Opinion mining using combinational approach for different domains

Jyoti Deshmukh[1], Amiya Kumar Tripathy[2], Dilendra Hiran[3]
[1]Department of Computer Engineering, PAHER University, India
[2]Department of Computer Engineering, University of Mumbai, India
[3]Department of Computer Applications, PAHER University, India

Article Info

Article history:

Received Sep 24, 2018
Revised Mar 27, 2019
Accepted Apr 8, 2019

Keywords:

Information systems
Knowledge discovery
Machine learning
Opinion mining
Text mining

ABSTRACT

An increase in use of web produces large content of information about products. Online reviews are used to make decision by peoples. Opinion mining is vast research area in which different types of reviews are analyzed. Several issues are existing in this area. Domain adaptation is emerging issue in opinion mining. Labling of data for every domain is time consuming and costly task. Hence the need arises for model that train one domain and applied it on other domain reducing cost aswell as time. This is called domain adaptation which is addressed in this paper. Using maximum entropy and clustering technique source domains data is trained. Trained data from source domain is applied on target data to labeling purpose A result shows moderate accuracy for 5 fold cross validation and combination of source domains for Blitzer et al (2007) multi domain product dataset.

Corresponding Author:

Jyoti Deshmukh,
Department of Computer Engineering,
PAHER University, Udaipur, Rajsthan, India.
Email: jyoja2007@gmail.com

1. INTRODUCTION

Opinion mining is emerging area of research as popularity & availability of reviews increasing. Opinion mining is used to determine the polarity of a text such as positive, negative or neutral. Opinion represents the individual's ideas, judgments, assessments, beliefs about specific topic. Extraction of opinion or sentiment is very important task in business and academic world. Every manufacturer wanted to know the review about their products [1]. Decision making affects because individuals are rely on what others are thinking. Therefore, sentiment analysis is popular stream, which extracts sentiments and analyze it [2-3].

Classification of documents depending on polarity is a key activity in opinion mining. Documents are written in positive, negative or in neutral orientation. Words polarity is impotant feature in opinion mining. As words are domain dependent, the knowledge of domain is required to find out whether word is positive or negative. Reusability of knowledge of one domain in another domain is key issue in opinion mining.

Doamin adaptation can be the solution for this problem in which knowledge can be transferred from one domain to another reducing cost in terms of time and money. Consider the problem, where task is to automatically classify the reviews on Electronics domain into positive and negative orientation. For this task, first we have to collect many reviews of the domain. Then train a classifier on the reviews with their subsequent labels. Large amount of reviews are needed to maintain good classification performance. Labeling reviews for each domain is time consuming as well as expensive process. Hence, domain adaptation need arises which can uses knowledge of one domain to another one [4]. Structural correspondence learning (SCL) proposed to extend structural learning. SCL defines pivot features, which are common to both source and target domain. This method tries to find the correlation between pivot features, and non-pivot features.

Whitehead et al., [5] proposed a method for building ensemble models, using lexicon similarity, that yield a high classification accuracy for domains in which no training was performed. It is reported that an adjusted form of cosine similarity between domain lexicons can be used to predict which models will be effective in a new target domain. Jialin et al., [6] proposed a general framework for cross-domain sentiment classification. Spectral feature alignment (SFA) creates meaningful clusters with the help of common words. Bipartite graph is constructed between common or domain independent and uncommon words of both source and target domains. Mutual information used to select common words and binary classifier is trained for classification. Experimental results shows effective performance of approach on both document and sentence level classification. Liu et al., [7] proposes a method for co-extracting opinion targets and opinion words by using a word alignment model. Main focus is on detecting opinion relations between opinion targets and opinion words. As compared to previous methods based on nearest neighbor rules and syntactic patterns, proposed method captures opinion relations more precisely. An Opinion Relation Graph is constructed to model all candidates along with a graph co-ranking algorithm to estimate the confidence of each candidate. The items with higher ranks are extracted out. The experimental results for three datasets with different languages and different sizes prove the effectiveness of the proposed method. In future work, authors planned to consider additional types of relations between words, such as topical relations, in Opinion Relation Graph. Balamurali A. R. et al., [8] proposes approach for cross domain sentiment tagging. A method for creating high in-domain classifier using simple low level features is introduced.

