International Journal of Technology and Engineering System (IJTES)
Vol 7. No.4 2015 Pp. 268-273
©gopalax Journals, Singapore
available at : www.ijcns.com
ISSN: 0976-1345

DUAL DOMAIN MINER FOR FEATURE BASED OPINION SUMMARIZATION FROM PRODUCT REVIEWS

[1]R. ABIRAMI, [2]S. RAMESH
[1]PG Student of Computer Science and Engineering,
[2]Faculty of Computer Science and Engineering,
Anna University Regional Office Madurai, Tamilnadu, INDIA
[1]abiucev@gmail.com, [2]**itz_ramesh87@yahoo.com**

ABSTRACT

Opinion mining is the way of gathering the people's thought about a particular concept. It is to improve the decision making of new user in various domains such as product, movie, news media, social networking shares etc. Feature based opinion mining rely only on single domain corpus in most of the existing methodology. Feature based opinion mining in two different domain corpuses is complex. This paper proposes the Dual Domain Miner methodology, the features is extracted from different domain using inter dependent domain relevance (IDDR) score and Opinion is classified using Appraisal Identifier. The Inter dependent domain relevance technique use removal of redundant features and pruning of irrelevant features from two different domains with the help of the IDDR score and threshold value. An opinion prediction method is classified into three categories like opinion score generation, conjunction based and finally negation based prediction. The opinion word is initially producing the opinion score based on the online tool. The sentence contains more than one opinion term go for the connecting word opinion. Finally the opinion is predicted based on presence of negative word. The summary of two different domain's feature with respect to their opinion is generated. The comparative analysis with single and contrast domain proves the effectiveness of the Dual Domain Miner.

Key Words – Contrast Domain, Dual Domain, Feature Pruning, IDDR, Mapping, Opinion, Single domain.

I . INTRODUCTION

Large volume of online information such as documents, files, web pages, books, news media present in the web. Due to this information web mining carries in three different kinds of process such as web content mining, web usage mining and web structure mining.

Web content mining is the process of extracting knowledge from web page content; Web usage mining is the extraction of the models and patterns store the activities of the user and gather the user requirements; Web structure mining is way to discover the knowledge of hyperlinks to maximize the relation between the web pages [13].

The opinion mining is from the web content mining. It performs the prediction of sentiment of the new document or sentence or review through the gathering of emotions, sentiments, thoughts from the previous reviews, documents and sentences.

The feature and opinion words are identified through Part-Of-Speech (POS) tagging methodology. POS tagging is process of identify the part-of-speech of given input sentence. Based on this POS outcome we have to identify the features and opinion word. Normally feature in the form of noun and opinion word in the form of adjective and verb. In addition to that the connecting word and negative words are also extraction for the prediction of opinion's nature. The vast majority of the existing method use single domain corpus to perform the feature based opinion mining. Different domain needs different method to perform feature extraction and opinion prediction. Dual Domain Miner performs the feature extraction in two different domains to reduce the complexity of feature extraction in different domain. The feature extraction and pruning is the first steps of the feature based opinion mining using inter dependent domain relevance. These approaches extract the features of two different domains at the same time. The extraction is depends on domain relevance score and threshold value.

The opinion prediction is done with the help of opinion score from the online tool. It contains the corresponding scores of each opinion words. In some cases the sentences contain two opinion word, based on the connecting word the opinion of the one word can predict using another word. The opinion words connected using 'and' or 'both' means, they have same opinion; 'but' or 'neither or nor' tends to reversed opinion. The negation plays an important role in opinion prediction. It reverses the nature of the opinion word. Based on the negative word, final opinion is predicted. The summarization is the final step of the feature based opinion summarization. The summarization is in the form of each feature with their corresponding positive, negative and neutral opinion word or sentence. The paper is organized as the following sections. Section II describes the related work of Feature based Opinion mining. Section III depicts the Methodology. The Experimental analyses are shown in the Section IV.

