Sentiment Classification using Cross-Domain Data with N-gram Method

Prakasini K[1], Asha Baby[2]

[1*] PG Scholar, Computer Science and Engineering, Vimal Jyothi Engineering College, KTU, Kannur, India
[2*] Assistant Professor, Computer Science and Engineering, Vimal Jyothi Engineering College, KTU, Kannur, India

e-mail: prakasini.k424@gmail.com,ashababy@vjec.ac.in

Abstract— Sentiment classification is a data mining method that widely different from other types of traditional information extraction. It includes only two basic categories (positive/negative and stars) compare to text classification. Sentiment analysis is one of the most prominent case of natural language processing. Sentiment classification will predict the polarity of reviews automatically either positive or negative with respect to sentiment polarity of sentence. This paper presents a classification method extended with feature clusters using n-gram words as features. Using a Spectral Feature Alignment algorithm those n-gram features are aligned as domain specific and domain independent for feature clustering. Domain independence achieved using cross domains. N-gram features can provide much better results for polarity classification within smaller false positive and false negative rates.

Keywords— Cross domain, sentiment analysis, Feature Extraction, Classification, Machine Learning

I. INTRODUCTION

With the massive growth in web services, large amount of useful data are generated by people. For example, through various social websites like Facebook, Twitter, various blogs on diverse topic. Sentiment classification is widely known as a domain-dependent task, since different expressions are used to express sentiments in different domains [1]. Supervised learning algorithms that require labeled data have been successfully used to build sentiment classifiers for a given domain. However, sentiments are expressed differently in different domains, and it is costly to annotate data for each new domain in which it would like to apply a sentiment classifier. If considering single domain for classification the classifier is effective in terms of sentiment polarity. But extending to cross domains a classifier trained on one domain might not perform well on another domain because it fails to learn the sentiment of the unseen words.

The Cross-Domain Sentiment Classification focuses on the challenge of training a classifier from one or more domains (source domains) and applying the trained classifier on a different domain (target domain). The sentiment classifier trained in one domain may not perform well in another domain. Since there are massive domains, it is impractical to annotate enough data for each new domain. Thus, cross-domain sentiment classification which transfers the sentiment knowledge from domains with sufficient labeled data (i.e., source domain) to a new domain with no or scarce labeled data (i.e., target domain). In Cross-Domain Sentiment Classification, the features or words that appear in the source domain do not always appear in the target domain. So the words which are non-significant for classification in the source domain do not contribute to the target domain in supervised cross-domain sentiment classification. Hence, identification of significant words in the source domain restricts the transfer of irrelevant information to the target domain. For example, in "Electronics" domain "easy" is usually used in positive reviews, e.g., "this camera is very easy to use". However, it is frequently used as a negative word in "Movie" domain [2]. For instance, "the climax of this movie is easy to guess". Thus, the sentiment classifier trained in one domain usually, can't be applied to another domain directly.

Table 1. Reviews for cross domain sentiment classification

	Movies	Products
+	**Excellent** cast, story line, performances.	**Great** phone. **Excellent** product for the price.
+	Twilight is a **great** film. I, really **liked** that.	The headset fulfills my requirements so I am **happy** with my purchase.
-	Very **disappointing**. The acting, is **terrible**, and the writing is w**orse**.	This is by far the **worst** purchase I've made on Amazon.

Since sentiment is expressed differently in different domains Some feature mismatch occurs in between domain words. So that a classifier trained in one domain need not perform well on another domain. In Table 1 shows that, some user sentiment reviews from movie and product domains. In the movie domain, the words "excellent", "great", and "liked" are used to express positive sentiment and "disappointing", "terrible" and "worse" are used to express negative sentiment. While in product domain "great", "excellent" and "happy" are used to express positive sentiment and "far" and "worst" express negative sentiment. In existing methods, reviews are classified at various levels i.e. word level, sentence level and document level. The proposed work is based on sentence level which classifies the sentence as positive or negative, based on overall sentiment words expressed in the sentence.

The rest of the paper is organized as follows: Section II contain the related works. Section III contain the some methodology to implement proposed system. Section IV describes results and discussion. Section V concludes research work with future directions.

II. Related Work

Supervised sentiment classification is determine the sentiment of texts according to their opinion and attitude for a given entity. It got more and more attention because of its applications. However, supervised sentiment classification requires that labeled and unlabeled data should be under the same distribution, so that the classifier built by using the labeled data could be well applied to the unlabeled data. But in cross-domain classification field, the labeled and unlabeled data are from different domains, and often have different distributions.