A generic classifier based on meta-classification approach coupled with this high in-domain classifier is used to create labeled data for a new domain from domains having labeled data. Results showed considerable improvement in cross domain sentiment tagging accuracy if domains are similar. In case of dissimilar domains system exceeds the baseline accuracies by substantial margins. Bollegala D. et al., [9] proposed a method creating thesaurus which is receptive to sentiment words from different domains. Author used both labeled and unlabeled data. Created lexicon vocabulary was expanded at train and test times in a classifier. Proposed method compared with many baseline methods which reveal a good performance. Shoushan et al., [10] proposed active learning in which source and target classifiers are trained separately. Using Query by Committee (QBC) selection strategy informative samples are selected and classification decision made by combining classifiers. Label propagation is used to train both classifiers. Result demonstrates significantly outperforms than the baseline methods. Like ensemble classifiers graph based methodology also used for domain adaptation. Inderjit S. et al., [11] proposed the graph based domain adaptation method. Similarity graph constructed between features from all domains, if these features are similar then edge exist between them. All labeled features used in metric-learning algorithms. Graph is constructed using data-dependent metric and weight is calculated for each edge. An experimental result demonstrates reduction in classification error. S. Bhatt et al., [12] proposes an algorithm to adapt classification model by iteratively learning domain specific features from the unlabeled test data. Moreover, this adaptation transpires in a similarity aware manner by integrating similarity between domains in the adaptation setting. Cross-domain classification experiments on different datasets, including a real world dataset, demonstrate efficacy of the proposed algorithm over state-of-the art.

Many open source lexicons are available which serve as a database for extracting the polarity values of opinion words. However, these generic polarity lexicons reveal the general sentiment of opinion words. An opinion word could be context dependent or domain specific. The word like “small” may represent a negative orientation in a hotel domain but if used in mobile applications it is a positive. Same way “freezing” is good for a refrigerator but negative for software applications. The variation of opinion possess by a same word in different domains restricts the usage of generic lexicons as it contains generalize polarity of a word. So need of domain adaptable lexicons emerges. Domain adaptability is major issue in sentiment analysis which has been addressed in proposed framework. A proposed approach attempts in building a classifier which uses maximum entropy classifier with clustering based on point wise mutual information between words.

2. PROPOSED METHOD

Opinion lexicon is a word or group of words in review. Identification of opinianated words or lexicons is an important task. In proposed method different tasks are discussed. Data collection is crucial task as large data is available online. For proposed approach Amazons multi product review dataset is used. After data collection cleaning of data is necessary. Stop word removal method is applied on data collected. Part of speech tagging is used to tag words as adjective, adverb. After this preprocessing step classifier is applied on cleaned data. Maximum entropy classifier is used alongwith clustering as shown in Figure 1.

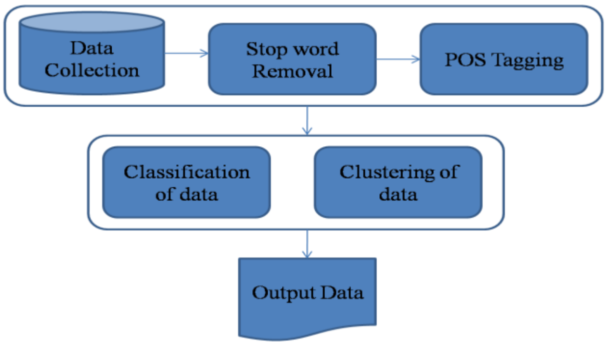


Figure 1. Flow of proposed method

Input: Source and target domains preprocessed review documents. For all features extracted from input dataset

1. Initialize $\lambda_i = 0$ for $i = 1$ to $n+1$ (where n is total number of features)
2. Repeat until convergence. Calculate the probability of class *c*

$$P(c|d,\lambda) \overset{def}{=} \frac{\exp\sum_i \lambda_i f_i(c,d)}{\sum_{c'\in C}\exp\sum_i \lambda_i f_i(c',d)} \ldots\ldots\ldots \tag{1}$$

where, $\lambda = \lambda_i + \delta_i$ (λ is calculated by Iterative scaling algorithm)

δ_i is increment quantity

$f_i(d,c) = \textit{tf-idf}$ value of each word considered as feature.

