ISSN No. 0976-5697

Volume 8, No. 5, May-June 2017

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

A Comparative Study on Cross Domain Sentiment Classification

Vrindha Aravind
PG Student
Department of Computer Science and Engineering
FISAT Engineering College
Ernakulum Kerala, India

Jyothish K John
Assistant Professor
Department of Computer Science and Engineering
FISAT Engineering College
Ernakulum, Kerala, 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.

Keywords: Cross Domain Sentiment classification, SCL, SFA, Sentiment Sensitive Thesaurus, Sentiment Embedding.

I. INTRODUCTION

As the influence of internet in our daily life increased, people began expressing their needs and opinions in public media which transform World Wide Web to a more creative and participative space. The large amount of information available in the internet is a valuable source of information. As stated by Pang *et al*. [1], the users started expressing their opinion about product and services in social media, shopping sites, blogs etc. This helps other customers to choose their desired products and services. But this data is huge in size. It is impractical to read all those data and take decision. Thus the need for sentiment analysis arises, which extracts all these data and produce summarised result. An automatic sentiment classifier classifies these reviews into positive and negative, based on the sentiments of the expressed words. According to Fan *et al*. [2], the main idea behind the concept of sentiment analysis is to determine the attitude of customers about some product or services and thus generate an overall polarity of the document or text.

Sentiment analysis, otherwise known as opinion mining is the task of determining the attitude and emotions behind the content. As the interest in the area increased, the amount of user generated data in the web also increased. In the paper, A Survey on Sentiment Analysis Algorithms for Opinion Mining [3], the task of sentiment analysis can be classified based on three levels:

i) Document level- At document based level, only the overall sentiments of the document is considered.
ii) Sentence level- At sentence level, the task is to determine whether the sentence is positive negative or neutral.
iii) Aspect level- But at the aspect level, the task directly considers each opinionated word [4].

According to the paper, A Survey on Cross Domain Sentiment Classification Techniques [5], the process of sentiment classification deals with categorising each word in the document to a pre-defined set of sentiment classes. Supervised binary sentiment classification deals with manually labelling each word on reviews with positive or negative labelled tags. But it is infeasible to manually annotate each word in every product document. Thus a sentiment classifier is developed which is trained using labelled reviews for one product, to classify the sentiments of a different product. An unsupervised sentiment classification deals with evaluating a document with pre-existing knowledge.

Sentiment analysis uses feature vectors, which is a collection of words particular to a domain. But each sentiment hold different meaning in different domain and it is expensive to annotate data in different domain. Thus, **Cross Domain Sentiment Classification** is considered a solution to this problem. As explained in the paper, A Survey paper on Cross Domain Sentiment Analysis [6], a classifier trained in one domain may not work well in other domain. Due to the mismatch in domain specific words, methods such as feature extraction, finding relatedness among words in both the domains, are to be performed prior to applying trained classifier on target domain.

The main task in cross domain sentiment analysis is to train a classifier in one or more domains and to apply it in a previously unknown target domain. To achieve this, according to Blitzer *et al*. [7], one should first identify the source domain features that are related to the target domain features. There are different techniques to perform this cross domain sentiment classification which yields a classification algorithm which perform equally accurate in target domain as in source domain.

II. CROSS DOMAIN SENTIMENT CLASSIFICATION TECHNIQUES

A. *NO ADAPTATION METHOD*

According to the author Bollegala *et al*. [11], the No-Adaptation method simply trains a binary classifier using

unigrams and bigrams from the labelled documents of source domain as features and do not include any feature expansion step. Then it applies the trained classifier to the target domain. This method does not include any adaptation and hence considered a lower bound.

