International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 4 | ISSN : 2456-3307

Cross-Domain Opinion Classification Exploitation Enhanced Sentiment Sensitive Thesaurus Aware Embeddings

Bade Ankammarao*[1], Mande Padma[2]
[1]Assistant Professor, Department of MCA , St. Mary's Group of Institutions, Guntur, Andhra Pradesh, India
[2]PG Student, Department of MCA , St. Mary's Group of Institutions, Guntur, Andhra Pradesh, India

ABSTRACT

Users can express their opinion and sentiments in various review sites in the internet. Sentiment classification deals with the extraction of useful information from unstructured data, which can be used in various applications. Sentiment classification predicts the polarity of each opinionated review. It helps the customers to choose and the manufacturer to rate their product/services. Cross domain sentiment classification helps in classifying the reviews across various domains at much lower cost and time. This paper presents a short survey on various techniques used to implement cross domain sentiment analysis. Unsupervised Cross-domain Sentiment Classification is the task of adapting a sentiment classifier trained on a particular domain (source domain), to a different domain (target domain), without requiring any labeled data for the target domain. By adapting an existing sentiment classifier to previously unseen target domains, we can avoid the cost for manual data annotation for the target domain. We model this problem as embedding learning, and construct three objective functions that capture: (a) distributional properties of pivots (i.e., common features that appear in both source and target domains), (b) label constraints in the source domain documents, and (c) geometric properties in the unlabeled documents in both source and target domains. Unlike prior proposals that first learn a lower-dimensional embedding independent of the source domain sentiment labels, and next a sentiment classifier in this embedding, our joint optimisation method learns embeddings that are sensitive to sentiment classification. Experimental results on a benchmark dataset show that by jointly optimising the three objectives we can obtain better performances in comparison to optimising each objective function separately, thereby demonstrating the importance of task-specific embedding learning for cross-domain sentiment classification. Among the individual objective functions, the best performance is obtained by (c). Moreover, the proposed method reports cross-domain sentiment classification accuracies that are statistically comparable to the current state-of-the-art embedding learning methods for cross-domain sentiment classification.

Keywords: Cross Domain; Deep Learning; Embedding Learning; Enhanced Sentiment Sensitive Thesaurus.

I. INTRODUCTION

With the increase in internet based services, people express their opinions about products online. Such sentiment information obtained from the customers is growing exponentially. Thus making it difficult for the manufacturer to classify the nature of 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, market analysis, opinion summarization and contextual analysis.[1]

Specific domain is used in sentiment analysis to provide greater accuracy. Sentiment analysis uses feature vector that has a collection of words which are limited and specific to particular domain (domain can be consider as student, school etc.). However sentiments hold different meanings in different domains and it is costly to annotate data for each new domain in which we would like to apply a sentiment classifier. Cross domain sentiment analysis can be considered as the solution to this problem but the

problem is that classifier trained in one domain may not work well when applied to other domain due to mismatch between domain specific words. So before applying trained classifier on target domain some techniques must be applied like feature vector expansion, finding relatedness among the words of source and target domain, etc. A different technique gives different analysis, result and accuracy which depend on the documents, domain taken into consideration for classification.

In literature, Sinno Jialin Pan et al.[2] 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.
In 2013, Danushka Bollegala et al. [5] developed a technique which uses sentiment sensitive thesaurus (SST) for performing cross-domain sentiment analysis. They proposed a cross-domain sentiment classifier using an automatically extracted sentiment sensitive thesaurus. To handle the mismatch between features in cross-domain sentiment classification, they 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. 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.

P.Sanju et a1.[1] proposed cross domain sentiment classification by creating enhanced sentiment sensitive thesaurus which 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.

Danushka Bollegala et al. [3] proposed embedding learning, constructing three objective functions that capture: (a) distributional properties of pivots (i.e. common features that appear in both source and target domains), (b) label constrains in the source domain documents, and (c) geometric properties in the unlabeled documents in both source and target domains.

