Cross Domain Sentiment Classification: Current Solutions

Neethu Kurian

***Abstract*— Cross-domain sentiment classification is the task of classifying documents in a domain of interest into appropriate subjective classes by utilizing information obtained from other domains. Methods to sentiment classification aim at bridging the gap between the source domain and target domain. This helps in adapting a classifier trained on source domain to predict the sentiment polarities of documents in target domain without the need for annotating the target domain data. This paper tries to give a brief overview on the existing methods of cross-domain sentiment classification.**

***Index Terms*— Sentiment classification, Cross-domain sentiment classification.**

I. INTRODUCTION

With the widespread of use internet and related services, opportunities for people to share information has increased largely. People around the world have found Internet as a new platform to share their experiences, opinions and knowledge on various topics. Online discussion groups, blogs, social networks, review sites etc. are only some of these social media over Internet. One basic property of text shared over these social media sites is their sentiment. Identifying the sentiment conveyed by these texts and classifying them accordingly is one of the hot topics of research in recent years. These tasks fall into the area of sentiment analysis or opinion mining.

The proliferation of social media has made online opinions an important aspect in determining the acceptability of products. Customers rely on online opinions to decide on the quality of products. Manufacturers rely on them to determine the customer opinion. Automated sentiment analysis has therefore become an important challenge.

Most of the approaches to sentiment analysis rely on labeled data from the same domain or product type. Annotating the text manually for each new domain is an expensive and time-consuming task. These difficulties have opened up the doors to cross-domain sentiment classification. Cross-domain sentiment classification aims at developing methods by which labeled data from one domain can be used to identify the sentiment orientation of opinionated texts in a different domain. Cross-domain sentiment classification is therefore useful in the case of domains where no labeled data or few labeled data is available.

Manuscript received May,2014
Neethu Kurian, Department of Computer Science and Engineering, Rajagiri School of Engineering and Technology. Cochin, India.

In this paper, I try to discuss the various approaches to sentiment classification and cross-domain sentiment classification. Section 1 introduces the topic. Section 2 and 3 describes sentiment classification and cross-domain sentiment classification.

II. SENTIMENT CLASSIFICATION

Sentiment analysis is the task of identifying and extracting subjective information from textual documents using natural language processing and data mining techniques. The key focus of sentiment analysis is to determine the attitude of a speaker or a writer towards some topic, or the overall contextual polarity of a document. The attitude may be the judgment or evaluation made by the author, or the emotional effect the author desires to have on the reader. In reviews, people use simple terms to express sentiment about a product or service. The major challenge in front of sentiment analysis algorithms is to identify the words contributing to the overall sentiment polarity of the text correctly. In addition to these cultural factors, linguistic variations and differing contexts makes sentiment classification an extremely difficult task.

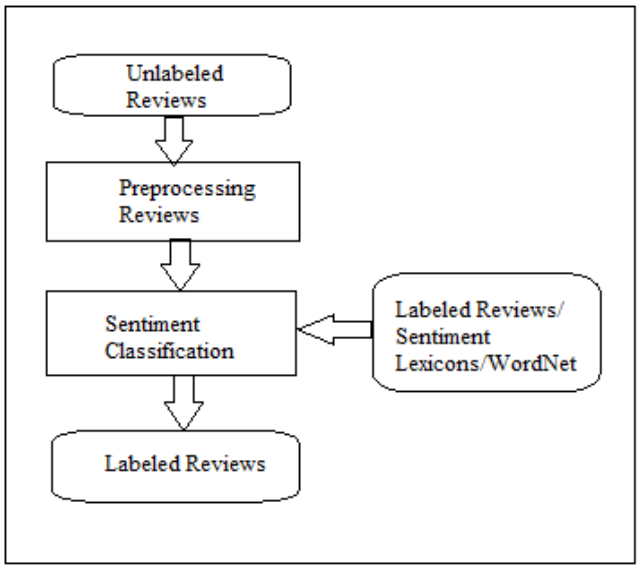

Fig 1 General Sentiment Classification Model

A general architecture of sentiment classification system is shown in Fig 1. The input to a sentiment analysis system consists of documents of any format whose sentiment labels

are not known (unlabeled documents). These documents preprocessed using a variety of NLP techniques such as stemming, tokenization, part of speech tagging, and dependency parsing. Sentiment classification systems utilize machine learning algorithms, lexicons or other linguistic resources to predict the sentiment labels. The output will be the documents annotated with sentiment labels. The annotations may be attached to whole documents (for document-level sentiment classification), to individual sentences (for sentence-level sentiment classification) or to specific aspects of entities (aspect-level sentiment classification) [6].

Machine learning techniques are being widely used for sentiment classification. These methods range from fully supervised techniques that utilize a labeled training corpus to unsupervised methods that make use of sentiment word lexicons like SentiWordNet and grammatical properties of text. Semi supervised methods that exploit the relatedness between words in target document and training corpus is also used for sentiment classification.

