Sentiment Analysis for Opinion Mining Using Cross-Domain Classifier

Pravin Jambhulkar[1]
M.tech. Scholar, Dept. of Computer Science & Engg, RCOEM, Nagpur, India [1]
jambhulkar.pravin29@gmail.com

Smita Nirkhi[2]
Asst. Prof., Dept. of Computer Science & Engg, RCOEM, Nagpur, India [2]
smita811@gmail.com

Abstract: Now a day's sentiment analysis is important for various task and applications like market analysis, opinion mining, contextual advertising, etc. Domain generalization remains a challenge in sentiment analysis hence we proposed a methodologies to perform cross-domain sentiment analysis. In cross-domain sentiment analysis, classifier trained on one domain is used to classify other domain. We create a glossary using labeled data from source domain and unlabeled data from both source and target domain. This glossary is used to handle a feature mismatch problem, and contains clusters of semantically similar words. For generating a glossary, first we calculate the co-occurrence matrix by point wise mutual information (PMI) [1] and using distributional hypothesis [2] we efficiently create a glossary. At test time, this glossary will be used to find the similar words, and hence solve the feature mismatch problem. Proposed methodologies will really outperform and achieve accuracy near to domain adaptation.

Keywords: cross-domain, sentiment glossary, feature vector, distributional relatedness,

*****

I. Introduction

With the massive growth in web services, huge amount of useful data are generated by people. For example, through various social websites like facebook, twitter, etc various forums, blogs on diverse topic. This data is useful for both consumer as well as producer point of view. Also people's opinion is useful for taking decision of various social aspects. But it is not possible to manually read and analyze these data hence sentiment analysis classify that data which can help people to understand this massive data in abstract. As a result, sentiment analysis has attracted much attention recently, for example, opinion summarization, opinion integration and review spam identification, etc.

Most of the sentiment analysis is carried out by targeting a particular domain to achieve higher accuracy. But collecting a training data is expensive and time consuming for each new domain because sentiment expressed differently in different domain. Domain generalizations still a big challenge in sentiment analysis because the words used in one domain may or may not use in another domain. Hence most of the features are unseen to that classifier which is trained on another domain. For performing a cross-domain sentiment analysis we require a framework to incorporate the information regarding relatedness among the features, hence created a glossary [3].

In this paper, for solving the feature mismatch problem of cross-domain sentiment analysis we are creating a glossary which contains the words that are semantically similar. This glossary will then used to extend the features [4] that are present in the review of target domain. Extended features are appended to the original features of review and then by following the bag of words, classifier will accurately classify that features.

The rest of the paper is organized as follows; in the next section first we see the review of related work in cross-domain sentiment analysis. Then we explain the idea behind our proposed method in section 3. Finally we conclude our work in section 4.

II. Literature Review

In [2] 2013, Danushka Bollegala et al. [4] 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 overcome the feature mismatch problem 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 extend feature vectors during train and test times for a binary classifier.

Spectral feature alignment (SFA) method is first proposed by Pan et al. [5] in 2010. In this, features are classified as to domain-specific or domain-independent 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 co-occur in some feature vector. Spectral clustering is conducted to identify feature clusters. Finally, a binary classifier is trained using the feature clusters to classify positive and negative sentiment.

A semi-supervised (labeled data in source, and both labeled and unlabeled data in target) extension to a well-known supervised domain adaptation approach is proposed [23]. This semi-supervised approach to domain adaptation is extremely simple to implement, and can be applied as a pre-processing step to any supervised learner. However, despite their simplicity and empirical success, it is not theoretically apparent why these algorithms perform so well. Compared to single-domain sentiment classification, cross-domain sentiment classification has recently received attention with the advancement in the field of domain adaptation.

SCL-MI. This is the structural correspondence learning (SCL) method proposed by Blitzer et al. [6]. This method 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, binary classifiers are trained 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.

III. OUR APPROACH

In this section, we describe our proposed methodology and techniques for performing cross-domain sentiment analysis. In cross-domain sentiment analysis, feature mismatch problems need to be handle and for that we are creating an automatic glossary which contains the semantically similar words. We then extend a feature vector at training and testing time which find the similar features from glossary. Fig.1 represents the architecture of our approach, in which Li and Ui denotes the labeled instances and unlabeled instances respectively.