3. For all classified words weight is assigned using Point wise Mutual Information (PMI) with (2)

$$PMI(Word_1, Word_2) = \log_2\left(\frac{p(Word_1 \wedge Word_2)}{p(Word_1)\,p(Word_2)}\right) \tag{2}$$

4. Common and uncommon words from source and target domains are clustered using weight value.
Output: positive and negative classified documents and word clusters.
Iterative Scaling Algorithm:
Input: features functions $f_1, f_2, \ldots\ldots. f_n$, Empirical Distribution $\tilde{p}(c,d)$.

Output: Optimal parameter values, optimal model p^*.

1. Start with $\lambda_i = 0$ *for all* $i \in \{1,2,\ldots.,n\}$
2. Do for each $i \in \{1,2,\ldots.,n\}$
   a. Let $\Delta\lambda_i$ be the solution to $\sum_{c,d} \tilde{p}(c)p(d|c)f_i(c,d)\exp(\Delta\lambda_i f^{\#}(c,d) = \tilde{p}(f_i)$

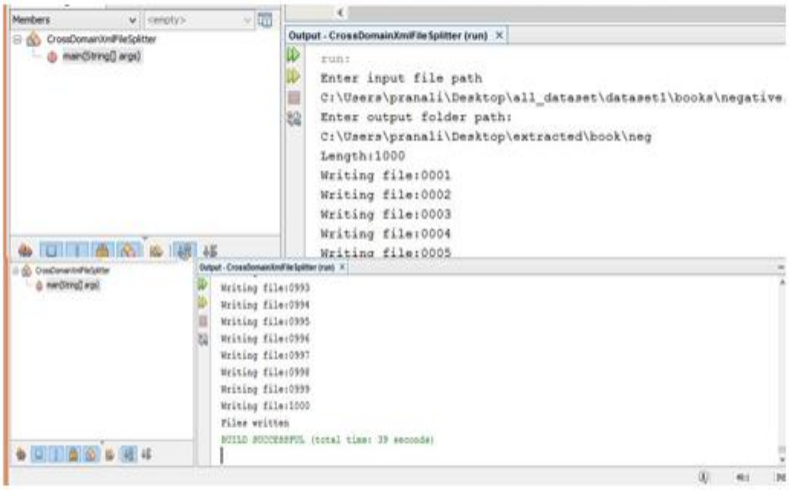
      where, $f^{\#}(c,d) = \sum_{i=1}^{c,d} f_i(c,d)$

   b. Update the value of λ_i according to $\lambda_i \leftarrow \lambda_i + \Delta\lambda_i$
3. Go to step 2 if not all the λ_i have converged. $f^{\#}(c,d)$ is the total number of features which are active for particular (c,d) pair. Iterative scaling algorithm is used to calculate λ value [13].

3. RESULTS AND ANALYSIS

Blitzer et al., [14] Multi Domain dataset is used for evaluation of proposed method. Results of each step recorded in following section. Preprocessed data is used for classification and clusteing process. In first experiment source domain is divided into 5 parts and target domain taken as it is. Second experiment done on combination of source domains.

The Multi-Domain Sentiment Dataset contains product reviews taken from Amazon.com from many product domains [14]. This dataset contains three types of files positive, negative and unlabeled in XML format. Each line in form of: feature:<count> .... feature:<count> #label#:<label>. These files are extracted using XML file splitter and reviews are written into text file as shown in Figure 2.

