ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 1, January 2014

A Survey Paper on Cross-Domain Sentiment Analysis

Pravin Jambhulkar[1], Smita Nirkhi[2]

M.tech. Scholar, Dept. of Computer Science & Engg, RCOEM, Nagpur, India [1]

Asst. Prof., Dept. of Computer Science & Engg, RCOEM, Nagpur, India [2]

Abstract: The Internet contains important information on its user's opinions and sentiments. The extraction of those unstructured data is known as opinion mining and sentiment analysis. Basically Sentiment classification aims to automatically predict sentiment polarity (e.g., positive or negative) of users publishing sentiment data (e.g., reviews, blogs). This paper presents a short survey on cross-domain sentiment analysis techniques suggested and implemented recently. These techniques are then compared on the basis of feature expansion, number of source domain used, labeled and unlabeled data etc. These are then summarised and analysed. General challenges for performing cross-domain sentiment analysis are also discussed.

Keywords: cross-domain sentiment classification, support vector machine, domain adaptation, SentiWordNet.

I. INTRODUCTION

The recent transformation of the World Wide Web into a more participative and co-creative, users express their opinions about products or services they consume in blogs posts, shopping sites, or review sites like about book (amazon.com), automobiles (caranddriver.com), movies (imdb.com), hotels (tripadvisor.com), restaurants (yelp.com) etc. The tremendous amount of information available on these sites is now a valuable knowledge source. For example a person looking for a restaurant in a particular city may see the reviews of available restaurants in that city while taking a decision to select one of them. But this data is huge in size, so it is not possible for one to manually read all those data, hence sentiment analysis is used to extract this data and produce a summarised result. Basically sentiment analysis is to classify the polarity of text in document or sentence whether the opinion expressed is positive, negative or neutral. Polarity of all reviews is aggregated to obtain an overall opinion toward the given object.

Sentiment analysis performs on specific domain to achieve higher level of accuracy. The feature vector used in sentiment analysis contains a bag of words which are limited and should be specific to particular domain (domain can be consider as book, hotel, electronics etc.). However sentiment 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. Hence the solution can be to perform cross domain sentiment analysis 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. Cross-domain classification is nothing but to make a sentiment analysis from domain specific to generalise. A different technique gives different analysis, result and accuracy which depend on the documents, domain taken into consideration for classification.

Labeled / Unlabeled data

Given a specific domain D, the sentiment data xi and a yi denoting the polarity of xi, xi is said to be *positive* if the overall sentiment expressed in xi is positive ($yi = +1$), while xi is *negative* if the overall sentiment expressed in xi is negative ($yi = -1$). A pair of sentiment text and its corresponding sentiment polarity $\{xi, yi\}$ is called the *labeled sentiment data*. If xi has no polarity assigned, it is *unlabeled sentiment data*. Besides positive and negative sentiment, there are also neutral and mixed sentiment data in practical applications. *Mixed polarity* means user sentiment is positive in some aspects but negative in other ones. *Neutral polarity* means that there is no sentiment expressed by users.

Given two specific domains D_{src} and D_{tar}, where D_{src} and D_{tar} are referred to as a source domain and a target domain respectively, The set of labeled instances from the source domain, contains pairs of (t, s) where a review, t, is assigned a sentiment label, s. Here, s belongs to {+1, -1}, and the sentiment labels +1 and -1, respectively, denote positive and negative sentiments.

This paper primarily presents a comprehensive evaluate account of performance of all the available techniques for cross-domain classification.

II. SENTIMENT SENSITIVE THESAURUS

In 2013, Danushka Bollegala et al. [1] 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

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 1, January 2014

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.

A. Relatedness Measure

The procedure to construct sentiment sensitive thesaurus is, to first split the review into individual sentences and construct part-of-speech tagging and lemmatization using the RASP system [15]. Then apply a simple word filter based on POS tags to filter out function words, which only keep nouns, verbs, adjectives, and adverbs, because these are identified as good indicators of sentiment. It considers reviews as a bag of words and then select unigrams and bigrams from each sentence. Next from each source domain labeled review create sentiment elements by appending the label of the review to each lexical element.

They represent a lexical or sentiment element *u* as a feature vector ***u***, in which all lexical or sentiment element *w* that co-occurs with *u* in a review sentence contributes a feature in *u*. However, the value of the *w* in vector *u* is denoted by *f (u, w)*. The vector *u* can be seen as a compact representation of the distribution of an element *u* over the set of elements that co-occur with *u* in the reviews. In the distributional hypothesis, it is state that features that have similar distributions are semantically similar words [16].