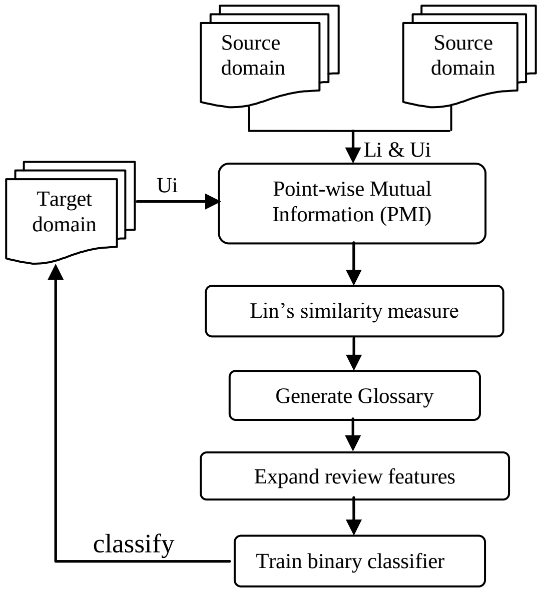

Fig.1 System architecture

I. Sentiment Glossary

Features from one domain may or may not come in other domain hence feature mismatch problem have to be solved in cross-domain sentient analysis. Hence to overcome this problem we are creating a sentiment glossary. Sentiment glossary contains the semantically similar words which is generated using point-wise mutual information technique (PMI) and distributional hypothesis. Next we discuss the procedure to construct a sentiment glossary.

A. Lexical and Sentiment element

Given a dataset, first we split them into sentences and then preprocess by POS tag and lemmatization technique using RASP system [7]. In POS tagging we append one of the tag to each word of the sentence which then used to retain functional words in simple word filter process. Lemmatization [8] is the process by which we can lemmatize the inflected form of word to its lemma. We have both labeled and unlabeled instances from different domains. So we apply point-wise mutual information technique on these domains to calculate the frequency of co-occurrences of words which will further used to find the similarity among the words by using distributional hypothesis.

Tables 1 illustrate the generation of lexical and sentiment elements for a review sentence in Books domain.

Steps	Example
A review sentence	This is an interesting and well researched book
POS tagging & lemmatization	DT /This VBZ /is DT/ an JJ /interest+ ing CC /and RB /well VBN/ research+ed NN/ book ./
Noun, verb, adjective & adverb	Interesting, well, researched
Unigrams & bigrams	Interesting, well, researched, interesting+well, well+researched
Sentiment element	Interesting*p, well*p, researched*p, interesting+well*p, well+researched*p

Table 1 Example for Books domain

B. Word Co-occurrences

Co-occurrences of words are calculated by using point-wise mutual information technique. Point-wise mutual information for weighting the features are seen very useful for various natural language processing task such as word clustering [9], similarity measurement [10], etc. Point-wise mutual information is represented as,

$$PMI(w_i) = p(f \,\&\, w_i) \,/\, p(w_i)*p(f)$$

Here, w_i denote the ith word from the available corpus (w_1, w_2, w_3,……, w_n) and *f* represents the problem word. In this form, $p(f \,\&\, w_i)$ represent the probability that *f* (problem word) and w_i (choice) occurs together. Whereas $p(f)$ * $p(w_i)$ denotes the probability of feature *f* and word w_i occurs alone. If *f* and *w* are independent then the probability that they co-occur are smaller than they occur alone and hence the smaller PMI score. If they are not independent then they have a more tendency to co-occur and hence the probability that they co-occur is greater than they occurs alone, so the greater PMI score.

C. Distributional Relatedness

Two words are semantically related if they have many common co-occurring words and it is not necessary that they are syntactic in relation. If two words are related then they have more common co-occurring words. For example, *cancer is a disease* and *diabetes is a disease*. Both cancer and diabetes have similar co-occurring words, hence they are semantically similar.

For computing the relatedness among the words we are using similarity measure proposed by Lin [11]. Next, for two words *s* and *t*, we compute the relatedness score of the element *t* to that of *s* as,

$$sim(t,s) = \frac{\sum_{w \in \{z|f(t,z) \cap f(s,z) > 0\}} (f(t,w) + f(s,w))}{\sum_{w \in \{z|f(t,z)>0\}} f(t,w) + \sum_{w \in \{z|f(s,z)>0\}} f(s,z)}$$

Here, $z|f(t, z)>0$ is the set of features of *z* that has positive point-wise mutual information value for the element *t*. $f(t, w)$ is the point-wise mutual information between an element *t* and a feature *w*, similarly for *s*. Lin's similarity measure perform so well for word clustering task gives the efficient output as compared to other numerous similarity measures.

Lin's similarity measure gives cluster of similar word as an output and using approximate vector similarity computation technique [12], we can efficiently create a glossary which will further used to extend the features of reviews.

D. Extend review features

Next task is that we append the additional features (selected from the sentiment glossary which we have created in previous section) to the review features.