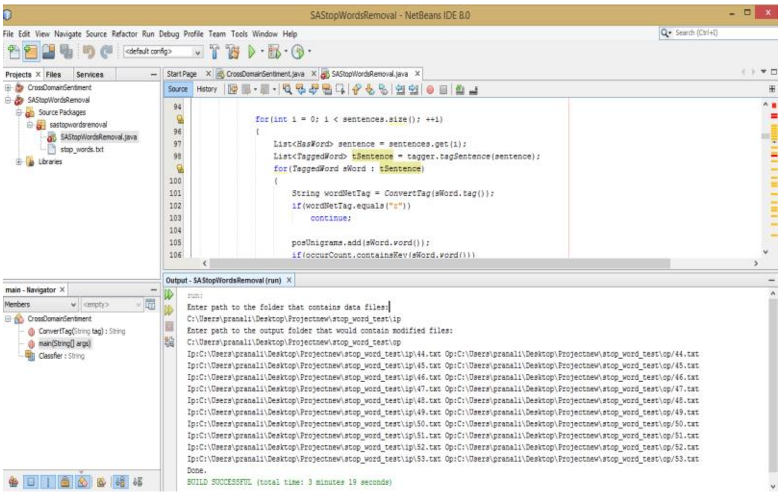

Members
<empty>
CrossDomainXmlFileSplitter
main(String[] args)
Output - CrossDomainXmlFileSplitter (run)

```
run:
Enter input file path
C:\Users\pranali\Desktop\all_dataset\dataset1\books\negative
Enter output folder path:
C:\Users\pranali\Desktop\extracted\book\neg
Length:1000
Writing file:0001
Writing file:0002
Writing file:0003
Writing file:0004
Writing file:0005
```

```
Writing file:0993
Writing file:0994
Writing file:0995
Writing file:0996
Writing file:0997
Writing file:0998
Writing file:0999
Writing file:1000
Files written
BUILD SUCCESSFUL (total time: 39 seconds)
```

Figure 2. XML file splitter

The dataset contains 1000 positive files and 1000 negative files for each domain. On this dataset preprocessing step is applied to remove noisy data. 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 English stop words as shown in Figure 3.



