DOMAIN ADAPTABLE WEIGHTED SEMI-SUPERVISED LEARNING FOR OPINION CLASSIFICATION

PRANALI WAGH [1], JYOTI DESHMUKH [2], SWATI DESHPANDE [3]

[1] PG Student, Shah & Anchor Kutchhi Engineering College, Chembur, Mumbai

[2] Ass.Professor, Shah & Anchor Kutchhi Engineering College, Chembur, Mumbai

[3] Ass.Professor, Shah & Anchor Kutchhi Engineering College, Chembur, Mumbai

sakec.pranaliw@gmail.com1,sakec.jyotid@gmail.com2,sakec.swati@gmail.com3

ABSTRACT **:** *Due to the exponential increase in the availability of online reviews makes sentiment classification an interesting topic in academic and industrial research. Customer reviews are significant source of information resource which are useful for both potential customers and product manufacturers. Whenever we need to make a decision, we want to know other opinions. Businesses and organizations always want to find consumer or public opinions about their products and services. To gather annotated training data for different domain is difficult because reviews can span so many different domains. Same sentiment word can express different meaning in different domains and annotating corpora for every possible domain of interest is impractical.*

Automatic sentiment classification is the task of classifying a given review with respect to the sentiment expressed by the user. Sentiment is expressed differently in different domains therefore the sentiment classifier that is trained to classify opinion polarities in a domain may produce miserable results when the same classifier is used in another domain. Therefore, Domain Adaptation is needed. The proposed Weighted Semi-Supervised Learning(WSSL) Algorithm is developed in which a classifier is trained on labeled reviews from source domain & target domain. Clustering is used to reduce the gap between domain-specific words of the two domains by putting them into unified clusters with the help of domain-independent words as a bridge.

INTRODUCTION

The growing popularity & availability of reviews increasing more attentation towards the research of sentiment analysis which is also known as an opinion mining. Opinion mining is used to determine the polarity of a text such as positive, negative or neutral .Opinion represents the individual's ideas, judgements, assessments, beliefs about specific topic. It is the private state of an individual which has a great impact on and provide guidance to individuals, organization duing decision making process. There are three main components that constitute an opinion. These components are used for opinion identification. First, is the opinion holder or source of an opinion. An object on which opinion is expressed & finally the evaluation. Opinion Mining is the study of people's opinions, attitudes and emotions toward an entity. An entity can be an individuals, events or topics. These topics are most likely to be covered by reviews. Opinions about products or services are expressed by user who consume them in blog posts, shopping sites, or review sites. Reviews on a wide variety of commodities are available on the Web. Reviews are useful to know what general public think about a particular product or service for consumer as well as for producer. Sentiment classification is the task of classifying a given review with respect to the sentiment expressed by the author of the review. A sentiment classifier that is trained to classify opinion polarities in a domain may produce miserable results when the same classifier is used in another domain. Sentiment is expressed differently in different domains. For instance, consider two domains, digital camera and car. The way in which customers express their thoughts, views and prospective about digital camera will be different from those of cars. But some similarities may also be present. So Sentiment analysis is a problem which has high domain dependency. Therefore cross domain sentiment analysis is a challenging problem that has to be unfolded. It has been shown that sentiment classification is highly sensitive to the domain from which the training data is extracted. A classifier trained using opinion documents from one domain often performs poorly on test data from another domain. The reason is that words used in different

domains for expressing opinions can be quite different ie. the same word in one domain may mean positive but in another domain may mean negative. Thus, domain adaptation is needed.

RELATED WORK

Classifying product reviews is a common problem in opinion mining and varieties of techniques have been used to address the problem. These techniques can be classified into two main approaches, lexicon based approach and machine learning based approach.

Pang and Lee [1] presented survey on sentiment analysis and opinion mining. In that survey they explained opinion oriented information access, challenges, opinion classification and summarization. Mikalai Tsytsarau, Themis Palpanas [2] also have presented Survey on opinion mining. In that survey author explained opinion mining, opinion aggregation and subjectivity analysis. Their study mentioned different work performed on this issue and their comparisons.

Many researchers used machine learning methods for sentiment analysis [3] [4] that involve training of classifier on datasets and use the trained model for new document classification. Some authors suggested another method such as dictionary of word lexicons [5].

Structural correspondence learning (SCL) method proposed by Blitzer et al. [7]. This method utilizes both labeled and unlabeled data in the benchmark dataset. It selects pivots using the mutual information between a feature (unigrams or bigrams) and the domain label. Next, linear classifiers are learnt to predict the existence of those pivots. The learnt weight vectors are arranged as rows in a matrix and singular value decomposition 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.

Spectral feature alignment (SFA) method proposed by Pan et al. [8]. 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 domainindependent 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.