First we take a review as a bag-of-words model in which we retain only functional words $\{w_1, w_2,\ldots,w_n\}$ and find the features from the glossary which are similar to the features of the review. The additional feature $\{f_1, f_2,\ldots, f_n\}$ are selected from the glossary iff there are more words in the reviews that are also listed as a neighbor of base entry in a glossary. We select those entries and extend the feature vector at training and testing time. Now this feature vector $\{ w_1, w_2,\ldots, w_n \}+\{ f_1, f_2,\ldots, f_n \}$ is used to train a binary classifier from a source domain which will accurately classify the target domain. The above procedure is used to extend the review feature at training and testing time to predict the sentiment of the target domain.

IV. Experimental Setup

The proposed work is carried out on Intel core i5 processor with 4GB RAM. The development environment used for implementing the proposed model is Python which runs on Ubuntu 12.04.3 LTS 64-bit operating system. The proposed system uses RASP system for POS tagging and lemmatization. The RASP system supports XML files for tagging. The system uses MLIB python Library for vector computation. Next we will explain the RASP system and machine libraries (MLIB) used in proposed system.

A. RASP system

RASP system is used to perform POS tagging and lemmatization process. The tokenized text is tagged with one of 150 part-of-speech (PoS) and punctuation labels (derived from the CLAWS tagset). This is done using a first-order ('bigram') hidden markov model (HMM) tagger implemented in C. The analyzer takes a word form and CLAWS tag and returns a lemma plus any inflectional affixes.

B. MLIB

It is a collection of various machine learning algorithms and data mining algorithms implemented in python. It was implemented by Danushka Bollegala, a Senior Lecturer in the University of Liverpool for a research purposes. MLIB uses Python version 2.7 or later. MLIB is released under BSD License for non-commercial academic use.

C. Classias Binary Classifier

We are using classias binary classifier which is a collection of machine-learning algorithms for classification. It supports L1/L2-regularized logistic regression (Maximum Entropy). L1 regularization is shown to produce a sparse model, where most irrelevant features are assigned a zero weight [22]. This enables us to select useful features for classification in a systematic way without having to preselect features using heuristic approaches.

D. Datasets

We use the cross-domain sentiment classification data set prepared by Blitzer et al. [6] 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.

Table 2: Number of reviews in Amazon Product Reviews Dataset

Domain	Positive	Negative	Unlabeled
Kitchen	800	800	16746
DVDs	800	800	34377
Books	800	800	13116
Electronics	800	800	5947

For each domain, there are 1000 positive and 1000 negative examples, the same balanced composition as the polarity dataset. The dataset also contains some unlabeled reviews for the four domains. The proposed work randomly selects 400 positive and 400 negative labeled reviews from each domain as training instances (total number of training instances are 400*3 = 1200). To conduct experiments, the proposed system selects each domain in turn as the target domain, with one or more other domains as source domain.

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:

$$Accuracy = \frac{no. of\ totally\ classified\ target\ review}{total\ no. of\ reviews\ in\ target\ domain}$$

V. RESULT

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

No adapt: In this method there is no any feature expansion perform. 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 domain. This gives the lowest accuracy in cross-domain sentiment analysis process.

Proposed method: This is the proposed method described in this paper. We use the sentiment glossary created using the procedure described in Section III and use the glossary for extending the reviews feature in a binary classifier.

In-domain. In this method, we train a binary classifier using the labeled data from the target domain. This method gives the best result for the cross-domain sentiment analysis. This upper baseline demonstrates the classification accuracy we get this accuracy if we had labeled data for the target domain. But note that this is not a cross-domain classification setting.

Table 3. Comparison of proposed method with other methods.

Domain	No Adapt	**Proposed**	In-Domain
kitchen	0.7261	**0.8318**	0.8770
DVDs	0.6897	**0.7626**	0.8240
Electronics	0.7053	**0.8086**	0.8440
Books	0.6272	**0.7332**	0.8040

Table 3 shows the accuracy of the above proposed methods for each of the four domains on the basis of data set used as the target domain. Moreover, we are showing the proposed method accuracy in boldfaces for sentiment classification results. From the results in Table 4, we see that the proposed method returns the best cross-domain sentiment classification accuracy for all four domains.

A. Multiple Source Domains

The proposed method uses multiple source domains to train a binary classifier. The accuracy will vary according to the domain used for training purpose. Fig. 2 shows the comparison of using multiple source domains and single source domain.