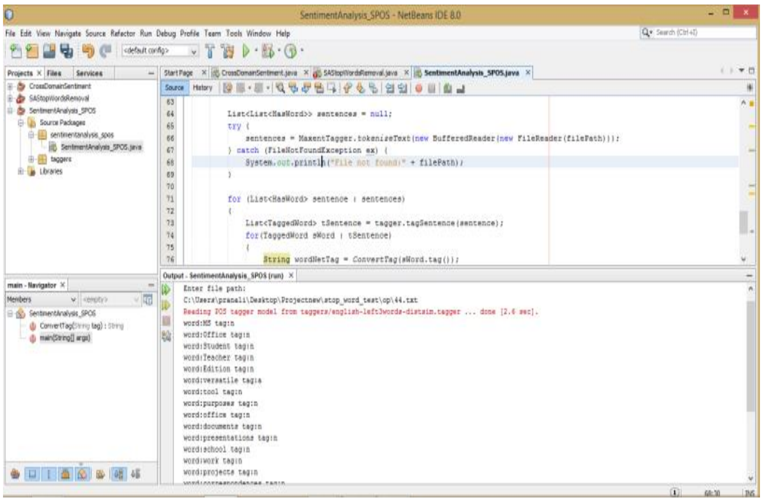
```
        for(int i = 0; i < sentences.size(); ++i)
        {
            List<HasWord> sentence = sentences.get(i);
            List<TaggedWord> tSentence = tagger.tagSentence(sentence);
            for(TaggedWord sWord : tSentence)
            {
                String wordNetTag = ConvertTag(sWord.tag());
                if(wordNetTag.equals("z"))
                    continue;

                posUnigrams.add(sWord.word());
                if(occurCount.containsKey(sWord.word()))
```

```
run:
Enter path to the folder that contains data files:
C:\Users\pranali\Desktop\Projectnew\stop_word_test\ip
Enter path to the output folder that would contain modified files:
C:\Users\pranali\Desktop\Projectnew\stop_word_test\op
Ip:C:\Users\pranali\Desktop\Projectnew\stop_word_test\ip\44.txt Op:C:\Users\pranali\Desktop\Projectnew\stop_word_test\op/44.txt
Ip:C:\Users\pranali\Desktop\Projectnew\stop_word_test\ip\45.txt Op:C:\Users\pranali\Desktop\Projectnew\stop_word_test\op/45.txt
Ip:C:\Users\pranali\Desktop\Projectnew\stop_word_test\ip\46.txt Op:C:\Users\pranali\Desktop\Projectnew\stop_word_test\op/46.txt
Ip:C:\Users\pranali\Desktop\Projectnew\stop_word_test\ip\47.txt Op:C:\Users\pranali\Desktop\Projectnew\stop_word_test\op/47.txt
Ip:C:\Users\pranali\Desktop\Projectnew\stop_word_test\ip\48.txt Op:C:\Users\pranali\Desktop\Projectnew\stop_word_test\op/48.txt
Ip:C:\Users\pranali\Desktop\Projectnew\stop_word_test\ip\49.txt Op:C:\Users\pranali\Desktop\Projectnew\stop_word_test\op/49.txt
Ip:C:\Users\pranali\Desktop\Projectnew\stop_word_test\ip\50.txt Op:C:\Users\pranali\Desktop\Projectnew\stop_word_test\op/50.txt
Ip:C:\Users\pranali\Desktop\Projectnew\stop_word_test\ip\51.txt Op:C:\Users\pranali\Desktop\Projectnew\stop_word_test\op/51.txt
Ip:C:\Users\pranali\Desktop\Projectnew\stop_word_test\ip\52.txt Op:C:\Users\pranali\Desktop\Projectnew\stop_word_test\op/52.txt
Ip:C:\Users\pranali\Desktop\Projectnew\stop_word_test\ip\53.txt Op:C:\Users\pranali\Desktop\Projectnew\stop_word_test\op/53.txt
Done.
BUILD SUCCESSFUL (total time: 3 minutes 19 seconds)
```

Figure 3. Output of data preprocessing

After stop word removal, using parser part of speech of sentence like noun, adjective, adverb, verb, etc. are extracted. Parsing is vital step as it gives opinion words as an output. 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 (POS) tagging. The part-of-speech is a category used in linguistics that is defined by a syntactic or morphological behavior of a word. The traditional English language grammar classifies POS in the following categories: verb, noun, adjective, adverb, pronoun, preposition, conjunction and interjection. The reason why POS tagging is so imperative 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 key role as an adjective. Firstly, text review is divided into sentences. Stanford parser 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 as shown in Figure 4.



```
        List<List<HasWord>> sentences = null;
        try {
            sentences = MaxentTagger.tokenizeText(new BufferedReader(new FileReader(filePath)));
        } catch (FileNotFoundException ex) {
            System.out.println("File not found:" + filePath);
        }

        for (List<HasWord> sentence : sentences)
        {
            List<TaggedWord> tSentence = tagger.tagSentence(sentence);
            for(TaggedWord sWord : tSentence)
            {
                String wordNetTag = ConvertTag(sWord.tag());
```

```
Enter file path:
C:\Users\pranali\Desktop\Projectnew\stop_word_test\op\44.txt
Reading POS tagger model from taggers/english-left3words-distsim.tagger ... done [2.6 sec].
word:MS tag:n
word:Office tag:n
word:Student tag:n
word:Teacher tag:n
word:Edition tag:n
word:versatile tag:a
word:tool tag:n
word:purposes tag:n
word:office tag:n
word:documents tag:n
word:presentations tag:n
word:school tag:n
word:work tag:n
word:projects tag:n
word:correspondences tag:n
```

Figure 4. Output of POS tagging

From each domain 1000 review files i.e. 500 negative and 500 positive files are taken for experiment. Using cross validation technique Source domain is divided into 5 parts and each part is taken as input. Target domain is fully taken. Results from each part are averaged. Table 1 shows accuracy from same technique for 12 classification tasks.