Using Sentiment Sensitive Embedding [1] Bollegala create thesaurus for classification. An embedding technique was developed for training phase. Distributional properties of pivots and label constraints in source domains and geometric properties in unlabeled source and target domains.

A method is proposed to automatically create a sentiment sensitive thesaurus [2] that is sensitive to sentiment words from different domains. The created thesaurus is used to expand the feature vector in training and testing a binary
Classifier.

For Classification another concept is POS taggers. Xia in [7] make use of POS based ensemble model to efficiently integrate features with different types of POS tags to improve the classification performance. A POS tag based method, such as adjectives, adverbs, nouns etc. Finding the significance of these adverbs and adjectives from cross domain, and noun become less important. The POS information is supposed to be a significant indicator of sentiment expression.

The two classifiers are employed to select informative samples with the selection strategy of Query By Committee (QBC). Finally, the two classifier is combined to make the classification decision. Importantly, the two classifiers are trained by the unlabeled data in the target domain using the Label Propagation (LP) algorithm. Active learning in in-domain classification normally contains only a small amount of labeled data while active learning in cross-domain classification normally contains abundant labeled data in the source domain.

Word polarity is informative in the case of classification of text. So some word polarity can't identifies without the knowledge of domain. Detecting word polarity is a challenging topic in multi domain. Overcome disadvantages of transfer learning technique Yoshida [6] proposed a novel Bayesian probabilistic model to handle multiple source and multiple target domains. In this model, each word is associated with three factors: Domain label, domain dependence/independence and word polarity. Transfer learning utilizes the results learned in a certain source domain to solve a similar problem in another target domain.

III. Methodology

A. Overview

A Cross-Domain Sentiment Classification system must overcome two main challenges. First, it should identify which source domain features are related to which target domain features. Second, it requires a learning framework to incorporate the information regarding the relatedness of source and target domain features. Given a set of labeled reviews $D_s = \{x_i, y_i\}$ from source domain where x_i represents features and y_i represent sentiment label $y_i \in \{+1,-1\}$. To predict the label in unlabeled target domain $D_t = \{x_j\}$ where x_j represent features in target domain. Classifier is trained by labeled reviews of source domain and it is applied to classify the reviews of unlabeled target domain. To implement a classification technique extended with clustering of trained data to identify feature clusters and the shorthand notations for classification. The proposed method each domain have their own datasets. The tasks are divided into two sub tasks.

- To identify domain-independent features and
- To align domain-specific features.

In the first subtask, it is aimed to learn a feature selection function to select domain-independent features, which occur frequently and act similarly across domains and these domain independent features are used as a bridge to make knowledge transfer across domains possible. After identifying domain independent features, can be used to denote a feature selection function for selecting domain-specific features, which can be defined as the complement of domain-independent features. Spectral Feature Alignment (SFA) method classified as domain-specific or domain independent using the mutual information between a feature and a domain label. 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. SFA uses some domain-independent words as a bridge to construct a bipartite graph to model the co-occurrence relationship between domain specific words and domain-independent words. The idea is that if two domain-specific words have connections to more common domain-independent words in the graph, they tend to be aligned together with higher probability. Similarly, if two domain-independent words have connections to more common domain-specific words in the graph, they tend to be aligned together with higher probability. The proposed approach can be divided into four main stages such as: preprocessing stage, feature selection stage, classification stage and result interpretation as shown in Figure.1.

B. Data Collection

For cross domain amazon provides an excellent site for sentiment classification and analysis trail. We use the multi domain sentiment dataset from Amazon site. The dataset contains multiple domains reviews. Since same words express different polarity rate in different domains, a domain independent (i.e., list of words common in both domains) sentiment word list used to start analysis. The features are about the structure or the content. Each review is assigned with a sentiment label, -1 (negative review) or +1 (positive review), based on the rating score given by the review authors. In each domain, there are more than 1000 positive reviews and 1000 negative ones.