SCL-MI is the structural correspondence learning (SCL) method proposed by Blitzer et al. [6]. In this method they utilizes both labeled and unlabeled 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.

II. LITERATURE SURVEY

Spectral Feature Alignment

In this algorithm a set of labeled data is considered from the source domain. A set of unlabeled data is obtained from the target domain for help to train the classifier for the target domain.SFA algorithm creates a new representation of data in order to reduce the gap between two domains.

Firstly, domain independent features are identified based on the frequency of occurrence and mutual information. Mutual information can be used to measure the dependency between features and domains. A domain specific feature will have high mutual information which will otherwise be domain independent. Also, the domain independent features must occur frequently.These domain independent words are then used to construct a biparite graph which acts as a bridge and models the co-occurrence relationship between domain specific and domain independent words. If two domain-specific words have connections to more common domain-independent words in the graph, they tend to be aligned together with higher probability. Similarly, if two domain-independent words have connections to more common domain-specific words in the graph, they tend to be aligned together with higher probability[2].A clustering algorithm based on graph spectral theory is then adapted on feature biparite graph to align domain specific features. We assume that (a) Two domain-specific features tend to be very related and will be aligned to a same cluster with high probability if they are connected to many common domain- independent features, (b) Two domain -independent features tend to be very related and will be aligned to a same cluster with high probability if they are connected to many common domain-specific features, (c) To reduce the gap between domains we can find a more compact and

meaningful representation for domain-specific features. Therefore, by applying graph spectral techniques on feature biparite graph, the mismatch problem between plural and domain specific features can be removed.

[2]The algorithm is as follows:

Input: labeled source domain data Dsrc = {(xsrci, ysrci)} nsrc for i=1, unlabeled target domain data Dtar = {xtarj}ntar

for j=1, the number of cluster *K* and the number of domain

independent

features *m*

Output: adaptive classifier *f: X Y*

1. Apply the criteria on *Dsrc* and D*tar* to select *l* domain-independent features. The remaining *m - l* features are treated as domain-specific features.

$$\emptyset DI = \begin{bmatrix} \emptyset DI(xsrc) \\ \emptyset DI(xtar) \end{bmatrix}, \emptyset DS = \begin{bmatrix} \emptyset DS(xsrc) \\ \emptyset DS(xtar) \end{bmatrix}$$

By using ΦDI and ΦDS , calculate (Di-words)-(DS-word) co-occurrence matrix M Є $R^{(m-l)xl}$

2. Construct matrix L = $D^{-1/2}$ $AD^{-1/2}$, Where

$$A = \begin{bmatrix} 0 & M \\ M^T & 0 \end{bmatrix}$$

3. Find the K largest eigenvectors of L, u1, u2, ...,uk, and form the matrix U = [u1 u2 ... uk] belongs to R m*k.

Let mapping φ (X_i) =X_i U[l:m-l,:].where X_i Є R $^{m-l}$ Return a classifier f, trained on

$$\{([x_{src_i}\ \gamma\varphi(\phi_{DS}(x_{src_i}))], y_{src_i})\}_{i=1}^{n_{src}}$$

III. ENHANCED SENTIMENT SENSITIVE THESAURAS

In this technique, an enhanced sentiment sensitive thesaurus is created which aligns semantically similar features from different domains and also sentiment features from wiktionary. The focus of this method is to extract more features from wiktionary ,with the help of java wiktionary library tool(JWKTL) ,which are then appended to ESST to provide a better performance.

Firstly, the sentences are split into parts and then parts of speech(POS) tagging is performed followed by Lemmatization using RASP[4]. Lemmatization converts singular words into base form and unwanted words are eliminated. With the help of POS tagging, unigrams and bigrams are extracted from the reviews. Next, sentiment features are created by appending the label of the review to each feature from each source domain labeled reviews. The notation *p to indicate positive features and *N to indicate negative features. Domain independent features are then extracted by computing mutual information between domain and features. If a feature has high mutual information with the domain then it is considered as domain specific whereas if it has less mutual information with domain then it is considered as domain independent. Using domain independent and domain specific features, a co-occurrence matrix is created. From the co-occurrence of the words found in documents, semantic meanings of the words are calculated. After that, values of the features in the co-occurrence matrix areweighted using point wise mutual inf ormation equation.