Danushka Bollegala et al. [9] proposed a method for cross domain sentiment classification in sentiment sensitive distributional thesaurus using labeled data for the source domains and unlabeled data for both source and target domains. Sentiment sensitivity is achieved in the thesaurus by incorporating document level sentiment labels in the context vectors used as the basis for measuring the distributional similarity between words, they created thesaurus to expand feature vectors during train and test times in a binary classifier.

PROBLEM SETTING

The WSSL Algorithm provides reliable method for constructing domain adaptable lexicon from the reviews collected from amazon dataset. Reviews often deal with various kinds of products or services for which vocabularies are different. A sentiment classifier that is trained to classify opinion polarities in a domain may produce miserable results when the same classifier is used in another domain. Sentiment is expressed differently in different domains. To deal with this, system creates a train model for which labeled and unlabeled data are available from the set of domains and then apply it to any target domain (labeled or unlabeled).

Domain D as a class of entities. For example, different types of products such as books, DVDs, cameras etc. are considered as different domains. Given a review written by a user on a product that belongs to a particular domain, the objective is to construct domain adaptable opinion lexicon such that the sentiment words can be used to express their polarity appropriately into their respective domain. 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 sentiment data in the source domain is represented by $L(Dsrc)$, and the set of unlabeled data in the target domain is represented by $U(Dtar)$.Model is proposed to construct domain adaptable lexicon for both source & target domain.

PROPOSED WORK

Figure 1. shows the general framework for constructing domain adaptable lexicon which consist of following phases.

Collection of Reviews

Reviews are collected from amazon dataset. This dataset consist of positive as well as negative review statements for four different categories of product like book, dvd, kitchen, electronics etc.

Data Pre-processing

In this stage it is required that the text which is collected that will be analyzed for detection of opinions. In this phase, pre-processing is done to eliminate unnecessary words called as stop words. This is important because the irrelevant data from the reviews could be eliminated. This eliminates the processing overheads of a large amount of textual data.

Most of the English sentences include words like “a, an, of, the, I, it, you, and etc”. Such words do not carry particular meaning. Information extraction from natural language can be done effectively and clearly by avoiding those words which occurs very often. To

remove stop words from sentences text file is used which consists of list of stop words.

Sentence Parsing

Sentence Parsing involves assigning different parts of speech tags such as noun, preposition, verb, adjective and adverbs to a given text are known as Part-Of-Speech tagging. It is a special application of natural language processing. The part-of-speech is a category used in linguistics that is defined by a syntactic or morphological behaviour of a word. The traditional English language grammar classifies parts-of-speech in the following categories: verb, noun, adjective, adverb, pronoun, preposition, conjunction and interjection. The reason why POS tagging is so important to information extraction is the fact that each category plays a specific role within a sentence. Nouns give names to objects, or entities from reviews. An adjective describes opinion. Also some adverbs can play important role as an adjective.

Examples:

the/DT battery/NN life/NN on/IN the/DT iphone/JJ 4S/CD is/VBZ amazing/JJ

It/PRP is/VBZ a/DT bit/RB expensive/JJ

Got/VBD some/DT problem/NN in/IN battery/NN

this/DT phone/NN is/VBZ very/RB slow/JJ

Firstly, text review is divided into sentences. Stanford parser [11] is used to generate the POS tagging of each word present in the sentence. It is very essential as it helps in finding general language patterns.

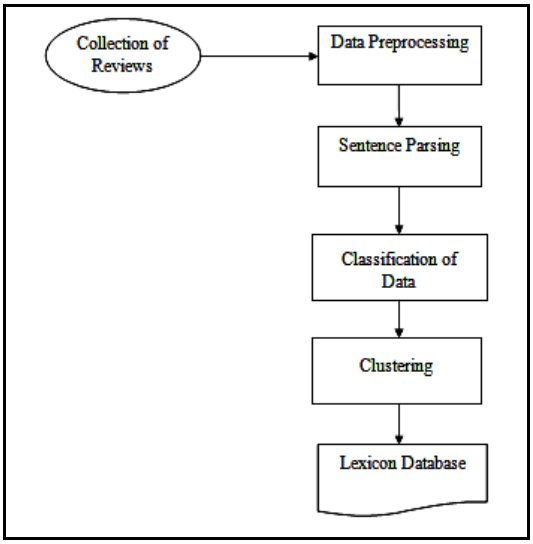

Figure 1.General framework for opinion lexicon construction

Classification of Data

Classification is the task of choosing the correct class label for a given input. In basic classification tasks, each input is considered in isolation from all other inputs, and the set of labels is defined in advance. Once the pos tagging is done maximum entropy classifier algorithm will be applied on those parsed words using which words can be classified as positive & negative [10]. Along with the classification polarity score is assigned to the classified words [9]. Once the words are classified into their respective class these words are then used to find domain-independent & domain-specific words from both the domain.

Steps for Maximum Entropy Classification algorithm is as follows.