II . RELATED WORK

The extractions of features from two different domain using inter dependent domain relevance mechanism [4] using camera and iPod domain. The features are extracted based on the domain relevance score. The score is measured using dispersion and deviation of the each term present in the review corpus. The intrinsic and extrinsic domain method [2] extracts the common features present in the two different domains. The domain like hotel and camera is used for feature extraction. These methods use two different threshold value to extract the common features such as intrinsic threshold value and extrinsic threshold value. Based on the threshold value and domain relevance score the features are extracted.

The features extraction is one of the important tasks in opinion mining. The product reviews [3] are gathered and find their opinion; the rating is the best method to express the opinion of the product. Hotel reviews are considered and found the opinion about the particular hotel. Another important task in opinion mining is the opinion prediction. The opinion of the product and political candidate are predicted using the lexical resource called SentiWordNet [1]. The automatic extraction of opinion based on the three different numeric score like obj(s), pos(s), neg(s). Initially the given opinion is split into subjective and objective then the subjective is split into positive and negative.

The unsupervised learning method is focused on another form of classification of reviews like recommended (thumbs up) and not recommended (thumps down) [5]. This method is worked with the help of the semantic orientation of the given phrase present in the reviews. The average semantic orientation is calculated by summation of the each phrase semantic orientation. If the value is greater than zero or positive then the given review is recommended otherwise the given phrase is not recommended. The comparative analysis is made between the rating from the website and recommendation predication using semantic orientation.

The features and opinions are extracted jointly using joint structure tagging [6]. Instead of linear chain, linguistics representation is incorporated into modular representation. The tree structure describe the jointly extraction of features and opinion. The new type of opinion mining is to identify the opinion with its holder and topic [7]. Normally the online news media text is using this kind of prediction. The FrameNet data is knowledge to predict the holder and topic of the opinion.

Phrase level sentiment analysis [11] classifies the phrase into neutral or polar, polar is classified as positive or negative. T. Wilson et.al create corpus and add contextual polarity judgment to the existing annotations in the multi-perspective question answering (mpqa) opinion corpus annotations of subjective expressions. Sentiment expressions are any word used to express anthought, emotion, evaluation, opinion, speculation etc. Annotators were informed to tag the polarity of sentiment expression as positive, negative, both or neutral.

The movie review [14] summarization uses the WordNet, statistical analysis and domain knowledge. The summarization is in the form of each features present in the movie with respect to their positive and negative opinion. Most of the review mining and summarization is concentrate on product reviews. But here focus on different domain called movie review. It has unique characteristics. The user wrote a comment for a particular movie not only a movie element (e.g. screenplay, vision effects, music) and also movie-related people (director, actor and screenwriter).

III . METHODOLOGY

In the Feature based opinion mining summarization consists of four tasks. Each task have different step to deal with their operations. The tasks are feature extraction, mapping sentence, polarity (nature of the opinion) prediction and summarization.