They compute *f (u, w)* as the point wise mutual information

between a sentiment *u* and a feature *w* as follows:

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

Here, *c(u, w)* denotes the number of reviews in which a lexical element *u* and a feature *w* co-occur, *n* and *m*, respectively, denote the total number of lexical elements and the total number of features.

Next, for two lexical or sentiment elements *u* and *v* (represented by feature vectors *u* and *v*, respectively), they compute the relatedness *T (u, w)* of the element *v* to the element *u* as follows:

$$T(v,u) = \frac{\sum_{w\in\{x|f(v,x)>0\}} f(u,w)}{\sum_{w\in\{x|f(u,x)>0\}} f(u,w)}$$

The relatedness score *T (u, w)* can be interpreted as the proportion of pmi-weighted features of the element *u* that are shared with element *v*.

They use the relatedness measure to construct a sentiment sensitive thesaurus in which, for each lexical element *u* it list

up lexical elements v that co-occurs with *v* in the descending order of the relatedness values *T (u, v)*. For example, for the word excellent the sentiment sensitive thesaurus would list awesome and delicious as related words. In this co-occurrences are computed over both lexical and sentiment elements extracted from reviews. To construct the sentiment sensitive thesaurus, it must compute pair wise relatedness values for numerous lexical elements. Moreover, to compute the point wise mutual information values in feature vectors, it must store the co-occurrence information between numerous lexical and sentiment elements. Sparse matrix format and approximate vector similarity computation techniques [21] is used to create a thesaurus from a large set of reviews. They avoid computing relatedness values between lexical elements that are likely to have very small relatedness scores with the help of approximate vector similarity computation techniques thus they are unlikely to become neighbors of a given base entry.

B. Feature expansion

They propose feature expansion method to overcome the fundamental problem of cross-domain sentiment classification that words that comes in the source domains do not always appear in the target domain. Hence in feature expansion method they append a feature vector with additional related features selected from the sentiment-sensitive thesaurus.

In this method, they model a review d using the set *{w1 . . . ; wn}*, where the elements *wi* are either unigrams or bigrams that appear in the review *d*. Then it represent a review *d* by a real-valued term frequency vector *d*, where the value of the *jth* element *dj* is set to the total number of occurrences of the unigram or bigram *wj* in the review *d*. It defines a ranking score *(ui, d)* for each base entry in the thesaurus which will find the particular candidates to expand a vector *d* for the review *d*, as follows:

$$score(ui,d) = \frac{\sum_{j=1}^{N} djT(wj,ui)}{\sum_{l=1}^{N} dl}$$

According to this definition, given a review d, a base entry *ui* will have a high ranking score if there are many words *wj* in the review *d* that are also listed as neighbors for the base entry *ui* in the sentiment-sensitive thesaurus. To expand a vector, d, for a review *d*, they first rank the base entries, *ui* using the ranking score and select the top k ranked base entries. Then it extend the original set of unigrams and bigrams by the base entries create a new vector d for a review d. The values of the extended vector d are set as follows:

The values of the first N dimensions that correspond to unigrams and bigrams *wi* that occur in the review d are set to di, their frequency in d. The subsequent k dimensions that correspond to the top ranked base entries for the review d, and they are assign a weight according to their ranking score. However, both relatedness scores as well as normalized term-frequencies should be small, which leads to very small absolute ranking scores. While the expanded features must have lower feature values compared to that of the original features in particular feature vector. They

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 1, January 2014

set the feature values for the original features to their frequency in a review. This expanded feature vector is now more accurately classify the target domain without facing the mismatch problem of unseen words.

C. Multiple Sources

Traditionally, in cross-domain classification single source classifier taken into consideration to classify target domain, but Danushka Bollegala et al. [1] uses multiple source domain to trained classifier for target domain which improves the accuracy. But while using multiple source domains, size of the data or words is limited hence when uses single domain it will select 1000 positive and 1000 negative words (say) while using two source domains, it will select 500 positive and 500 negative from each domain and so on.

III. SPECTRAL FEATURE ALIGNMENT

Spectral feature alignment (SFA) method is first proposed by Pan et al. [2] 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. Between a domain-specific and a domain independent feature in the graph an edge is formed if those two features co-occur in some feature vector. After that, spectral clustering is conducted to identify feature clusters. Finally, a binary classifier is trained using the feature clusters for classification of positive and negative sentiment.

A. Domain-Independent Feature Selection

SFA uses some domain-independent words which help to construct a bipartite graph that model the co-occurrence relationship between domain-specific words and domain-independent words. There are several strategies available for selecting domain-independent features.

