International Journal of Electrical and Computer Engineering (IJECE)
Vol. 7, No. 2, April 2017, pp. 967~974
ISSN: 2088-8708, DOI: 10.11591/ijece.v7i2.pp967-974

An Approach for Big Data to Evolve the Auspicious Information from Cross-Domains

Preeti Arora[1], Deepali Virmani[2], P. S. Kulkarni[3]
[1,2]Departement of Computer Science Engineering, BhagwanParshuram Institute of Technology, New Delhi, India
[3]Departement of Mathematics, College of Engineering Roorkee, Roorkee, India

Article Info

Article history:

Received Oct 10, 2016
Revised Feb 19, 2017
Accepted Mar 8, 2017

Keyword:

Bag-of-words
Feature extraction
Labelled words
Opinion mining
Sentimental classification

ABSTRACT

Sentiment analysis is the pre-eminent technology to extract the relevant information from the data domain. In this paper cross domain sentimental classification approach Cross_BOMEST is proposed. Proposed approach will extract †ve words using existing BOMEST technique, with the help of Ms Word Introp, Cross_BOMEST determines †ve words and replaces all its synonyms to escalate the polarity and blends two different domains and detects all the self-sufficient words. Proposed Algorithm is executed on Amazon datasets where two different domains are trained to analyze sentiments of the reviews of the other remaining domain. Proposed approach contributes propitious results in the cross domain analysis and accuracy of 92 % is obtained. Precision and Recall of BOMEST is improved by 16% and 7% respectively by the Cross_BOMEST.

Corresponding Author:

Preeti Arora,
Departement of Computer Science Engineering,
GGSIPU, Bhagwan Parshuram Institute of Technology,
New Delhi, India.
Email: erpreetiarora07@gmail.com

1. INTRODUCTION

Sentimental Analysis is a way to process different blogs, reviews such as movie, beauty, online shopping sites, and etc to indicate their sentiments as positive or negative. Customer expresses their views related to product or facility they avail. By analyzing those views consumers can effectively improve their decision making. Classification of Sentiment has been applied in various areas such as analysis of market [1], opinion mining [2], opinion summarization [3]. There are various levels to express the sentiments i.e. Document-level, Sentence-level, and Aspect-level. In this paper, sentiment is used to determine the sentence in term of subjective or objective. If sentence is subjective then check the expression of sentence as positive or negative sentiments. In single-domain analysis, all the sentiments related to a specific domain only. This training data might not produce an ample scope for different domains. So, there is a need of a solution for cross domain analysis [4] which trains the classifier from one or more domains known as source domains and utilizes the trained classifier in a different domain known as target domain. While a review express †ve polarity in one domain may be act as −ve polarity in another domain therefore a classifier is needed to transfer the expertise in different domains for a better performance. In this paper an approach is proposed to reveal domain dependent words and infer independent words.

A virtuous approach and Cross BOMEST algorithm for the cross-domain sentiment classification is proposed to boast the cross-domain data, to minimize the gap between domains. This algorithm is the modified version of BOMEST [5] which works efficiently on single domain with the accuracy of 78 %. For the cross-domain analysis reviews for various products (Baby, Beauty, Electronics, and Health products) are gathered from Amazon. Our proposed Cross_Boms classifier is used to model the relationship between domain-specifics and self-sufficient words by various phases explained in detail in further Section.

The paper is organized as, Section 2 comprises the literature and study conducted. Section 3 and 4, elaborate our proposed approach and Cross BOMEST Algorithm for Amazon dataset. Finally, the results of our experiments are validated in Section 5 and Section 6 defines the future scope of the work.

2. LITERATURE REVIEW

Hu and Liu [1] introduced feature based target extraction on consumer electronics reviews treated as the first work of Target Extraction. They proposed an approach on a statistical analysis of the review terms based on association mining. Manually selected subset of product features yields a precision of 0.72 and a recall of 0.80. Blitzer et al. [4] focuses on cross domain classification and focuses on the challenges of training a classifier from source domains and applying the trained classifier in a target domain as identification of feature selection and the learning framework to find the significance of source and target domain features. Pan and Ni [6] proposed a method for sentiment classification to bridge the gap between the domains, using spectral feature alignment (SFA) algorithm to *align* domain-specific words from different domains into unified clusters. These clusters can be used to reduce the gap between domain-specific words from the two domains, hence enhance the sentiment classifier.