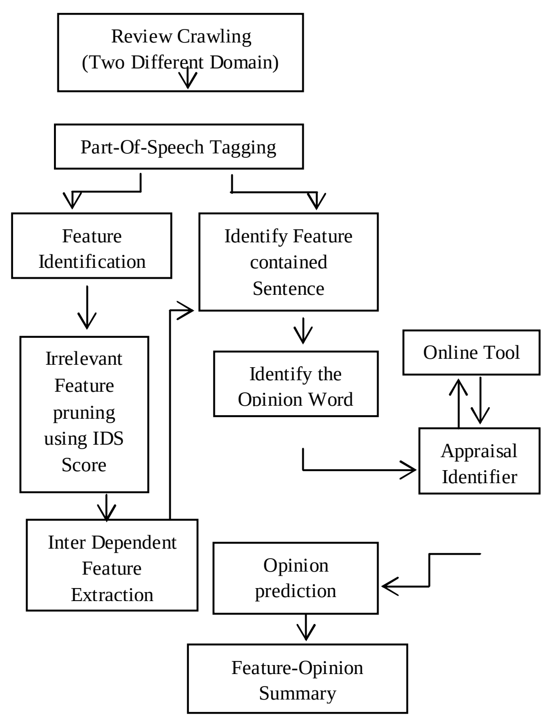

Fig.1 the Overview of Dual Domain Miner

The overview of the Dual Domain Miner (DDM) is explained by fig. 1. Initially the Reviews are crawled from the two different domain corpuses. Use part-of-speech Tagging, the Features and Opinion is identified. The collected features are grouped to form feature list. The valid set of features is extracted from the features list using Inter Dependent Domain Relevance mechanism. The Valid set of Features is used to identify the relevant sentence from the domain corpus. The opinion words are identified using POS tagging. The Identifier is generates the score from the online sentiment prediction tool. The identified opinion word is sent to the identifier from where the opinion is predicted. The Appraisal identifier classifies the opinion word into positive, negative and mixed. Based on the features and predicted opinion the summary has been generated.

A.Feature Extraction

Normally features in the form of noun, it can gather from the reviews using POS tagging tool. The review taken from the Domain corpus is send to the tool, it produce the part-of-speech of all terms with respect to their word like good_JJ, book_NN etc.

Feature List $= \cup \cup \cup \ldots \cup$
New Feature List $= \cup \cup \ldots \cup$
Feature Count $= \cup \cup \ldots \cup$
Feature Count FC_i Corresponds to NF_i

ALGORITHM 1: Eliminate redundant features
Input: A list of Features F
Output: List of New Features NF and Count FC
for each Feature F_i do
set j as increment of i
for each Feature F_j do
if(F_i same as F_j)
remove feature F_j from F
incrementfeature count FC_i of F_i.
addF_i to NF with Corresponding FC_i
incrementNF_i
return NF

The feature extraction have two tasks one is eliminate the redundant features and prune the irrelevant features. After extracting the features, the review contain the features may occur more than once.so we have to consolidate the features in the given reviews using Alg. 1.

ALGORITHM 2: Valid Feature Extraction
Input: A list of features NF Feature Count FC
Output: A list of Valid feature VF
for each valid features NF_i do
for each Review R_j in Domain corpus
calculate weight Tw_{ij} by (1)
calculateScatterWholeSw_i by (4)
calculateScatterInSI_i by (6)
calculate the InterDependent Score IDS_i by (7)
if($IDS_i \geq$threshold)
addNF_i to VF
return VF

The feature list is now ready to perform the extraction of valid features from the list using Alg. 2. Initially the weight is assigned to each feature in each reviews using (1).

$$= \begin{cases} 1 + \log(1 + \quad (\quad)) & > 0 \\ 0 & h \end{cases} \text{—> (1)}$$

The average weight of the feature across all document is calculated using (2), The Corpus contain X number of reviews.

$$= -\sum \quad \text{--> (2)}$$

$$= \frac{\overline{\sum \ (\qquad)}}{} \text{--> (3)}$$

The ScatterWhole (SW) is how spread the term across all reviews present in the corpus.

$$= \frac{}{} \text{--> (4)}$$

The average weight of the document present in the review corpus is calculated using (5), the Review Contain Y features.

$$= -\sum \quad \text{--> (5)}$$

The ScatterIn (SI) is how each term occur in each review corpus is calculated using (6)

$$= \quad - \quad \text{--> (6)}$$

The features are extracted using inter dependent score (IDS) is calculated using (7).

$$= \quad \times \sum \quad \text{--> (7)}$$

Finally the extraction is based on the threshold value and IDDR score comparison.

B. Mapping sentence

The opinion based on the features is identified using mapping concept. Mapping Sentence is the process to identify the sentence contains the valid feature.