A first strategy is to select domain-independent features based on their frequency in both domains, i.e. it select those features that occur more than k times in both source and target domain. A second strategy is based on the mutual dependence between features and labels on the source domain data. Third strategy motivated by supervised feature selection criteria, it can use mutual information which measure any correlation or dependency between features and domains. The feature is domain specific if a feature has high mutual information, otherwise it is domain independent.

B. Bipartite Feature Graph Construction

With the help of above strategies, it can easily identify which features are domain-specific and which are domain-independent, and with this it constructs a bipartite graph G between them. Bipartite graph is used to model the intrinsic relationship between domain-specific and domain-independent features. In this graph vertices represent the domain-specific *(Vds)* and domain-independent *(Vdi)* words. The edge E connects two vertices in *Vds* and *Vdi*. Each edge is associated with weight *mij*. The score of mij measures the relationship between words in source and target domain.

Besides using the co-occurrence frequency of words within documents, they also adopt more meaningful methods to estimate m_{ij}. They define a reasonable "window size". If a domain-specific word and a domain-independent word co-occur within the "window size", then there is an edge connecting them. Furthermore, they also use the distance between *wi* and *wj* to adjust the score of m_{ij}. The weight it assigns is larger to the corresponding edge if their distance is small.

C. Spectral Feature Clustering

Based on the graph spectral theory [9], they assume if two domain-specific features are connected to many common domain-independent features, then they tend to be very related and will be aligned to a same cluster with high probability, and if two domain-independent features are connected to many common domain-specific features, then they tend to be very related and will be aligned to a same cluster with high probability.

Given the feature bipartite graph *G*, the goal is to learn a feature alignment mapping function: R^{m-l} -> R^k, where *m* is the number of all features, *l* is number of domain-independent features and *m - l* is the number of domain-specific features.

1. Form a weight matrix M belong to $R^{(m-1)*l}$, where M_{ij} corresponds to the co-occurrence relationship between a domain-specific word w_i belong to W_{DS} and a domain-independent word w_j belong to W_{DI} .
2. Form an affinity matrix A of the bipartite graph, where the first *m - l* rows and columns correspond to the *m - l* domain-specific features, and the last *l* rows and columns correspond to the *l* domain-independent features.
3. Form a diagonal matrix D, where $D_{ii} = \sum A_{ij}$, and construct the matrix $L = D^{-1/2} AD^{-1/2}$.
4. Find the *k* largest eigenvectors of L, $u_1, u_2, ..., u_k$, and form the matrix $U = [u_1u_2.....u_k]$ belong to R^{m*k}.
5. Define the feature alignment mapping function.

Though the goal is only to cluster domain-specific features, it is proved that clustering two related sets of points simultaneously can often get better results than only clustering one single set of points [19].

Algorithm 1: Spectral Domain-Specific Feature Alignment algorithm for Cross-Domain Sentiment Classification

Input: labeled source domain data $D_{src} = \{(x_{srci}, y_{srci})\}^{nsrc}$ for i=1, unlabeled target domain data $D_{tar} = \{x_{tarj}\}^{ntar}$ 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.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 1, January 2014

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

2. By using $\emptyset DI$ and $\emptyset DS$, calculate (Di-words)-(DS-word) co-occurrence matrix $M \in R^{(m-l)\times l}$

3. Construct matrix L = $D^{-1/2}AD^{-1/2}$, Where $A = \begin{bmatrix} 0 & M \\ M^T & 0 \end{bmatrix}$

4. Find the K largest eigenvectors of L, u1, u2, ..., uk, and form the matrix U = [u1 u2 ... uk] belongs to R^{m*k}. $Let\ mapping\ \varphi(xi) = xiU_{[1:m-l,:]}, where\ xi \in R^{m-l}$

Return a classifier f, trained on

$$\{([xsrci\ \gamma\varphi(\emptyset DS(xsrci))], ysrci)\}_{i=1}^{nsrc}$$

D. Feature Augmentation

A tradeoff parameter is used in this feature augmentation to balance the effect of original features and new features.

IV. STRUCTURAL CORRESPONDENCE LEARNING

SCL-MI. This is the structural correspondence learning (SCL) method proposed by Blitzer et al. [3]. 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.