A basic task in sentiment analysis is classifying given text according to its sentiment polarity, i.e., whether the expressed opinion in a document, a sentence or feature/aspect of an entity is positive, negative, or neutral [1]. Another variation of sentiment classification is to classify according to the emotional states expressed, such as angry, sad, happy etc. A different approach to sentiment classification is the use of a scaling system whereby words commonly associated with having a negative, neutral or positive sentiment with them are given an associated number on a predefined scale [12]. Feature/aspect-based sentiment analysis is a more fine-grained instance of sentiment classification. It refers to determining the opinions or sentiments expressed on different features or aspects of entities.

A. Document Level Sentiment Classification

Two main approaches to document level sentiment analysis are, supervised learning and unsupervised learning. The supervised approach requires the availability of training data for each class. The training data is used by the system to learn a classification model, which can be used to predict the sentiment labels of input documents. Pang et al. [1] proposed a supervised method in which sentiment classification is viewed as a special case of traditional text classification problem. They use three machine-learning algorithms (SVM, Naive Bayes, Maximum entropy) to classify the documents. Unsupervised approaches to document- level sentiment analysis are based on determining the semantic orientation (SO) [2] of specific phrases within the document. The semantic orientation of the document is calculated as the average of phrase-level semantic orientations. If it is above some predefined threshold, the document is classified as positive and otherwise as negative. Turney [2] presented a simple unsupervised learning algorithm for classifying a review as recommended or not recommended. He determined the semantic orientation of words by computing the words point wise mutual information (PMI) [11] for their co-occurrence with a positive seed word (excellent) and a negative seed word (poor). PMI(P,W) gives the statistical dependence between the word P and the word W.

B. Sentence Level Sentiment Classification

Sentence level sentiment analysis provides a more fine-grained view of the different opinions expressed in the document. It aims at extracting sentiments associated with a sentence or phrase. Yu and Hatzivassiloglou proposed a method [15], which first creates prior-polarity lexicons. Sentiment label is assigned to a sentence by averaging the prior semantic orientations of instances of lexicon words in the sentence. A unique approach based on the minimum cuts was proposed in Pang and Lee [10]. This approach is based on the assumption that neighboring sentences should have the same subjectivity classification.

C. Aspect Level Sentiment Classification

Aspect-based sentiment analysis (also called feature-based sentiment analysis) is the research problem that focuses on the recognition of all sentiment expressions associated with the aspects within a document. A feature or aspect is an attribute or component of an entity, e.g., the screen of a cell phone, or the picture quality of a camera. Aspect level sentiment classification requires, identifying relevant entities, extracting their features/aspects, and determining whether an opinion expressed on each feature/aspect is positive, negative or neutral. A commonly used approach to identify aspects is to extract all noun phrases (NP) and measure for each candidate NP, the PMI with phrases that are tightly related to the product category (like phones, printers, or cameras) [13]. Another approach to aspect identification is to use a phrase dependency parser [14] that utilizes known sentiment expressions to find aspects.

III. CROSS DOMAIN SENTIMENT CLASSIFICATION

The supervised classification methods cannot perform well when training data and test data are from different domains. The main reason for this variation in performance is that training data do not have the same distribution with test data. As each domain has its own vocabulary to express sentiment, the sentiment information shared between two unrelated domains will be less. The poor performance found in the above-mentioned scenario can hence be accounted to the following two basic reasons. First, each domain has its own domain specific words and it will be different from one domain to another domain. For instance, the word warmhearted frequently appears in hotel reviews, but it hardly appears in electronics reviews. Secondly, words that have high correlations with certain class labels in the training domain may not have the same degree of correlations with the same class labels in the target domain. For instance, the word portable may be positive in electronics reviews, but it means nothing in hotel reviews.

Therefore, training a classifier with the labeled data in the same domain is found to give an upper bound in terms of performance. Hence, labeled data is considered as the most valuable resource for the sentiment classification. In some traditional domains or, plenty of labeled sentiment data are freely available on the web, but in other domains, labeled sentiment data are scarce and it involves much human labor to manually label reliable sentiment data. A sentiment classifier trained with the labeled data from one domain normally performs unsatisfactorily in another domain. Therefore, the challenge is how to utilize labeled sentiment

data in one domain (that is, source domain) for sentiment classification in another domain (that is, target domain). This resulted in the concept of cross-domain sentiment classification.