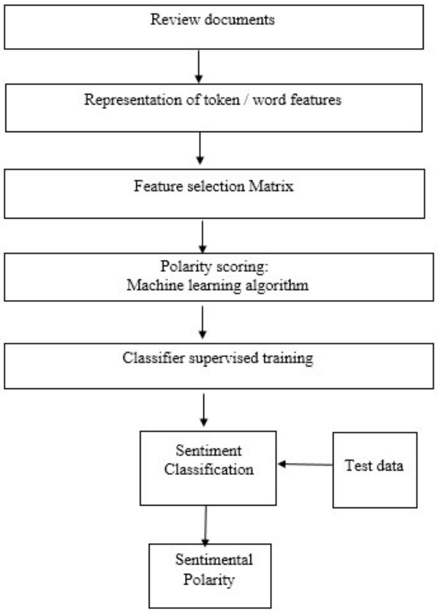


Figure. 1. Steps Involved in Sentiment classification

C. Cross-domain Sentiment Classification using N-Grams

Given two specific domains D_{src} and D_{tar}, where they are called as source domain and target domains respectively. Suppose there is a set of labelled sentiment data $\{x_i, y_i\}$ in D_{src} and unlabelled data $\{x_i\}$ in D_{tar}. The task of cross-domain sentiment classification is to learn an accurate classifier to predict the polarity of unseen sentiment data from D_{tar} using cross-domain words.

Table 2 shows the example of unigrams, bigrams and sentiment elements extracted from the reviews. First, model the review as a bag of words using tokenization and then extract unigrams, bigrams from each sentence. Bigrams are necessary for sentiment classification since semantic orientations of sentences are identified by bigrams. Then, from each source domain labelled reviews, sentiment features are created by appending the label of the review to each feature. The notation *P to indicate positive features and *N to indicate negative features.

Table 2. Sentiment feature from positive review

Steps	Example
A review sentence	Great phone. Excellent Product for the price.
Tokenization	Great // phone // excellent // product // for // the // price
Nouns, Verbs, Adjectives/ Adverbs	Great,phone,excellent, product, price
Unigrams and Bigrams	Great, phone, excellent, product, price, great+phone, phone+excellent, excellent+product, product+price
Sentiment Elements	Great*P, phone*P, excellent*P, product*P,price *P, great+phone *P, phone+excellent *P, excellent+product *P, product+price*P

D. Sentiment Classification

After the identification of the sentiment elements we need to determine a term, which is indeed a marker of opinionated content and is done with Sentiwordnet, a lexical resource in which each WordNet synset s is associated to three numerical scores Obj(s), Pos(s) and Neg(s), describing how objective, positive, and negative the terms contained in the synset are. The method used to develop Sentiwordnet is based on the quantitative analysis of the glosses associated to synsets, and on the use of the resulting vectorial term representations for semi-supervised Synset classification [18]. The list of words in Sentiwordnet is used to initialize the analysis. The bag of words are used to map the domain independent words. By getting the domain independent words, it is easy to calculate domain specific words. i.e., complement if independent words

Next the domain specific words mapped to domain independent words using a bipartite graph for a cross domain analysis. The clustering is done using Spectral Feature Alignment algorithm.

Algorithm: Construction of cross-domain sentiment classifier using Spectral Feature Alignment with Sentiwordnet.

Input: Labeled source Domain data $D_{src} = \{x_{src}, y_{src}\}$ and unlabeled target Domain $D_{tar} = \{x_{tar}\}$ and Sentiwordnet file.

Output: Predict the label of the target domain.

1. Pre-process the selected review file into the bag of words.
2. Extract Domain Independent features and domain-specific features from the given reviews on D_{src} and D_{tar} to select l domain-independent features. The remaining m-l features are treated as domain-specific features.
3. Create co-occurrence matrix $M \in R^{(m-l)*l}$ between domain independent features with domain specific features.
4. Create edges of the graph, if an edge in E connects two vertexes in the source domain and target domain respectively.
5. Find a new representation of feature vector is used to train a sentiment classifier to predict the label of the target domain.
6. Test the classifier in the target domain.

IV. Results and Discussion

The proposed work can be carried out using Intel processor with 2GB RAM and 300GB Hard disk. The development environment used for implementing the proposed model is Java which runs on 32-bitWindows operating system. The proposed system uses a dumped file called Sentiwordnet.

A. Parameter for Evaluation

The benchmark dataset is collected from the large online shopping site Amazon.com, which is familiarly used in many cross-domain sentiment analysis tasks. Amazon reviews of four different domains are discussed for cross-domain sentiment classification: Movies (M), Products (P), Books (B), and Magazines (Ma). In the dataset, each review assigned a -1 (negative) and +1 (positive) label according to the score given by the product users. In this dataset, we can construct up to 12 cross-domain sentiment classifications. M→P, M→B, M→Ma, P→B, P→Ma, P→M, B→Ma, B→M, B→P, Ma→M, Ma→P, Ma→B. In this, the letter preceding the arrow is source domain and after the arrow term as target domain. For each pair of domain classification, we calculate the accuracy of system performance.