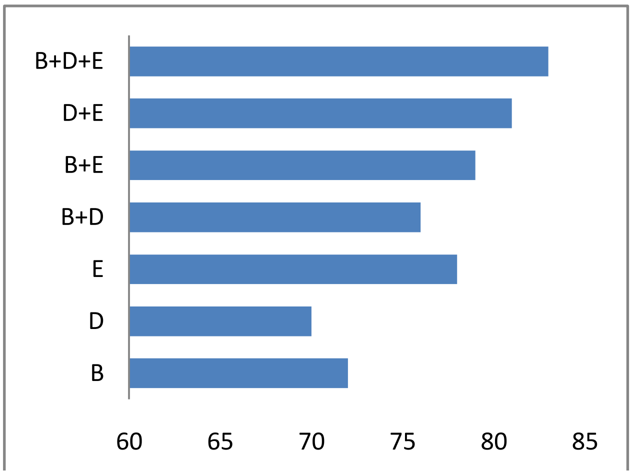

Fig. 2 Comparison with multiple source domains

Accuracy in each case show variation and it is highest when we combine all three domains to train a classifier. So we conclude that as we combine more domains we get more accuracy.

VI. CONCLUSION

We present an approach of cross-domain sentiment analysis in which we create a sentiment glossary which will used to handle the feature mismatch problem of cross-domain sentiment classification. We created a glossary using labeled and unlabeled instances of source and target domain in which we apply PMI and distributional relatedness measure to compute the co-occurrences and similarity among the words. Finally we present how to extend review features which will further used to train a binary classifier. Next we explain a softwares and tools used for experimental purpose. Finally we evaluate and compare our result with existing approaches.

REFERENCES.

[1] Yan Xu et al. “A Study on Mutual Information-based Feature election for Text Categorization” Journal of Computational Information Systems 3:3 (2007) 1007-1012

[2] P. Pantel and D. Ravichandran, “Automatically Labeling Semantic Classes,” Proc. Conf. North Am. Ch. Assoc. for Computational Linguistics: Human Language Technologies (NAACL-HLT ’04), pp. 321-328, 2004.

[3] Gregory Grefenstette, “Automatic Thesaurus Generation from Raw Text using Knowledge-Poor Technique” Making sense of Words, 9th Annual Conference of the UW Centre for the New OED and Text Research, 1993.

[4] 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.

[5] Sinno Jialin Pan, Xiaochuan Niz, Jian-Tao Sunz, Qiang Yangy, Zheng Chen, “Cross-Domain Sentiment Classification viaSpectral Feature Alignment”, 19th Int’l Conf. World Wide Web (WWW’10).

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

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

[8] T. Joachims, “Text Categorization with Support Vector Machines: Learning with Many Relevant Features,” Proc. 10th European Conf. Machine Learning (ECML ’98), pp. 137-142, 1998.

[9] D. Lin, “Automatic Retrieval and Clustering of Similar Words,” Proc. Ann. Meeting of the Assoc. Computational Linguistics (ACL ’98), pp. 768-774, 1998.

[10] P. Turney, “Similarity of Semantic Relations,” Computational Linguistics, vol. 32, no. 3, pp. 379-416, 2006.

[11] D. Lin, “Automatic Retrieval and Clustering of Similar Words,” Proc. Ann. Meeting of the Assoc. Computational Linguistics (ACL ’98), pp. 768-774, 1998.

[12] S. Sarawagi and A. Kirpal, “Efficient Set Joins on Similarity Predicates,” Proc. ACM SIGMOD Int’l Conf. Management of Data, pp. 743-754, 2004.

[13] M. Hu and B. Liu, “Mining and Summarizing Customer Reviews,” Proc. 10th ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining (KDD ’04), pp. 168-177, 2004.

[14] S. Xie, W. Fan, J. Peng, O. Verscheure, and J. Ren. Latent space domain transfer between high dimensional overlapping distributions. In 8th International World Wide Web Conference, pages 91–100, April 2009.

[15] J.M. Wiebe, “Learning Subjective Adjective from Corpora,” Proc. 17th Nat’l Conf. Artificial Intelligence and 12th Conf. Innovative Applications of Artificial Intelligence (AAAI ’00), pp. 735-740, 2000.

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

[17] D. Lin, “Automatic Retrieval and Clustering of Similar Words,” Proc. Ann. Meeting of the Assoc. Computational Linguistics (ACL ’98), pp. 768-774, 1998.

[18] N. Jindal and B. Liu. Opinion spam and analysis. In Proceedings of the international conference on Web search and web data mining, pages 219–230, Palo Alto, California, USA, 2008. ACM.

[19] S. Xie, W. Fan, J. Peng, O. Verscheure, and J. Ren. Latent space domain transfer between high dimensional overlapping distributions. In 8th International World Wide Web Conference, pages 91–100, April 2009

[20] 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.

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

[22] A.Y. Ng, “Feature Selection, l1 vs. l2 Regularization, and Rotational Invariance,” Proc. 21st Int’l Conf. Machine Learning (ICML ’04), 2004.

[23] Daum´e III.H, Abhishek.K, Avishek.S(2010), ‘Frustratingly Easy Semi-Supervised Domain Adaptation’, Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing, ACL 2010 pp. 53–59.