Several approaches to cross-domain sentiment classification have been proposed. A good number of them are based on sentiment transfer across source and target domains, where knowledge obtained from labeled instances of source domain are transferred to target domain. A Cross-domain sentiment classification model [Fig 2] utilizes labeled data from source domain as well as unlabeled data from target domain to perform sentiment transfer

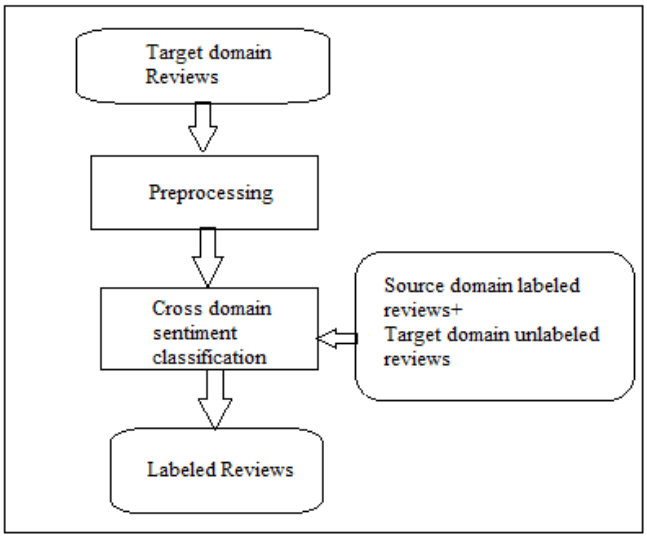

Fig 2 General Cross-Domain Sentiment Classification Model

Aue and Gamon [19] surveyed four different approaches to adapt a sentiment classification system to a new target domain in the absence of large amounts of labeled data. In the first approach, they trained a single classifier using equal amounts of training data from each of the domains where labeled data are available. This all-purpose classifier, being trained on multiple domains, will be less domain-specific than a classifier that has only seen data from one domain. A modification to this approach is to limit the features used during training to those that appear in the target domain. The third approach suggested using an ensemble of classifiers. The fourth approach exploits large amounts of unlabeled data available in the target domain. In this approach, small amounts of labeled target domain data were combined with large amounts of unlabeled data from the same domain in order to learn the model parameters for a generative classifier. The fourth approach provided the best classification accuracy of the four, since it is able to take advantage of unlabeled data in the target domain.

Blitzer et al [4] used structural correspondence learning algorithm for domain adaptation [16] in sentiment classification. In SCL, pivot features are selected based on their common frequency in both domains and point wise mutual information. Pivot features are used as the link for sentiment transfer between source and target domains. SCL is based on a multitask learning algorithm, alternating structural optimization [5]. SCL models the relationship between pivot features and non-pivot features by constructing a set of related tasks. Hence, the transfer ability of SCL for cross-domain sentiment classification is limited by the number of related tasks that could be constructed.

Pan et al [7] proposed spectral feature alignment that exploits domain independent features to construct a bipartite graph. The bipartite graph is used to model the co occurrence relationship between domain independent and domain specific features. Feature clusters are created by co-aligning domain independent and domain specific features. The clusters can be used to reduce the mismatch between domain specific words of both domains. The feature vectors augmented with elements from clusters are used to train a classifier to predict sentiment labels of target domain documents.

Tan et al [8] proposed MIEA method in which iterative reinforcement of sentiment scores is performed to determine the sentiment polarity of documents. Here sentiment transfer is achieved by exploiting all possible relationships between documents and words in both source and target domains. These relationships are modeled as graphs. The documents and words in the graph are assigned sentiment scores using graph-ranking algorithms. Based on the graphs iterative reinforcement approach computes the sentiment scores for the unlabeled documents based on which their sentiment class is identified.

Wu and Tan [9] proposed a two-stage framework for cross-domain sentiment classification. At the first stage, they tried to build a bridge between the source domain and the target domain. Graph ranking algorithm [17] was used for this purpose, which assigned a sentiment score to the documents. The sentiment scores were utilized to choose a set of most confidently labeled documents as seeds. At the second stage, they used the intrinsic structure revealed by the seed documents to calculate the sentiment score of each document. They employ the manifold-ranking algorithm [18] to spread the seeds ranking scores to their nearby neighbors to compute the ranking score for every unlabeled document. The target domain documents were then labeled based on these scores.

An active learning approach is proposed in [22], which performs cross-domain sentiment classification by actively selecting a small amount of labeled data in the target domain. In this approach, initially two individual classifiers are trained on the labeled data from the source and target respectively. The classifiers then use the active learning strategy called Query by Committee [20] to select a small amount of informative samples. A label propagation (LP) algorithm [21] is then applied to make the classification decision for a sample, both with the help of the source and target classifiers. LP is a graph based ranking algorithm, which propagates the labels from the labeled data to the unlabeled data.

Bollegala et al [23] used a semi-supervised approach that uses multiple source domains for sentiment classification. They proposed a method that uses a sentiment sensitive thesaurus coupled with feature expansion. The thesaurus captures related sentiment words for every entry base on a relatedness measure. During feature expansion, the thesaurus is used to expand feature vectors for the unlabeled

target domain documents. The candidates for feature expansion are selected using a ranking score. The expanded vectors where then given as training data to classifier.