3. CROSS_BOMEST

3.1. Proposed Approch

In this paper we propose an approach for Cross Domain Analysis. Till date, the existing approaches for the cross domain deal with single source domain and a classifier to predict target domain. The block diagram for Cross_BOMEST is shown in Figure. 1 is consists of two phases, detailed explanation of each phase is as follows:

Phase 1 deals with forming of Lexical_Boms_Dictionary and used this dictionary to increase the †ve polarity reviews. Number of steps involved and the output of all these steps are shown Section 4. The first step of our proposed approach Cross_BOMEST starts with gathering all the reviews. The Dataset used for implementation is Amazon data set http://jmcauley.ucsd.edu/data/amazon/. Used data set contains ***1,60,792*** reviews of Baby product , ***1,98,502*** reviews of Beauty product, ***3,46,355*** reviews of Health product and ***16,89,188*** reviews of Electronics product from May 1996 to July 2014, for the analysis. Then Junk data such as hyperlinks, <div>, <p>,
 etc are removed. Also all the repetitive words, stop words, images, url, videos and audios which do not contribute to the meaning of the sentence are removed. After that the Porter Stemming Algorithm is used to remove the suffix from the words and bring it to the root. After stemming, reviews are tokenized using BOW (Bag of words). Then BOMEST [5] is called for the POS tagging which effectively identify the nouns, verbs, adverbs and adjectives that have effect on the text. Therefore this module is capable of creating an indexed data, assign the score to it and store it into the trained dictionary [7].

In the last step, word is taken from the list created in above step and identified all synonyms of that word available in the reviews using Ms Word Introp. All the synonyms were replaced with the word and total occurrence of the word is calculated. For example, dataset contain "Even" as †ve, "Bad" as −ve polarity with their number of occurrences as shown in Table 1.

Table 1. †ve, −ve word with synonyms & occurrences

Polarity	Word	Synonyms	Polarity	Word	Synonyms
†ve (45,182)	Even (20854)	Still(22085), Smooth(1516), Constant(556), Equal(171)	−ve (6057)	Bad (4053)	Awful(576), Evil(31), Serious(378), Poor(1019)

"Even" is matched with its synonyms and all the synonyms are replaced with the word. Total count is evaluated as shown in Figure 2(b). Similarly "Bad" is compared to its synonyms and count is generated as shown in Figure 2(b).

Phase 2 is the vital part of an approach consists of following steps: Here, different combinations of domains are considered as the source domain such as Baby & Beauty, Beauty & Electronics and etc. Now merge all the †ve, −ve Lexical_Boms_Dictionary for the source domains and fetch all the self-sufficient words using Self-sufficient word collector. For the cross domain analysis Cross_Boms classifier is used to predict the †ve, −ve polarity of the target domain i.e. any domain except source. Domain-dependent words are used solely and monotonously in the document. Some of them do not contribute to the sentiments [8] for a domain therefore elimination of these words from the dataset will enhance the performance of classifier. In sentimental Classification, targets are the frequent words, POS, phrases or terms that have great effect on the opinion [9] to shows †ve, −ve polarity. Selection of proper target yields higher accuracy in classification by reducing the extensity of a text. These features are passed to Cross_Boms Classifier [10] to discover and eliminate unessential, inappropriate and redundant aspects from data that do not contribute to the accuracy. Therefore with the help of these features, in cross domain analysis precision, recall and accuracy is calculated as shown in result Section. The salient feature of Cross_Boms Classifier is the ***WSR*** (weight synonym replacement) scheme in place of the term frequency [11] of each word. The WSR focuses on reducing the weight of "very bad" class of source domain data and promote the weight of "very-good" class to generate more accurate results in target domain. At last classifier is trained on source domain and used to predict the †ve, −ve polarity reviews of target domain using steps mentioned in algorithm in next Section.