The classification Accuracy on target domain is used as a metric for evaluation. It is the fraction of the correctly classified target domain reviews from the total number of reviews in the target domain, and is defined as follows:

$$\text{Accuracy} = \frac{\text{Number of correctly classified target reviews}}{\text{Total number of reviews in the target domain}}$$

So the accuracy of prediction of target domain reviews by a classifier is computed using above equation. The accuracy measure gives the percentage of reviews that are correctly classified.

B. Analysis of Proposed Model

For experiments, among four domains, one domain is taken as source domain and another domain is selected as target domain. To study the effect of enhanced sentiment corpus for one source domain and one target domain, 12 different combinations of one source domain and one target domain is analysed.

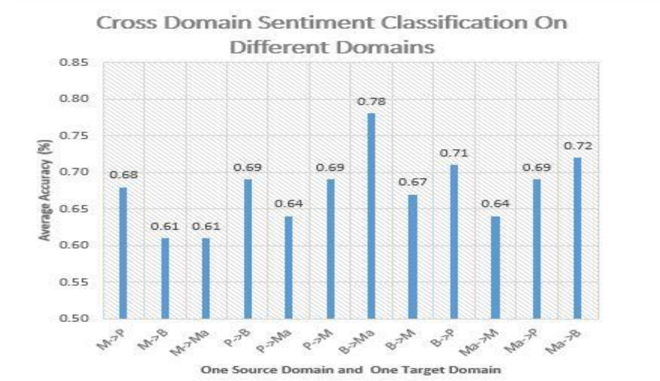


Figure. 2. Classification accuracy of one source and one target domain using proposed N-gram model

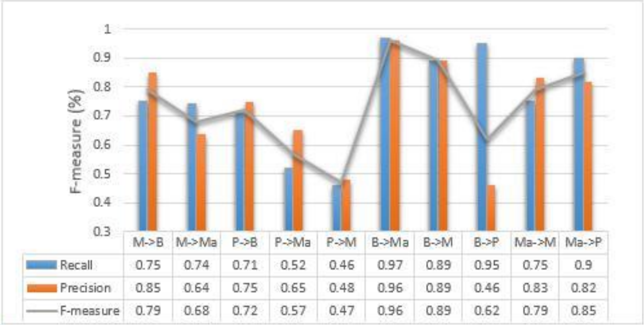
From the analysis, M→P, Movie domain is selected as source domain and Product domain is selected as target domain which produces accuracy 0.68%. Next M→B and M→Ma produces 0.61% accuracy each. Here B denotes books domain and Ma denotes Magazine domain.. Accuracy is used as an evaluation measure. Accuracy is the proportion of correctly classified examples to the total number of examples. On the other hand, error rate refers to incorrectly classified examples to correctly classified examples.

$$\text{Recall} = \frac{\text{Number of correct positive prediction}}{\text{Number of positive examples}}$$

$$\text{Precision} = \frac{\text{Number of correct positive prediction}}{\text{Number of positive predictions}}$$

$$\text{F-measure} = \frac{2\ \text{*Precision *Recall}}{\text{Precision + Recall}}$$

To evaluate the performance of sentiment classification, this paper has adopted precision, recall, SentiWordNet scores are exploited together with additional features to assign a polarity to a text using machine learning and F-Measure as a performance measure. Figure 3 shows that Analysis of the proposed method of recall and precision vs F-measure. This work was able to produce relatively good results for some of the domain, and it was able to handle only two classes with an acceptable accuracy. Domain-specific and domain-independent words compared to SentiWordNet shows average matching percent 67.77 %. The experimental results of proposed approach have shown a significant increase in accuracy for different domains over baseline approach as the proposed framework emphasizes on the granularity of the word. This is the major change in classifier in comparison to the traditional approach. Importance of each word that has more impact on results of the classifier was classified.