Table 1. Analysis of Accuracy from cross validation technique

Source→Target	Accuracy (%) Source1	Accuracy (%) Source2	Accuracy (%) Source3	Accuracy (%) Source4	Accuracy (%) Source5	Accuracy (%) Average
B→D	61.0	64.7	60.9	64.6	68.2	63.88
B→E	64.0	63.6	64.1	74.7	65.2	66.32
B→K	61.0	60.5	60.5	65.10	67.60	62.94
D→B	69.5	68.60	63.6	61.3	63.80	65.36
D→E	69.5	68.6	64.0	58.9	65.9	65.38
D→K	71.8	70.7	61.9	62.0	66.6	66.6
E→B	63.4	50.3	61.0	52.7	61.7	57.66
E→D	61.0	50.2	59.3	54.9	57.9	56.66
E→K	67.1	50.5	67.6	64.1	73.4	64.54
K→B	58.4	67.8	60.4	52.7	52.0	58.26
K→D	68.6	71.39	68.7	67.30	63.4	67.87
K→E	57.99	61.3	57.9	55.3	55.3	57.55

When size of training data is less than testing data accuracy values are decreasing compare to full training data. Accuracy getting after dividing target domain into parts and taking Source domain as full for each classification task is average 98.4%. Kitchen domain and DVD domain are showing average 67.87 % accuracy. Most of the words from these domains are similar. Electronics and DVD domain are showing less accuracy 56.66%. Training data is randomly divided. Fold 4 showing 74.7% accuracy for Book to Electronics domain whereas fold 5 shows 73.4% accuracy for Electronics to Kitchen domain as shown in

Figure 5. Figure 6 shows the effect of combining multiple source domains. We see that the combination of DVD and Electronics as well as Book and DVD as source domain gives highest accuracy. Other observation shows that when we use two source domains is always greater than the accuracy if we use single domain as a source. Experiment was done for 400 files.

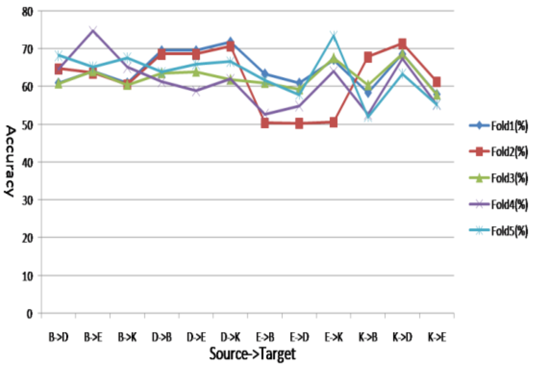


Figure 5. 5 Fold validation

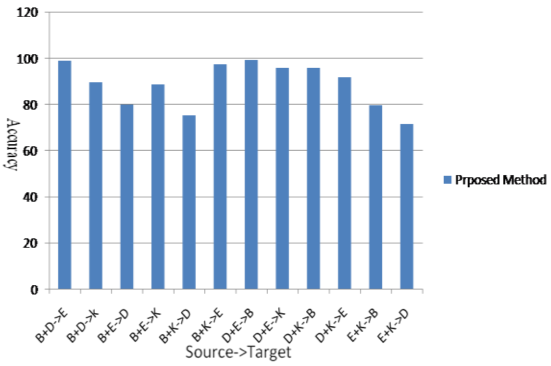


Figure 6. Accuracy analysis of combination of source domains

Combination of Book and Kitchen as source applied on DVD domain as target showing less accuracy that is 75.25% compared to others. Also Electronics and Kitchen, Book and Electronics as a source and DVD as a target producing 71.5% and 79.75% accuracy respectively. From observations accuracy for these domains is less compared to other combinations and DVD as target. It states that words from source domains are not matching with target domain as shown in Figure 6. Efficiency of model is dependent on domains similarity.