Inputs: A collection D of labeled documents and a set of feature functions f_i

Set the constraints

$$P(c \mid d) = \frac{1}{z(d)} \exp\left(\sum_i \lambda_i f_i(d,c)\right) \quad \text{Eq. (1)}$$

For every feature f_i

- Initialize the λ_i's to be zero.
- Iterate until convergence:

Calculate the expected class labels for each document with the current parameters,

$$z_\lambda(d) = \sum_c \exp\left(\sum_i \lambda_i f_i(d,c)\right) \quad \text{Eq. (2)}$$

Calculate δ_i from the following eq.

$$\delta_i = \frac{1}{M} \log \frac{\sum_{d \in D} f_i(d, c(d))}{\sum_c p_\lambda(c \mid d) f_i(c,d)} \quad \text{Eq. (3)}$$

- Set $\lambda_i = \lambda_i + \delta_i$ Eq. (4)

Output: A text classifier that predicts a class label.

Weights of the words are calculated using following equation.

$$f(u,w) = \log\left(\frac{\frac{c(u,w)}{N}}{\frac{\sum_{i=1}^{n} c(i,w)}{N} * \frac{\sum_{j=1}^{m} c(u,j)}{N}} \right) \quad (5)$$

Where, $c(u,w)$ denotes the number of review sentences 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, and $N = \sum_{i=1}^{n} \sum_{j=1}^{m} c(i,j)$.

Clustering

Clustering is a division of data into groups of similar objects. Each group, called cluster, consists of objects

that are similar between themselves and dissimilar to objects of other groups.

To co-align domain-specific and domain-independent words into a set of feature-clusters clustering is used. Clusters can be used to reduce the mismatch between domain-specific words of both domains which is helpful for training an accurate classifier for the target domain.

Performance of the model is calculated using the following formula which is the proportion of instances whose class the classifier can correctly predict.

$$\text{Accuracy} = \left(TP + TN \Big/ TP + TN + FP + FN\right)$$

Experiments & Evaluation

We use the cross-domain sentiment classification dataset prepared by Blitzer et al. [7] to compare the proposed method against previous work on cross-domain sentiment classification. This dataset consists of Amazon product reviews for four different product types: books, DVDs, electronics and kitchen appliances. These reviews are written in xml format, in each domain, there are 1000 positive review and 1000 negative one. The dataset also contains some unlabeled reviews for the four domains. From this dataset, we can construct 12 cross-domain sentiment classification tasks: B → D, B→ E, B → K, D → B, D→ E, D → K, E → B, E→ D, E → K, K → B, K → D, K → E, where the word before an arrow corresponds with the source domain and the word after an arrow corresponds with the target domain.

Following figures shows comparison result between SFA algorithm & WSSL algorithm. In which one particular domain is used as a target domain & the other domains used as a source domain.

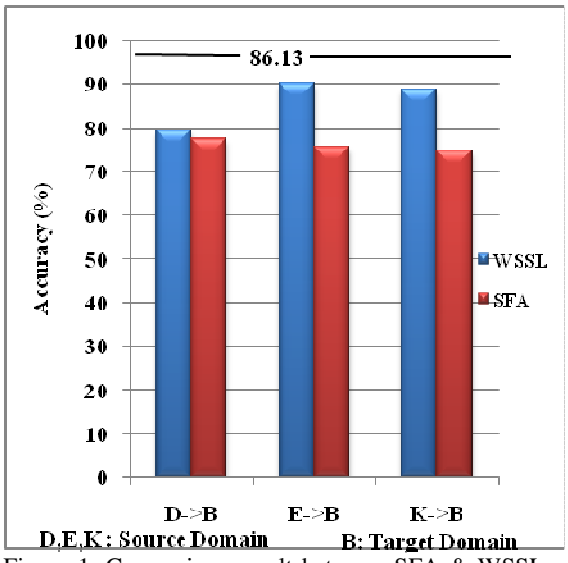

Figure 1. Comparison result between SFA & WSSL using Book (B) as a target & other domain as source domain.

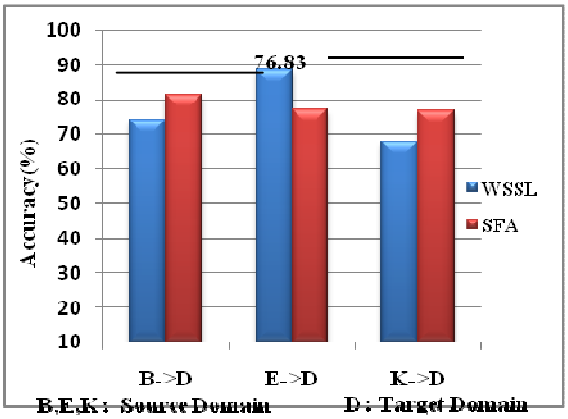

Figure 2. Comparison result between SFA & WSSL using Dvd (D) as a target & other domain as source domain.