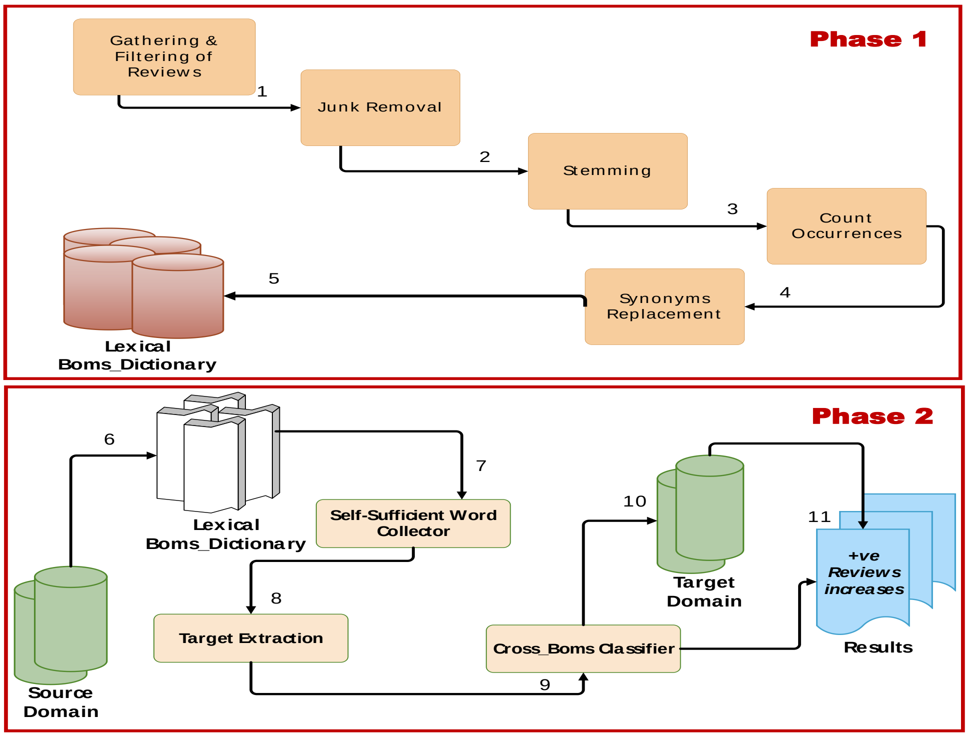
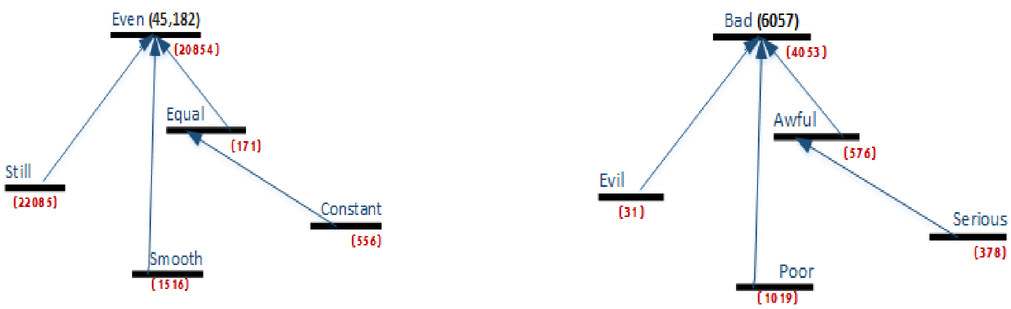


Figure 1. Block Diagram for Cross Domain Analysis



(a) (b)

Figure 2. (a) †ve Synonym Replacement, (b) −ve Synonym Replacement

3.2. Proposed Algorithm

Cross_BOMEST the, designed algorithm, which takes input as Amazon Reviews [12] datasets of Baby, Beauty, Health and Electronics products and gives better efficiency than BOMEST, which work on single domain. Expression of user's views are different in different domains, so to analyzing different domains is a time consuming and costly process as classifier has to be trained each time for a domain. Therefore there is a need of novel approach that can be work efficiently on the cross domain. Hence, Boms_Classifier is used to enhance the accuracy of the existing method by the combing two different source domains and predict the results of the target domain by the steps defined in algorithm.