Domain Corpus $= \cup \cup \cup \cup \ldots \cup$
Review $= \cup \cup \cup \ldots \cup$
Sentence $= \cup \cup \cup \ldots \cup$
Valid sentence $= \cup \cup \ldots \cup$

ALGORITHM 3: Mapping Valid Feature with opinion word
Input: Two Different Domain Corpus
Output: A set of Valid sentence VS_i
for each C_i do
for each R_i do
for each S_i do
for each VF_i do
    Match $VF_i$to S_i
If(VF_i present in S_i)
    Add S_i to VS_i
Return VS

Each feature is mapped into whole review corpus. If the sentence is present, then the features are extracted, otherwise leave that sentence. It can explain in Alg. 3.

C. Polarity prediction

After finish mapping, the opinioned word present in the sentences are gathered.

Algorithm 4: Gathering the Opinion word, Connecting word and Negative word
Input: VS_i
Output: OPW_i, NG_i, CC_i
For each VS_i do
For each W_i in VS_i
If (W_i as Adjective and related to Feature)
  Add W_i to OPW_i
Else If (W_i as verb and related to Feature)
  Add W_i to OPW_i
Else If (W_i as negative word)
  Add W_i to NG_i
Else if (W_i as connecting word)
  Add W_i to CC_i
Return OPW_i, NG_i, CC_i

Usually the opinion is in the form of adjective and verb. The opinion words, connecting word, Negative words are gathered using alg. 4.

Algorithm 5: Opinion Identification
Input: NS_i
Output: Nature of Opinion
Step 1: For each NS_i do
  Find OPW_i, NG_i,CC_i from alg 4
  Find score from training data for OPW_i as OPS_i
Step 2: If (NS_i contain single opinion 'OP_i')
  Goto step 6
Step 3: Else if (NS_i contain two opinion word 'OP_{ii}', 'OP_{ij}' with connecting word)
  Goto step 6 for OP_{ii}
  Goto step 7 for OP_{ij}
Step 4: If (NS_i contain NG_i)
  Op = reverse (Top)
 Else
  Op = Top
Step 5: return Op
Step 6: If (OPS between 0 and 50)
  Assign Top as Weakly Positive
 Else if (OPS between 50 and 100)
  Assign Top as Strongly Positive
 Else if (OPS between -50 and 0)
  Assign Top as weakly negative
 Else if (OPS between -100 and -50)
  Assign Top as Strongly Negative
 Else
  Assign Top as Neutral
Goto Step 4
Step 7: If (CC_i as 'and' or 'both')
  Assign Top to OP_{ij}
 Else if (CC_i as 'but' or 'neither or nor')
Assign reverse of Top to OP_{ij} Goto Step 4

The identifier is used to predict the polarity of the opinion words using score from the tool, connecting word and Negative term is explained in Alg. 5 have three steps they are.

1. Opinion score prediction
2. Connecting opinion
3. Negative word comparison

The opinion score is gathered from the training reviews. After gather the score, polarity like positive, negative and mixed are classified. To reduce the classification time, go for the connecting word. The sentence contain two opinioned word, if one of the word is identified, then the opinion word of the another can be predicted using connecting word like and, both, but etc.

If the connecting word is 'and' and 'both' then assign same opinion of identified word to another word or if the word like' but', 'neither or nor', then reverse of first word's opinion. Finally check the presence of negative word. If negative word is not present then assign the same opinion to their feature otherwise reverse the opinion.

Finally the feature with their corresponding like positive, negative and mixedare summarized like below example.

