for Dependency Parsing," Proc. CoNLL Shared Task Session of EMNLP-CoNLL (CoNLL '07), pp. 1051-1055, 2007.

[11] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, "Analysis of Representations for Domain Adaptation," Proc. Advances in Neural Information Processing Systems Conf., 2006.

BIOGRAPHY

Manjunathan Nanjundan is a Assistant Professor in Department Computer Science and Engineering, Asan Memorial college of Engineering and Technology, Anna University. He received Master of Engineering (ME) degree 2009 from Anna university, Chennai, India. Her research interests are Computer Networks (wireless Networks),Database, Data Mining etc.

ISSN(Online):2320-9801
ISSN(Print): 2320- 9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 5, May 2014