Algorithm 1: BOMEST on Cross Domain

Input : − *RDS*1, *RDS*2, *RDS*3 & *RD*4 *Review DataSet of Beauty, Baby, Health & Electronics*
Output: −*Analysis of all RDS in terms of Accuracy*

1. *Begin*
2. *Filtering of Reviews*

   Read the JSON to fetch review text from the raw data.
3. *Junk Removal of the reviews*
   Stop word Removal
   Images, video, audio, url, symbols & hashtags
4. *Apply Porter Stemmer for stemming the data*
5. *Tokenization of the data using BOW*
6. *BOMEST* ( )
   POS Tagging , Generate Polarity ¶ as † ve, −ve, ± for each RDS
   Assign the score to the trained Dictionary
7. *WSR*( )
8. *Prepare † ve, −ve & ± Lexical_Boms_Dictionary*
9. *Cross_BOMEST*( )
10. *Print the result for Precision, Recall & Accuracy*
11. *End*

Algorithm 2: WSR

Input : ¶ *as* † *ve, −ve for each RDS*
Output : *Lexical_Boms_Dictionary*

1. *Begin*
2. *Using Linq we group the data and prepare † ve, −ve array*
3. *For* $i = 1$ *to len(groups)*
   a. *Get all synonyms using Ms Word Introp*
   b. *Let Count := Currentgroup.Count*
   c. *For* $i = 1$ *to len (Synonyms)*
   d. *If (Synonym) exists in groups*
   {
   Count = count + Synonym group.Count
   }
   e. *Insert current group.name , total count into Result File*
4. *Return the Result File as Lexical_Boms_Dictionary*
5. *End*

Algorithm 3: Cross_BOMEST

Input : *Combination of any two RDS as Source Domain & Lexical_Boms_Dictionary*
Output: *Result for Recall, Persicion & Accuracy*

1. *Begin*
2. *Choose two domains let baby, beauty as source domain*
3. *Tokenize the words using BOWP*
4. *WSR*( )
5. *BOMEST* ( )
   POS Tagging , Generate Polarity ¶ as † ve, −ve, ± for RDS
6. *Result = S_S_W(FirstDomainPositive, SeconDomainPositive)*
7. *Result+= S_S_W(FirstDomainNegative, SeconDomainNegative)*
8. *for* $i = 1$ *to result.lenght*
   for $j = 1$ *to list.lenght*
9. *if word.startwith(list[j])then*
   continue loop step 8
10. *else Re- Calculate Weight*
    if (word is + ve)then
    $W = W - 0.65$
    else
    $W = W + 0.65$
11. *Assign the score & generate* 5 *classes as ery bad, Bad, Avg, Good and Very good*
12. *Boms- classifier to train the target domain*
13. *Store the result for Precision, Recall & Accuracy in a file*

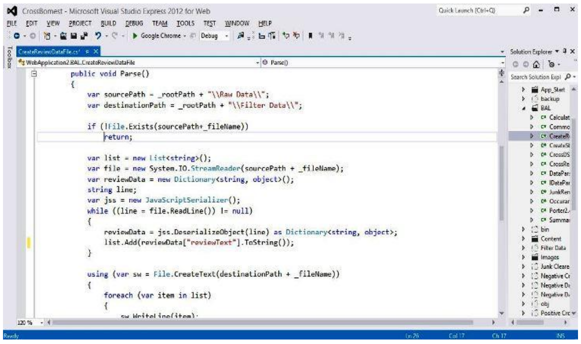
Algorithm 4: S-S-W (Self-sufficient words)

1. *Begin*
2. *for* $i =$ 1 *to FirstDomainList.Length*
3. *if* (*FirstDomainlist[i] exists in SecondDomainlist*)
4. *Result.Add(FirstDomainlist[i])*
5. *Calculate the threshold value for feature by counting diff count & sum count*
6. *feature above the theshold value traeted as domain dependent otherwise as self sufficient words*
7. *Now assign the weight to the words*
8. *End*