Table 1: Overview on different techniques of cross domain sentiment classification

Technique	Features	Advantages	Disadvantages	Accuracy (%)
No Adaptation Method	Trains a classifier from unigrams and bigrams	Easy to build	Low accuracy	72.61
SCL	Calculates correlation between pivot features.	Weight vectors arranged in the form of matrices and hence easy to map features	Performance depends on pivots	78.83
SFA	Constructs bipartite graph to represent correlation	Better accuracy	Time consuming	81.48
Sentiment Thesaurus	A dictionary containing sentiment features is constructed	Thesaurus contains words from multiple domain	Thesaurus creation time and effort	83.63
Sentiment Sensitive Embedding	Classifier trained from lower dimensional embedded space	Uses three rules to obtain greater accuracy	Comparatively difficult to implement	85.86

B. *STRUCTURAL CORRESPONDENCE LEARNING*

Structural correspondence learning method selects features from different domain and then calculates the correlation between these pivot features. Blitzer *et al*. [8] proposed that this method involves both source domain and target domain. Pivot features are the features that occur frequently in both source and target domain and these features are selected by calculating mutual information between a feature and the domain label. SCL technique can be applied to feature based classification. Unlabelled data are obtained from both source and target domain whereas labelled data set is obtained only for source domain. This method is known as supervised learning. The first step in SCL is to select a set of pivot features from the unlabelled data of both domain. The algorithm uses a weight vector which contains values of the correlation between pivot and other features. Positive values in weight vector shows that non-pivots are highly associated with corresponding pivots. Linear classifiers are used to predict the existence of the features. These weight vectors are arranged in the form of matrices. From these feature matrix a mapping is made from the original feature space of both domain to a lower dimensional feature space by performing singular value decomposition. From these lower dimensional matrix, a binary sentiment classifier is trained.

The performance of the SCL algorithm depends on the selection of pivot features. This is the reason why most commonly occurring word from both domains are selected are pivots. The pivots should predict the source labels by calculating the mutual information between the features.

C. *SPECTRAL FEATURE ALIGNMENT*

Spectral Feature Alignment algorithm is an approach for cross domain sentiment classification. For this method, a set of labelled data is collected for the source domain. In order to train a classifier on the target domain, a set of unlabelled data is collected for the target domain. In general, Spectral Feature Alignment algorithm, proposed by Pany *et al.* [9], aims at reducing the gap between the source domain and the target domain. To represent the co-occurrence relationship between both domain specific and domain independent words, SFA constructs a bigraph. A spectral clustering algorithm is adapted based on graph theory, as proposed in the paper

Spectral Graph Theory [10] to achieve co-alignment between the domain-specific and domain-independent words.

The main idea behind this method is that, if two domain specific words have connection to a common domain independent word, then these words will be aligned with high probability. The same rule applies in the reverse case also. If two domain independent words have connection to a domain specific word, then they too will be aligned together with high probability. So, to align these domain specific and domain independent words into a feature cluster, a clustering algorithm is constructed which used the principles of graph theory. By co-aligning the features on bipartite graph, SFA fully exploits the relationship between domain-specific words and domain-independent words. Comparing to Structural Correspondence Learning, Spectral Feature Alignment algorithm is proved to maintain better accuracy.

D. *SENTIMENT SENSITIVE THESAURUS*

When applying a classifier trained in one domain, to an unknown domain, the problem arising much often is that the reviews in the target domain may be unknown to the training model. To overcome this issue, the sentiment sensitive thesaurus is created, which contain more related words from different domains.

To construct a sentiment sensitive thesaurus, Bollegala *et al.* [11] proposes that both labelled and unlabelled reviews are collected from different domains and POS tagging is applied. From each POS tagged reviews, words are normalised to form its lemma with a text classification process called Lemmatization. Function words are filtered out from these lemmas, selecting only the nouns, verbs, adverbs and adjectives. From the filtered reviews, lexical elements are created and then each lexical elements are

appended with its corresponding sentiment labels. Each lexical element is represented with its feature vector.