4. CONCLUSION

The commitment of this paper is to apply the semi supervised method to build domain adaptation model to extract features. Proposed approach utilizes maximum entropy classifier to classify reviews into two classes positive and negative. Clustering is appled on classified words using pointwise mutual information. The experiemnatl results deliver good accuracy value for 5 fold cross validation and combination of source domains. Blitzer et al. [14] multiproduct dataset is used for experiments. In the proposed framework, clustering for only unigrams is used. Bigrams can be used in future. Non word features can be included in approach.

REFERENCES

[1] Aciar S., Zhang D., Simoff S., Debenham J., "Informed Recommender: Basing Recommendations on Consumer Product Reviews," *In Intelligent Systems, IEEE,* Vol.22, Issue 03, pp.39-47, 2007.

[2] Bing Liu, “Sentiment Analysis & Opinion Mining”, *Morgan & Claypool Publishers,* Kindle Edition, 2012.

[3] Bermingham A. Conway M., McInerney L., O'Hare N., and Smeaton A., “Combining Social Network Analysis and Sentiment Analysis to Explore the Potential for Online Radicalisation,” *In Proc. of Int’l Conf. on Advances in Social Network Analysis and Mining,* Athens, Greece, July 20-22, pp. 231-236, 2009.

[4] Blitzer J, R. McDonald, F. Pereria, “ Domain Adaptation with Correspondence Learning,” EMNLP’06, In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA,USA, Pages 120-128, 2006.

[5] Matthew Whitehead , Larry Yaeger, “Building a General Purpose Cross-Domain Sentiment Mining Model,” *Computer Science and Information Engineering, 2009 WRI World Congress on* ,Volume- 4 , pp. 472-476, Los Angeles, CA, 2009.

[6] Pan Sinno Jialin, Ni Xiaochuan, Jian-tao Sun, Qiang yang, and Zheng Chen, “Cross Domain Sentiment Classification via Spectral Feature Alignment,” ACM, 19th International World Wide Web *Conference* ,Raleigh, North Carolina, USA, 2010.

[7] Kang Liu, Liheng Xu, and Jun Zhao, “Co-Extracting Opinion Targets and Opinion Words from Online Reviews Based on the Word Alignment Model,” *IEEE transactions on Knowledge Data Engineering,* vol. 27, no. 3, March, 2015.

[8] Balamurali A R, Debraj Manna, Pushpak Bhattacharyya, “Cross-Domain Sentiment Tagging Using Meta-Classifier and a High Accuracy In-Domain Classifier,” *Proceedings of ICON 2010: 8th International Conference on Natural Language Processing*, 2010.

[9] Bollegala D. Weir D., and Carroll J., "Cross-Domain Sentiment Classification using a Sentiment Sensitive Thesaurus," *In Knowledge and Data Engineering, IEEE Transactions,* Vol. 25 Issue: 8, pp.1719-1731, 2013.

[10] Li Shoushan, Yunxia Xue, Zhongqing Wang, Guodong Zhou, “Active Learning for Cross-domain Sentiment Classification,” *In IJCAI*, pp. 2127-2133, 2013.

[11] Dhillon Inderjit S., Subramanyam Mallela, Kumar Rahul, “A Divisive Information-Theoretic Feature Clustering Algorithm for Text Classification,” In *Journal of Machine Learning Research,* pp.1265-1287, 2003.

[12] Himanshu S. Bhatt, Deepali Semwal, Shourya Roy, “An Iterative Similarity based Adaptation Technique for Cross Domain Text Classification,” *Proceedings of the 19th Conference on Computational Language Learning*, pp.52-61, Beijing, Chiana, July 30-31, 2015.

[13] Nigam, Kamal, John Lafferty, Andrew McCallum, “Using maximum entropy for text classification,” In *IJCAI-99 workshop on machine learning for information filtering*, vol. 1, pp. 61-67. 1999.

[14] Blitzer J., M. Dredze, F. Pereira, “Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification,” In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 432-439, Prague, Czech Republic, 2007.