A. Algorithm

In SCL, it uses labeled data from source domain and unlabeled data from source and target domain, then it chooses *m* pivot features which frequently occurs in both domains. It will then find some correlation between the pivot features and other features by training linear predictor. It will predict occurrences of each pivot in the unlabeled data from source and target domains. It uses a weight vector w_l, in which positive entries mean that a non-pivot features is highly correlated with the corresponding pivot. It arrange weight vector into a matrix W. At training and test time, they observe a feature vector x, then apply the projection X_x to obtain k new real-valued features. Now it learns a predictor for the augmented instance (x, X_x). If X contains meaningful correspondence, then the predictor which uses X will perform well in both domains.

B. Selecting Pivots with Mutual Information

The performance of SCL depends on the selection of pivots; hence the frequently occurring words in both domains can be a good choice, as they are good indicators of part of speech. In SCL, pivots feature needs to be a good predictor of the source label, so they choose those features that have highest mutual information to the source label.

V. CONCLUSION

In this paper, we present a short survey on various techniques used for cross-domain sentiment analysis. We have mainly discussed three techniques viz sentiment sensitive thesaurus, spectral feature alignment, structural correspondence learning. All these three techniques are different from one other in the way of expanding the feature vector, measuring the relatedness among the words, and finally the classier used for classification. Some methods used for performing cross-domain classification uses labeled or unlabeled data or some uses both. Hence the technique used gives different result for different domain as well as purposes. We discussed pros and cons of each methods used and the challenges faced for cross-domain sentiment analysis.

ACKNOWLEDGMENT

I would like to thanks my respected project guide Prof. Smita Nirkhi who has put her careful guidance and timely advice which inspired me towards sustained effort for my concerned work.

Special thanks to Prof. Wanjari Co-ordinator M.tech CSE for his timely guidance.

One more personality, who is ever ready to listen to the difficulties of students and is always on his toes to help them out, I would like to thank our Head of Department Dr. M. B. Chandak and entire Computer Science and Engineering Department of RCOEM for their assistance.

REFERENCES

[1] 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.

[2] 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).

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

[4] Saeideh Shahheidari, Hai Dong, Md Nor Ridzuan Bin Daud, "Twitter sentiment mining: A multi domain analysis", 7th International Conference on Complex, Intelligent, and Software Intensive Systems.

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

[6] 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.

[7] 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.

[8] 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.

[9] M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177, Seattle, WA, USA, 2004. ACM.

[10] 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 Computational Linguistics (NAACL '09), pp. 281-289, 2009.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 1, January 2014

[11] Y. Lu, C. Zhai, and N. Sundaresan. Rated aspect summarization of short comments. In *Proceedings of the 18th international conference on World wide web*, pages 131–140, Madrid, Spain, 2009. ACM.
[12] T. Li, Y. Zhang, and V. Sindhwani. A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association of Computational Linguistics*, pages 244–252, Suntec, Singapore, August 2009. ACL.
[13] 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.
[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] T. Briscoe, J. Carroll, and R. Watson, "The Second Release of the RASP System," Proc. COLING/ACL Interactive Presentation Sessions Conf., 2006.
[16] Z. Harris, "Distributional Structure," Word, vol. 10, pp. 146-162, 1954.
[17] S. Sarawagi and A. Kirpal, "Efficient Set Joins on Similarity Predicates," Proc. ACM SIGMOD Int'l Conf. Management of Data, pp. 743-754, 2004.
[18] F. R. K. Chung. *Spectral Graph Theory*. Number 92 in CBMS Regional Conference Series in Mathematics. American Mathematical Society, 1997.
[19] I. S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the 7th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 269–274. ACM, 2001.
[20] 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.
[21] D. Lin, "Automatic Retrieval and Clustering of Similar Words," Proc. Ann. Meeting of the Assoc. Computational Linguistics (ACL '98), pp. 768-774, 1998.
[22] 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.
[23] 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.
[24] B. Chen, W. Lam, I. W. Tsang, and T.-L. Wong. Extracting discriminative concepts for domain adaptation in text mining. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 179–188, New York, NY, USA, 2009. ACM.
[25] J. Blitzer, R. McDonald, and F. Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of the Conference on Empirical Methods in Natural Language*, pages 120–128, Sydney, Australia, July 2006.
[26] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407, 1990.
[27] T.-K. Fan and C.-H. Chang, "Sentiment-Oriented Contextual Advertising," Knowledge and Information Systems, vol. 23, no. 3, pp. 321-344, 2010.
[28] Y. Lu, C. Zhai, and N. Sundaresan. Rated aspect summarization of short comments. In *Proceedings of the 18th international conference on World wide web*, pages 131–140, Madrid, Spain, 2009. ACM.
[29] B. Liu. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Springer, January 2007.
[30] H. Daumé III. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic, June 2007.
[31] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.