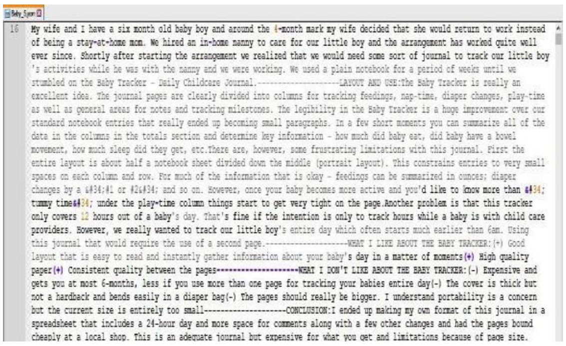
4. EXPERIMENT SETUP

The implementation of above algorithm is done by using C#.Net. Each phase is implemented by combination of user defined and inbuilt functions. Code and results of each step of phase 1 and phase 2 are provided below.

Step 1: Filtering of Reviews



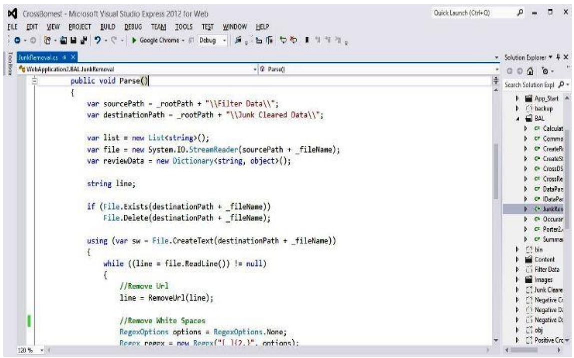
(a)



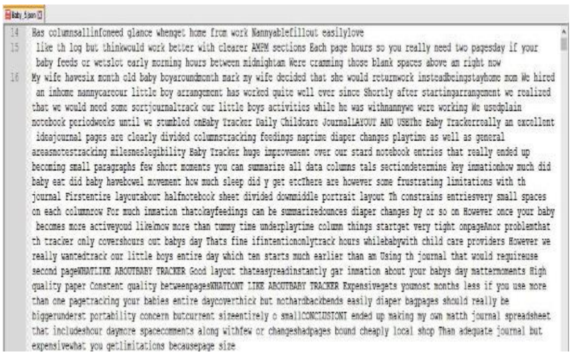
(b)

Figure 3. (a) Filter of JSON Reviews, (b) Raw Data of Reviews after Filtering

Step 2: Junk Removal



(a)



(b)

Figure 4. (a) Junk Removal from Reviews, (b) Reviews after Junk Removal

Step 3: Stemming using Porter Algorithm

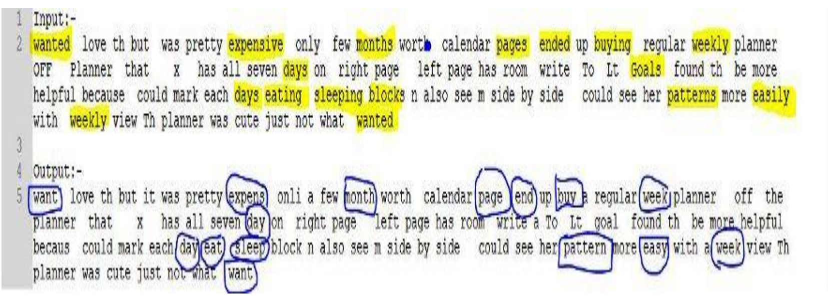
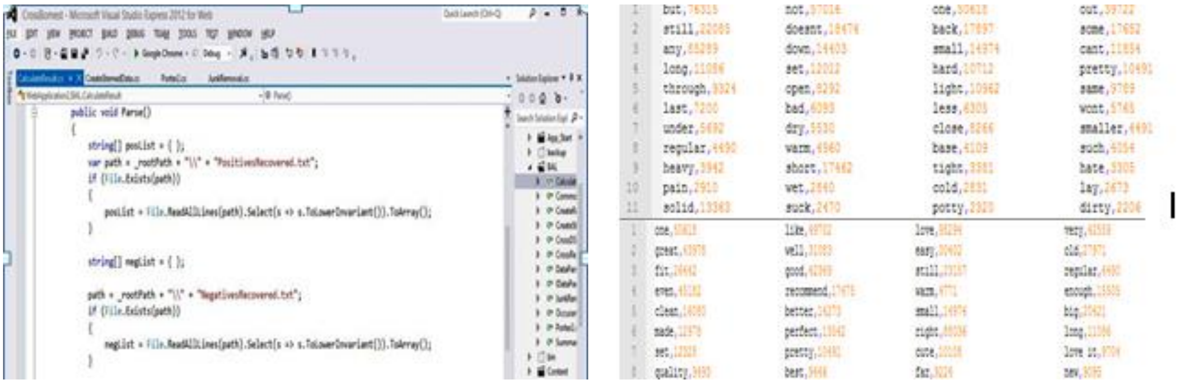