Most of the methods to cross-domain sentiment classification utilize a combination of labeled data from source domains and unlabeled data from source and target domains. These methods try to bridge the gap between the source and target domains in order to determine the sentiment labels of the documents in target domain.

IV. CONCLUSION

Cross-domain sentiment classification is the task of classifying sentiment documents in a target domain using labeled data from a different domain. Major challenge in cross-domain sentiment classification is that the sentiment is expressed using different words across different domains. This paper tried to identify the challenges in cross-domain sentiment classification and discussed the existing methods for cross-domain sentiment classification.

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)*, 2002, pp. 79-86.

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

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

[4] A. Aue and M. Gamon, "Customizing sentiment classifiers to new domains: a case study.", *Proceedings of International Conference on Recent Advances in Natural Language Processing*, 2005, pp. 207-218.

[5] J. Blitzer, M. Dredze, and F. Pereira, "Biographies, Bollywood, Boom-Boxes and Blenders: Domain Adaptation for Sentiment Classification", *Proc. 45th Ann. Meeting of the Assoc. Computational Linguistics (ACL 07)*, 2007, pp. 440-447.

[6] R. K. Ando and T. Zhang. "A framework for learning predictive structures from multiple tasks and unlabeled data.", *Journal of Machine Learning Research*, 6,2005,1817–1853.

[7] S. Pan, X. Ni, J.T. Sun, Q. Yang, and Z. Chen.," Cross-domain sentiment classification via spectral feature alignment.", *In Proceeding of the 19th International World Wide Web Conference (WWW 2010)*, 2010, pp. 751-760.

[8] Q. Wu, S. Tan, X. Cheng, and M. Duan," MIEA: a mutual iterative enhancement approach for cross-domain sentiment classification." *Proceedings of the 23rd International Conference on Computational Linguistics*, 2010, pp. 1327-1335.

[9] Q. Wu and S. B. Tan, "A two-stage framework for cross-domain sentiment classification", *Expert Systems with Applications*, vol. 38, Oct 2011, pp. 14269-14275,.

[10] Pang, B. and Lee, L, "A Sentimental Education: Sentiment Analysis using Subjectivity Summarization based on minimum cuts.", *In Proceedings of the Association for Computational Linguistics*, 2004, pp. 271-278.

[11] Turney, P.D., "Mining the Web for synonyms: PMI-IR versus LSA on TOEFL.", *Proceedings of the Twelfth European Conference on Machine Learning*, 2001, pp. 491-502.

[12] A. Agarwal, F. Biadsy and K. Mckeown,"Contextual Phrase-Level Polarity Analysis using Lexical Affect Scoring and Syntactic Ngrams.*", In Proceedings of ECACL 2009*, 2009, pp. 24-32.

[13] A.M Popescu, and O. Etzioni," Extracting product features and opinions from reviews. *", In Proceedings of Conference on Empirical Methods in Natural Language Processing*, 2005, pp. 339-346.

[14] Y. Wu, Q. Zhang, X. Huang, and L. Wu, " Phrase dependency parsing for opinion mining*.", In Proceedings of Conference on Empirical Methods in Natural Language Processing*, 2009, pp. 1533-1541.

[15] H. Yu and V. Hatzivassiloglou, "Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences.", *In EMNLP*,2003,pp. 129-136.

[16] J. Blitzer, R. McDonald, and F. Pereira, "Domain adaptation with structural correspondence learning*". In Empirical Methods in Natural Language Processing (EMNLP)*, 2006, pp. 120-128.

[17] Q. Wu, S. Tan, X. Cheng, "Graph ranking for sentiment transfer.*" In Proceedings of ACL*, 2009, pp. 317-320.

[18] D. Zhou, J. Weston, A. Gretton, O. Bousquet, B. Scholkopf, "Ranking on data manifolds*.", In Proceedings of NIPS*.,2003, pp. 169-176.

[19] Y. Freund, H. Seung, E. Shamir, and N. Tishby, "Prediction, and Query by Committee.", *In Proceeding of NIPS-92n*, 1992, pp.483-490.

[20] X. Zhu and Z. Ghahramani, "Learning from Labeled and Unlabeled Data with Label Propagation.", *CMU CALD Technical Report*,2002.

[21] S. Li, S. Ju, G. Zhou, X. Li, "Active learning for imbalanced sentiment classification." In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2012, pp. 139-148.

[22] D. Bollegala, D. Weir, J. Carroll, "Cross-Domain Sentiment Classification Using a Sentiment Sensitive Thesaurus", *IEEE Transactions on Knowledge and Data Engineering*, 2012, pp. 1719 - 1731