[1] 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 features. semantically similar domain specific features are aligned using ESST by computing similarity measure equation in PMI weighted matrix. ESST list up many domain specific features based on descending order of the similarity score for every domain specific feature.

ESST also collects more semantically similar features from wiktionary using JWKTL. The seed adjectives are extracted from the review sarefed into the dictionary and corresponding glossaries are obtained from it.The unigrams and bigrams for each adjective obtained from the glossaries are then added to ESST. The domain specific features of various domains are then augmented with 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-domainsentiment classification. This new representation of feature vector is used to train a sentiment classifier to predict the label of target domain.

The algorithm is as follows:

Input: labeled source Domain data Dsr= { Xsr,Ysr} and unlabeled source domain Dsr={Xi} and unlabeled target Domain Dtr = {Xtr}and Wiktionary dump file
Output: predict the label of target domain.

- Extract Domain Independent features and domain specific features from the given Reviews.
- Create co-occurrence matrix between domain independent features with domain specific features.
- Compute Point Wise Mutual Information for each features using equation (1).
- 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.
- 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 that are appended to ESST.
- Find a new representation of feature vector by Feature augmentation while training a classifier.
- Test the classifier in target domain.

IV. SENTIMENT SENSITIVE EMBEDDINGS

Projecting the source and the target features into the same lower-dimensional embedding, and subsequently learning a sentiment classifier on this embedded feature space is a popular approach to cross domain sentiment classification. It is useful only when there is little overlap between original source and target feature spaces.[3] A limitation of this two-step approach that decouples the embedding learning and sentiment classifier training is that the embeddings learnt in the first step is agnostic to the sentiment of the documents, which is the ultimate goal in cross-domain sentiment classification.

In the proposed technique,spectral embedding are used to project words and documents into the same lower dimensional embeddings in comparison to 318ptimizing each objective function separately.

V.CONCLUSION

In this paper three different techniques used for cross domain classification are studied. Spectral feature alignment technique using spectral feature alignment algorithm using spectral graphs. The Enhanced Sentiment Sensitive Thesaurus technique uses java wiktionary to align sentiment features and provide a better performance. Lastly, sentiment sensitive embeddings technique uses spectral embeddings to project words and documents into the same lower dimensional embeddings.

VI REFERENCES

[1]. P. Sanju, T.T.Mirnalinee.Cross Domain Sentiment Classification Using Enhanced Sentiment Sensitive Thesaurus (ESST). 2013 Fifth International Conference on Advanced Computing (ICoAC).

[2]. Sinno Jialin Pany, Xiaochuan Niz, Jian-Tao Sunz, Qiang Yangy and Zheng Chen.Cross-Domain Sentiment Classification via Spectral Feature Alignment

[3]. Danushka Bollegala,Tingting Mu,John Y. Goulermas.Cross domain Sentiment Classification using Sentiment Sensitive Embeddings

[4]. T. Briscoe, J. Carroll, and R. Watson, "The Second Release of the RASP System," Proc. COLING/ACL Interactive Presentation Sessions Conf., 2006.

[5]. Danushka Bollegala, David Weir, and John Carroll, "Cross-Domain Sentiment Classification Using a Sentiment Sensitive Thesaurus", IEEE transactions on knowledge and data engineering, VOL. 25, NO. 8, August 2013.

[6]. J. Blitzer, M. Dredze, F. Pereira, "Domain Adaptation for Sentiment Classification", 45th Annv. Meeting of the Assoc. Computational Linguistics (ACL'07).

[7]. Ms Kranti Ghag and Dr. Ketan Shah, "Comparative Analysis of the Techniques for Sentiment Analysis", ICATE 2013

[8]. B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," Foundations and Trends in Information Retrieval, vol. 2, nos. 1/2, pp. 1-135, 2008.

[9]. 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.