Figure 5. Result of Stemming

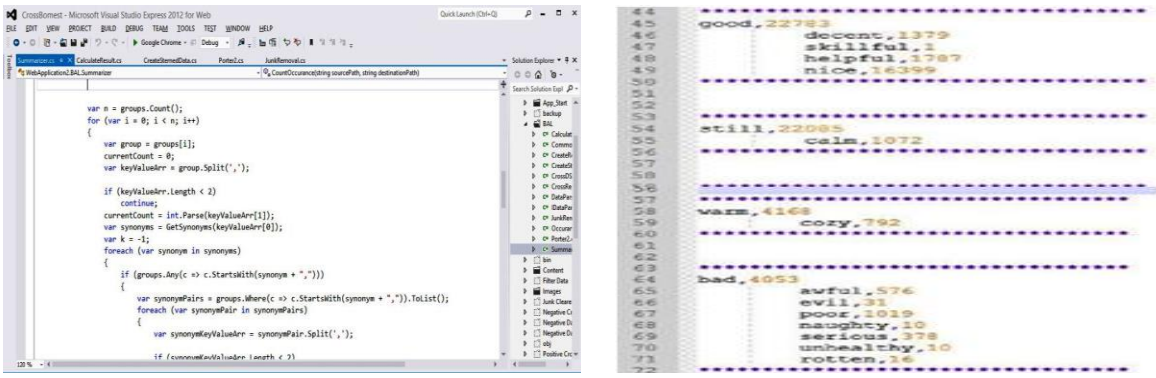
Step 4: Count Occurrence



(a) (b)

Figure 6. (a) Count Occurrence for †ve, −ve polarity, (b) †ve, −ve Count Occurrence

Step 5: Synonyms Replacement



(a) (b)

Figure 7. (a) Finding of Synonyms & Replacement, (b) † ve, − ve Synonyms Occurrences for †ve, −ve polarity

5. RESULT ANALYSIS

To generate the results for Cross_BOMEST approach for the cross domain sentiment analysis different product reviews from Amazon are used. For the implementation, randomly 1500 †ve and 1500 −ve polarity reviews are selected from each domain, out of which 1000 reviews of each polarity are used for training of classifier, the remainder is used for testing of classifier. In experiment, randomly combination of two domains act as source domain, remaining domain out of data set acts as the target domain such as ***Beauty + Baby→Electronics***, ***Beauty + Electronics→Health***, ***Electronics + Health→Beauty*** and ***Beauty + Health→Baby*** is called as A, B, C and D respectively for calculating the precision, recall and accuracy metric. The estimation metric is Cross_Boms classifier which effectively predicts the reviews of target domain correctly by determining domain-independent words from the source domain. It is clearly shown that the recall and precision metric increases for the Cross_BOMEST as compared to BOMEST. Using the features of ***TN+BOW*** 66% of precision is obtained whereas, ***STN+BM*** provides 77.25% of precision. Similarly ***TN+BOWN+BM*** provides 82.5% of recall whereas, ***BM+WSR+CBM*** shows 93.25% of recall. Also the Precision-Recall Graph for BOMEST and Cross_BOMEST is shown in Figure 8. and Figure 9. respectively.