Point wise mutual information between each lexical element u and its feature vector w, is obtained using formula,

$$f(u,w) = \log\left\{\frac{\frac{c(u,w)}{N}}{\frac{\sum_{i=1}^{n} c(i,w)}{N} \times \frac{\sum_{j=1}^{m} c(u,j)}{N}}\right\}$$

C (u,w) denotes number of reviews containing both u and w
n denote the total number of lexical elements
m denote total number of feature vectors.

Point wise mutual information is useful for similarity measurement, word classification etc. The relatedness between two lexical elements are computed which is needed to construct sentiment sensitive thesaurus.

P.Sanju *et al*. [12], creates an enhanced sentiment sensitive thesaurus that not only aligns features from different domains, but also from wikitionary. Here, in addition to calculating the mutual information, semantically similar features are also collected with the help of java wikitionary library and are added to the already created thesaurus.

E. *SENTIMENT SENSITIVE EMBEDDINGS*

In this method, both source and target domain features are projected into a lower dimensional space and a sentiment classifier is trained from this embedded feature space. This feature is only applicable when there is a small overlap between source and target feature space.

Bollegala et al [13], selected a set of pivots (say M) from the given pair of domains. A d-dimensional feature vector is used to represent *i*th pivot feature in domain A and an h-dimensional feature vector is used to represent the same in domain B, thereby resulting in an M*d and M*h feature matrices respectively. A k-dimensional embedded space is generated to define, the relationship between two domains and also the distribution of pivots in two domains.

For this purpose, three rules are generated: (1) same pivots form different domains should be mapped as close as possible in the embedded space; (2) friend closeness and enemy dispersion has to be enhanced for source domain in the embedded space; (3) local geometry between documents in same domain must be preserved.

The embedded space maintains the connection between documents within each domain and also arranges the source and target domains according to pivot features. This helps in achieving unsupervised sentiment classification in target domain with the help of supervised information from source domain.

IV. CONCLUSION

In this paper, a small survey on four basic methods used for cross domain sentiment classification is presented. All these methods are different from one another in all means from feature expansion mechanism used, to the final classifier trained for classification. Structural Correspondence learning generates feature matrix from which a classifier is trained. Spectral Feature Alignment algorithm uses spectral graph to train a classifier based on graph theory. Sentiment sensitive thesaurus creates a sentiment feature dictionary. As an extension of SST method, an Enhanced Sentiment Sensitive Thesaurus is created which uses features of java wikitionary to the created thesaurus. Sentiment Sensitive Embedding trains a classifier by projecting words and documents to a lower dimensional space.

V. REFERENCE

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

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

[3] Vidisha M. Pradhan, Jay Vala, Prem Balani, A Survey on Sentiment Analysis Algorithms for Opinion Mining, International Journal of Computer Applications, 133,(2016).

[4] Sebastian Ruder, Parsa Ghaffari, John G. Breslin. Deep Learning for Multilingual Aspect-based Sentiment Analysis, Proceedings of SemEval-2016, (2016)330-336.

[5] Kinnari Ajmera et al,"A Survey on Cross Domain Sentiment Classification Techniques" International Journal of Computer Science and Information Technologies, Vol. 6 (6) , 2015, 5170-5172

[6] Pravin Jambhulkar, Smita Nirkhi, “A Survey Paper on Cross-Domain Sentiment Analysis”, International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 1, January 2014

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

[8] J. Blitzer, R. McDonald, and F. Pereira, “Domain adaptation with Structural Correspondence learning,” in EMNLP, 2006, pp. 120 –128.

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

[10] F. R. K. Chung. Spectral Graph Theory. Number 92 in CBMS Regional Conference Series in Mathematics. American Mathematical Society, 1997

[11] 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

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

[13] Danushka Bollegala,Tingting Mu,John Y. Goulermas. “Cross domain Sentiment Classification using Sentiment Sensitive Embeddings” IEEE Transactions on knowledge and data engineering, Volume: 28, Issue :2, Feb. 12016.