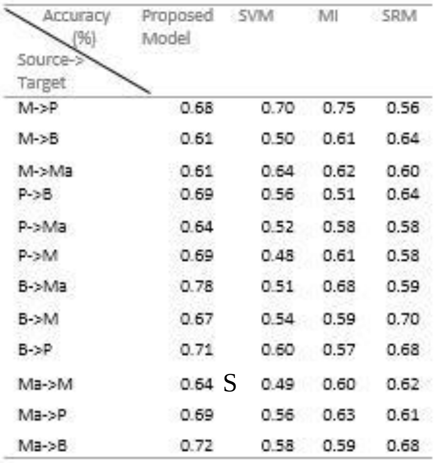


	M->B	M->Ma	P->B	P->Ma	P->M	B->Ma	B->M	B->P	Ma->M	Ma->P
Recall	0.75	0.74	0.71	0.52	0.46	0.97	0.89	0.95	0.75	0.9
Precision	0.85	0.64	0.75	0.65	0.48	0.96	0.89	0.46	0.83	0.82
F-measure	0.79	0.68	0.72	0.57	0.47	0.96	0.89	0.62	0.79	0.85

Figure. 3. Analysis of Precision & Recall vs F-measure

C. Comparison with Baseline Models

From Figure 4, it is evident that Book and Product, if considered as a source domain, achieve a good compatibility with Magazine and Movie domain, which is considered as target domain. Besides, Book and Magazines are compatible domains.



Accuracy (%) Source-> Target	Proposed Model	SVM	MI	SRM
M->P	0.68	0.70	0.75	0.56
M->B	0.61	0.50	0.61	0.64
M->Ma	0.61	0.64	0.62	0.60
P->B	0.69	0.56	0.51	0.64
P->Ma	0.64	0.52	0.58	0.58
P->M	0.69	0.48	0.61	0.58
B->Ma	0.78	0.51	0.68	0.59
B->M	0.67	0.54	0.59	0.70
B->P	0.71	0.60	0.57	0.68
Ma->M	0.64 S	0.49	0.60	0.62
Ma->P	0.69	0.56	0.63	0.61
Ma->B	0.72	0.58	0.59	0.68

Figure 4. Comparative analysis of accuracy

Baseline methods use in this study are SVM Method, Mutual Information Method (MI) and Semantic Representation Model (SRM). SVM achieved was between 50.0% and 70.0%, MI was between 51.0% and 75.0% and SRM was between 56.0% and 70.0%, whereas the accuracy of proposed algorithm was between 61.0% and 78.0%. Only the Movie, Product, and Book were considered as the source domain, while, Book, Magazines, and Movie as a target domain, producing comparatively less accurate results than the baseline methods [Figure 5].

The Figure 5 shows that the highest accuracy is achieved when the source domain is Book (Target as Magazine) using the Proposed method, that is, 78%. For the rest of the domains, proposed shows similar accuracies to those achieved using the baseline models.

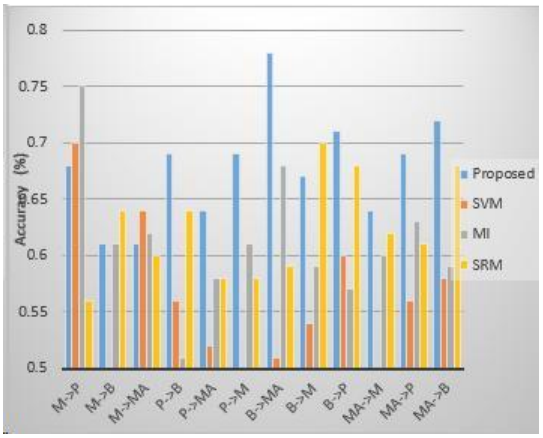


Figure. 5. Accuracy analysis

The comparison of the proposed model done with following models. First, The SVM classifier is trained on the matrices of each domain and then used to predict the classes of the remaining domains. Secondly, MI model relies on the mutual information values of features in the source domain. And finally, Semantic features representation relies on the features context to determine their relatedness to other features in the texts. The method that is used for this representation is called latent semantic analysis. It applies statistical techniques to a corpus to represent the "contextual-usage meaning" of features. LSA is usually applied on feature frequency matrices without-pre-processing of the source domain.

V. Conclusion and Future Scope

The Cross-domain sentiment classification is the task of classifying sentiment documents in a target domain using labelled data from a different domain. A major challenge in cross-domain sentiment classification is that the sentiment is expressed using different words across different domains. The proposed system develops a cross-domain sentiment classifier using an automatically extracted sentiment lexicons. To overcome the feature mismatch problem in cross-domain sentiment classification, it uses labelled data from multiple source domains and unlabeled data from the source and target domains to compute the improved relatedness of features and construct a Sentiment Corpus. The proposed system extends the feature vectors by using the created corpus. Then using these extended vectors, a binary classifier is trained from the source domain labelled reviews to predict positive and negative sentiment in reviews. In future, it can be extended to perform classification of positive, negative and neutral reviews. It can also be extended to overcome the problem of word Polysemy in cross-domains. Future studies can be taken up to determine the co-clustering of words and documents from a different domain.