Figure 8. Precision-Recall Graph

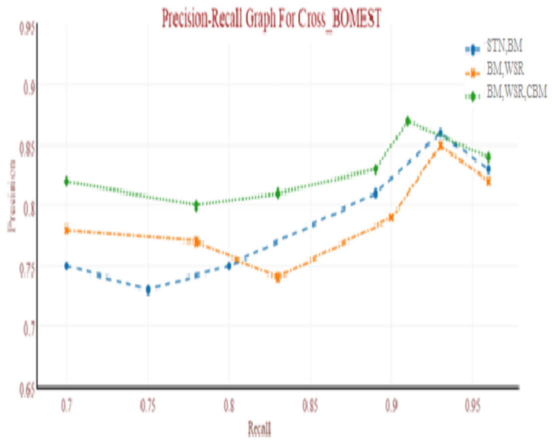


Figure 9. Precision-Recall Graph

Here, Table 2 & 3 presents the accuracy result for the BOMEST, Cross_BOMEST as shown in Figure 10 & Figure 11 respectively. By the result it is cleared that Boms_Classifier Using the ***BM+WSR+CBM*** provides the maximum accuracy of 92% which is higher than the BOMEST i.e. 81%.

Table 2. BOMEST Accuracy

BOMEST				
Target Extraction	Accuracy			
	A	B	C	D
TN,BOW	75	75	71	77
N,BOWN	78	79	74	80
TN,BOWN,BM	83	81	79	85

Table 3 Cross_BOMEST Accuracy

Cross_BOMEST				
Target Extraction	Accuracy			
	A	B	C	D
STN,BM	78	84	80	83
BM,WSR	86	85	89	90
BM,WSR,CBM	91	89	92	95

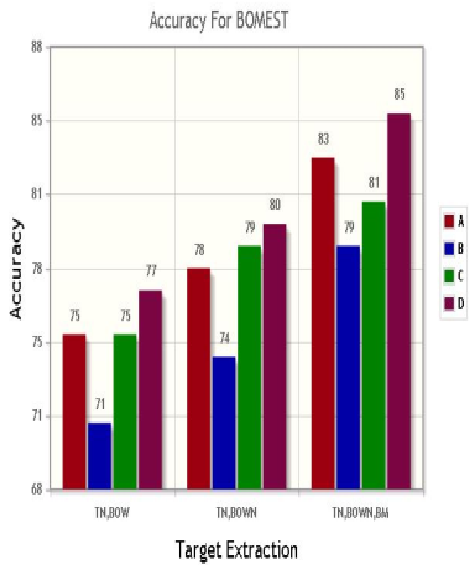


Figure 10. Accuracy Graph for BOMEST

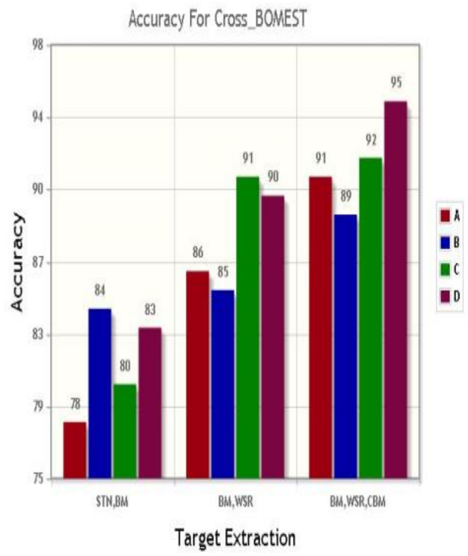
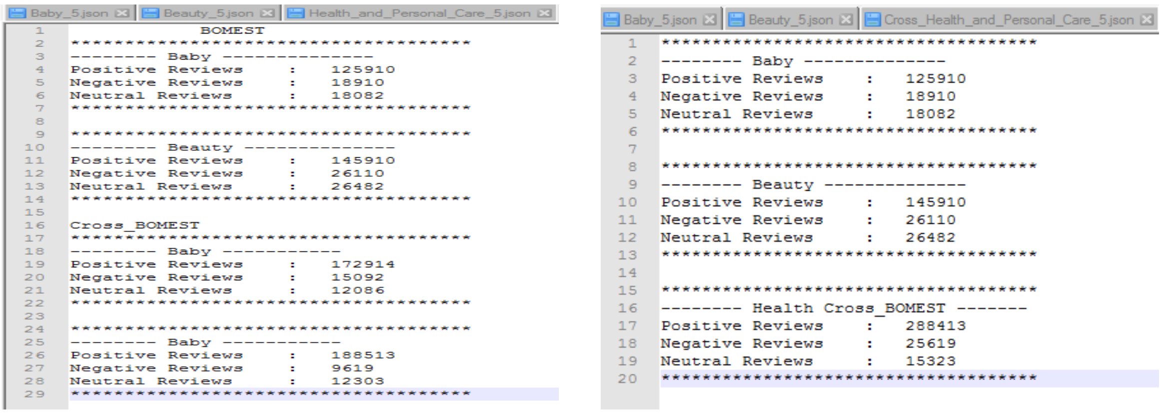