Example 1:
Feature: "Memory"
STRONGLY POSITIVE: 17
Sentence 1: The Memory capacity is Excellent
Sentence 2: I admire the Memory size
…
WEAKLY POSITIVE: 10
Sentence 1: The Memory of the Canon S100 is not bad.
Sentence 2: Canon S100's memory is fair
…
WEAKLY NEGATIVE: 6
Sentence 1: Need improvement in memory size
Sentence 2: Memory size is not enough
…
STRONGLY NEGATIVE: 5
Sentence 1: Memory size is too low.
Sentence 2: Its Memory capacity is very poor.
…

IV . EXPERIMENTAL ANALYSIS

The comparative analysis carried out with the usage of the single domain such as Intrinsic Domain Relevance (IDR), Extrinsic Domain Relevance (EDR) and contrast domain such as Intrinsic-Extrinsic Domain Relevance (IEDR) and Dual Domain Miner (DDM). The domain such as Canon S100 (camera), MicroMp3, Nokia 6600 (Mobile) and iPod are used for the predicting the feature extraction accuracy.

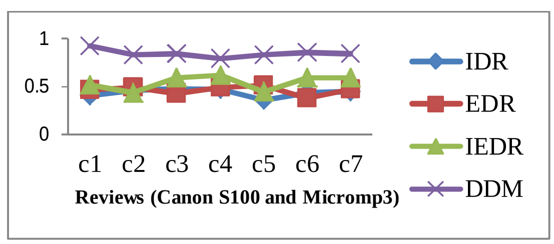

Fig.2 Graph for Camera and Mp3 domain F-Score.

The F-Score value of fig.2 is calculated based on the precision and Recall value of Camera and Mp3 player. The precision and recall proves that the DDM is more efficient this lead to the F-Score also shows that the DDM is more efficient than other three algorithms. The predication capacity of DDM is 60% more than the other three approaches.

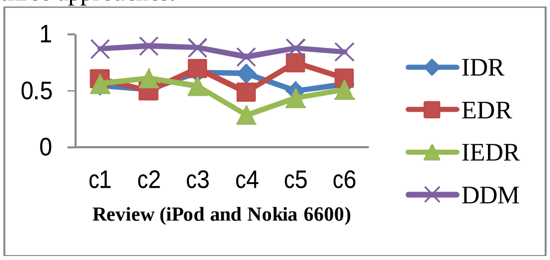

Fig.3 Graph for iPod and Mobile domain F-Score.

Fig. 3 shows The F-Score value of feature extraction for iPod and Mobile domain. The precision and recall proves that the DDM is more efficient, this lead to the F-Score shows that the DDM is more efficient than other three algorithms. The predication capacity of DDM is 50% more than the other three approaches.

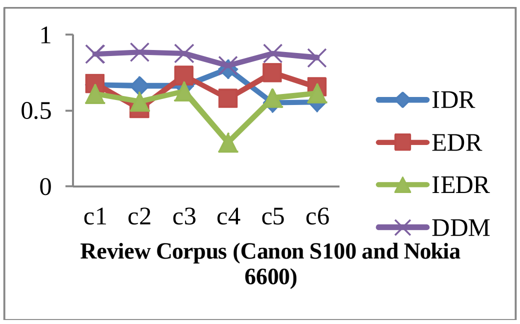

Fig.4 Graph for Camera and Mobile domain F-Score.

Fig. 4 shows The F-Score value of feature extraction for iPod and Mobile domain. DDM is more efficient than other three algorithms like IDR, EDR and IEDR. The predication capacity of DDM is 50% more than the other three approaches.

V . CONCLUSION

The Dual Domain Miner is summarized the feature-opinion pair in two different domain at the same time. Two different domain corpuses are considered and operate the tasks simultaneously in two corpuses. The IDDR algorithm is for efficient feature extraction method using two different tasks to perform their job. The removal of redundant features is eliminating feature with counting the occurrence of the features and pruning of irrelevant features using IDDR score. The IDDR algorithm is much better than existing single domain feature extraction in feature based opinion mining. The polarity prediction using appraisal identifier is simpler and effective using three approaches like scoring, connectivity and negation. Finally the summarization gives more effective explanation about the two different domain corpuses. The future enhance involve using more than two different corpus to extract the features. The implicit feature extraction is added to the feature extraction. The polarity like midly positive, midly negative, both also included.