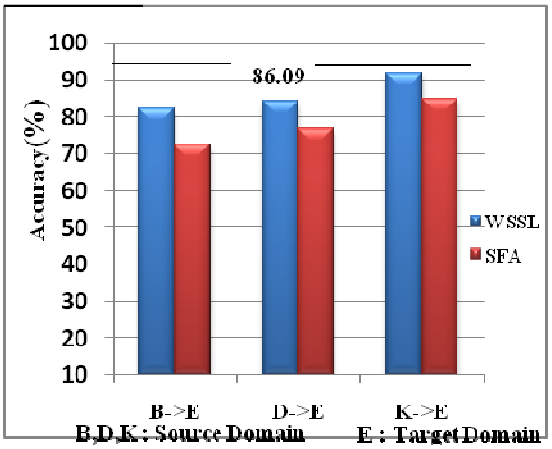

Figure 3. Comparison result between SFA & WSSL using Electronics (E) as a target & other domain as source domain.

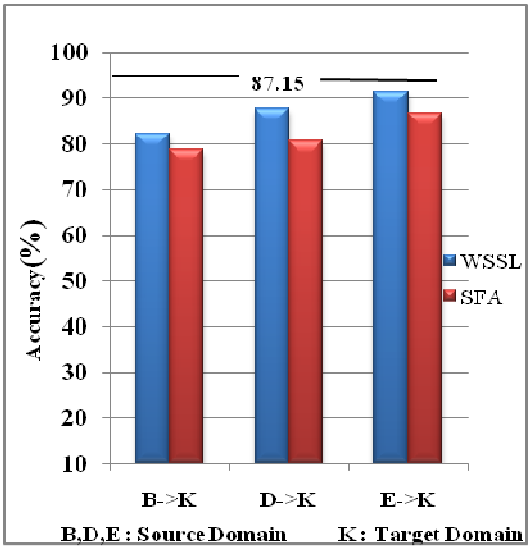

Figure 4. Comparison result between SFA & WSSL using Kitchen (K) as a target & other domain as source domain.

CONCLUSIONS & FUTURE WORK

From the above results it has been observed that book & dvd if considered as a source domain achieves a good compatibility with electronics & kitchen domain considered as target domain respectively,also by considering electronic & kitchen as a source domain displays better compatibility when adapting to kitchen & electronic target domain.

Accuracy achieved by SFA was between 72.5% to 86.75% whereas accuracy of proposed algorithm lies between 67.85% to 91.64%.

Currently system classifies reviews into positive & negative category, which can further be improved by classifying reviews into neutral category.

REFERENCES

[1] Bo Pang, Lillian Lee, "Opinion Mining and Sentiment Analysis", Foundations and Trends in Information Retrieval Vol. 2, Nos. 1–2 (2008).

[2] Mikalai Tsytsarau, Themis Palpanas "Survey on mining subjective data on the web", Data Mining Knowledge Discovery, Springer 2012, pp.478-514.

[3] Alvaro Ortigosa, José M. Martín, Rosa M. Carro, "Sentiment analysis in Facebook and its application to e-learning", Computers in Human Behavior Journal Elsevier 2013.

[4] Chien-Liang Liu, Wen-Hoar Hsaio, Chia-Hoang Lee, Gen-Chi Lu, and Emery Jou "Movie Rating and Review Summarization in Mobile Environment", IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews, Vol. 42, No. 3, May 2012, pp.397-406.

[5] Elena Lloret, Alexandra Balahur, José M. Gómez, Andrés Montoyo, Manuel Palomar, "Towards a unified framework for opinion retrieval, mining and summarization" Journal of Intelligent Information Systems Springer 2012, pp.711-747.

[6] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? sentiment classification using machine learning techniques," in EMNLP 2002, 2002, pp. 79–86.

[7] J. Blitzer, R. McDonald, and F.pereira, "Domain adaptation with structural correspondence learning," in EMNLP 2006.

[8] Sinno Jialin Pan Et Al." Cross-Domain Sentiment Classification Via Spectral Feature Alignment" *The International World Wide Web Conference Comimittee (IW3C2). WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA*.

[9] Bollegala, David Weir and John Carroll, "Cross-domain Sentiment Classification Using a Sentiment Sensitive Thesaurus", IEEE Transaction on Data and Knowledge Engineering, Vol. 25, No. 8, 2013.

[10] Kostas Fragos et al."A Weighted Maximum Entropy Language Model for Text Classification", *Proceedings of the 2nd International Workshop on Natural Language Understanding and Cognitive Science, NLUCS 2005, In conjunction with ICEIS 2005, Miami, FL, USA, May 2005.*

[11] http://nlp.stanford.edu:8080/parser.