Acknowledgment

I wish to record my indebtedness and thankfulness to all who helped me prepare this Project and present it in a satisfactory way. I am greatly indebted to God Almighty for being the guiding light throughout, with his abundant grace and blessing that strengthened me to pursue this endeavor with confidence. I am especially thankful to my guide for giving me valuable suggestions and critical inputs in the preparation of this report.

References

[1] D. Bollegala, T. Mu, and J. Y. Goulermas, "Cross-domain sentiment classification using sentiment sensitive embeddings," IEEE Transactions on Knowledge and Data Engineering, vol. **28**, no. **2**, pp. **398–410**, **2016**.

[2] Danushka Bollegala, David Weir, and John Carroll, "Cross Domain sentiment classification using a sentiment sensitive thesaurus," in IEEE Transactions on knowledge and data engineering, vol. **25**, no. S, August **2013**.

[3] .A.Kennedy and D.l nkpen, "Sentiment classification of movie reviews using contextual valence shifters," Computational Intelligence, vol **22**, pp.**110-125**, **2006**.

[4] N. X. Bach, V. T. Hai, and T. M. Phuong, "Cross-domain sentiment classification with word embeddings and canonical correlation analysis," in Proceedings of the Seventh Symposium on Information and Communication Technology. ACM, **2016**, pp. **159– 166**.

[5] L. Li, X. Jin, and M. Long, “Topic correlation analysis for cross-domain text classification.” in AAAI, **2012**.

[6] Y. Yoshida, T. Hirao, T. Iwata, M. Nagata, and Y. Matsumoto, "Transfer learning for multiple-domain sentiment analysis-identifying domain dependent/independent word polarity." in AAAI, **2011**.

[7] R. Xia and C. Zong, “A pos-based ensemble model for cross-domain sentiment classification.” in IJCNLP. Citeseer, **2011**, pp. **614–622**.

[8] A. Garcia-Fernandez, O. Ferret, and M. Dinarelli, "Evaluation of different strategies for domain adaptation in opinion mining," in Language Resources Evaluation Conference (LREC), **2014**.

[9] Y. Choi, Y. Kim, and S.-H. Myaeng, "Domain-specific sentiment analysis using contextual feature generation," in Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion. ACM, **2009**, pp. **37–44**.

[10] W. Medhat, A. Hassan, and H. Korashy, “Sentiment analysis algorithms and applications: A survey,” Ain Shams Engineering Journal, vol. **5**, no. **4**, pp. **1093–1113**, **2014**.

[11] J. Blitzer, R. McDonald, and F. Pereira, ”Domain adaptation with structural correspondence learning,” in EMNLP 2006,**2006**.

[12] S. J. Pan, X. Ni, J.-T. Sun, Q. Yang, and Z. Chen, “Cross-domain sentiment classification via spectral feature alignment,” in Proceedings of the 19th international conference on World wide web. ACM, **2010**, pp. **751–760**.

[13] Lun Yan and Yan Zhang, "News Sentiment Analysis based on cross-domain sentiment word lists and content classifiers," ADMA 2012, LNAI 7713, pp **577-5SS**, Springer-Verlag Berlin Heidelberg **2012**.

[14] B. Pang and L. Lee, "Opinion mining and sentiment analysis, "Foundations and Trends in Information Retrieval, vol. **2,** no. **1-2**, pp.I-**135,200S**.

[15] Aurangzeb Khan, Baharum Baharudin, Sentiment classification by sentence level semantic orientation using sentiwordnet from online reviews and Blogs," in InU Comp Sci. Emerging Tech, vol **2**, no.**4** August **20 I I**.

[16] S. Li, Y. Xue, Z. Wang, and G. Zhou, “Active learning for cross-domain sentiment classification.” in IJCAI, **2013**.

[17] A. Esuli and F. Sebastiani, ”Sentiwordnet: A publicly available lexical resource for opinion mining,” in LREC 2006, **2006**, pp. **417- 422**.

[18] Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. ”Sentiword-net 3.0: an enhanced lexical resource for sentiment analysis and opinion mining.” LREC. Vol**. 10**. No. **2010. 2010.**

[19] M. K. Borude and M. R. Singhal, “Efficient opinion mining for multiple domain.”