Figure 11. Accuracy Graph for Cross_BOMEST

Figure 12(A) shows the result for †ve reviews of different dataset in the single domain using BOMEST, Cross_BOMEST. Whereas Figure 12(B) shows the comparison results of Cross_BOMEST which effectively increases the number †ve polarity reviews when tested for cross domain analysis.



(a) (b)

Figure 12. (a) BOMEST, Cross_BOMEST, (b) Cross_BOMEST on target domain results on single Domain

6. CONCLUSION

In this paper, cross domain sentimental classification approach Cross_BOMEST is proposed. The proposed approach, works in two phases. Phase 1, is used to form Lexical_Boms_Dictionary. The dictionary is used to increase the †ve polarity reviews by filtering of reviews after junk removal and stemming. These reviews are then tokenized using BOW and BOMEST which is used for the POS tagging. Total count of the polarity of the reviews as †ve, and ‒ve, is stored in the indexed list, which is further used to replace all the synonyms with the matched word to escalate the polarity. In phase 2, two different source domains are trained to extract the reviews of the other remaining target domain. After that Cross-BOMEST approach is used to discover all the self-sufficient words that are used to bridge the gap between the sentences in different domains. With the help of, BM, WSR, CBM target extraction Cross_boms Classifier discover and eliminate unessential, inappropriate and redundant aspects from data that do not contribute to the accuracy. The accuracy of 92% is obtained by proposed Cross_BOMEST algorithm when applied to cross domains. Precision and Recall of BOMEST is improved by 16% and 7% respectively in single domain. Results, demonstrate that Cross_BOMEST shows 5% refinement in the precision and accuracy when compared to other existing techniques.

REFERENCES

[1] Hu M. and Liu B., "Mining and summarizing customer reviews," in *KDD*, pp. 168–177, 2004.
[2] Pang B. and Lee L., "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, pp. 1–135, 2008.
[3] Lu Y., *et al.*, "Rated aspect summarization of short comments," in *WWW*, pp. 131–140, 2009.
[4] Blitzer J., *et al.*, *"Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification,"* in Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, pp. 440–447, 2007.
[5] Jain V., *et al.*, "BOMEST a Vital Approach to Extract the Propitious Information from the Big Data," LNNS Springer's, 2016.
[6] Pan S., *et al.*, "Cross Domain Sentiment Classification via Spectral Features," in *WWW 2010, Raleigh, North Carolina, USA. ACM*, pp. 751–760, 2010.
[7] Fang H., "A re-examination of query expansion using lexical resources," in *ACL*, pp. 139–147, 2008.
[8] Bollegala D., *et al.*, "Cross-Domain Sentiment Classification Using a Sentiment Sensitive Thesaurus," *IEEE transactions on knowledge and data engineering*, 2013.
[9] Chen B., *et al.*, *"Extracting discriminative concepts for domain adaptation in text mining,"* in Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, New York, NY, USA, ACM, pp. 179–188, 2009.
[10] David B. S., *et al.*, "Analysis of representations for domain adaptation," in *Annual Conference on Neural Information Processing Systems 19, Cambridge, MA, MIT Press.*, pp. 137–144, 2007.
[11] Xie S., *et al.*, "Latent space domain transfer between high dimensional overlapping distributions," in *18th International World Wide Web Conference*, pp. 91–100, 2009.
[12] J. Mcauely, 2016. http://jmcauley.ucsd.edu/data/amazon/2016.