REFERENCES

[1] A. Esuli and FabrizioSebastiani.. “SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining”. *In Proceedings of LREC*. 2006.

[2] Zhen Hai, Kuiyu Chang, Jung-Jae Kim, and Christopher C. Yang “Identifying Features in Opinion Mining via Intrinsic and Extrinsic Domain Relevance,” *ieee transactions on knowledge and data engineering*, vol. 26, no. 3, march 2014.

[3] A. Popescu and O. Etzioni, “Extracting Product Features and Opinions from Reviews,” *Proc. Human Language Technology Conf.and Conf. Empirical Methods in Natural Language Processing*, pp. 339-346, 2005.

[4] R. Abirami, S. Ramesh “Extracting Features in Opinion Mining using Inter Dependent Domain Relevance”, *Proc. National Conference on Intelligence Computing-15*, pp 115-120, 2015.

[5] P.D. Turney, “Thumbs Up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews,” *Proc. 40th Ann. Meeting on Assoc. for Computational Linguistics*, pp. 417-424, 2002.

[6] F. Li, C. Han, M. Huang, X. Zhu, Y.-J. Xia, S. Zhang, and H. Yu, “Structure-Aware Review Mining and Summarization,” *Proc. 23rd Int’l Conf. Computational Linguistics*, pp. 653-661, 2010.

[7] S.-M. Kim and E. Hovy, “Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text,” *Proc. ACL/COLING Workshop Sentiment and Subjectivity in Text*, 2006.

[8] M. Hu and B. Liu, “Mining and Summarizing Customer Reviews,” *Proc. 10th ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining*, pp. 168-177, 2004.

[9] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up?: Sentiment Classification Using Machine Learning Techniques,” *Proc. Conf. Empirical Methods in Natural Language Processing*, pp. 79-86, 2002.

[10] L. Qu, G. Ifrim, and G. Weikum, “The Bag-of-Opinions Method for Review Rating Prediction from Sparse Text Patterns,” *Proc. 23rd Int’l Conf. Computational Linguistics*, pp. 913-921, 2010.

[11] T. Wilson, J. Wiebe, and P. Hoffmann, “Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis,” *Proc. Conf. Human Language Technology and Empirical Methods*

in Natural Language Processing, pp. 347-354, 2005.

[12] B. Liu, “Sentiment Analysis and Opinion Mining,” *Synthesis Lectures on Human Language Technologies,* vol. 5, no. 1, pp. 1-167, May 2012.

[13] F. Fukumoto and Y. Suzuki, “Event Tracking Based on Domain Dependency,” *Proc. 23rd Ann. Int’l ACM SIGIR Conf. Research and Development in Information Retrieval*, pp. 57-64, 2000.

[14] L. Zhuang, Feng Jing and Xiaoyan Zhu. “Movie Review Mining and Summarization.” *In Proceedings of CIKM* 2006.

[15] K. Dave, S. Lawrence & D. Pennock. “Mining the peanut gallery: opinion extraction and semantic classification of product reviews.” *WWW*’2003.

[16] Pang, B. and Lee, L. “A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts.” *In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (Barcelona, Spain, July 21 - 26, 2004). Annual Meeting of the ACL. Association for Computational Linguistics, Morristown, NJ,* 271, 2004.

[17] Whitelaw, C., Garg, N., and Argamon, S. “Using appraisal groups for sentiment analysis*.” In Proceedings of the 14th ACM international Conference on information and Knowledge Management (Bremen, Germany, October 31 - November 05, 2005). CIKM '05. ACM, New York, NY*, 625- 631, . 2005.

[18] N. Jakob and I. Gurevych, “Extracting Opinion Targets in a Single and Cross-Domain Setting with Conditional Random Fields,” *Proc. Conf. Empirical Methods in Natural Language Processing*, pp. 1035-1045, 2010.

[19]] V. Hatzivassiloglou and J.M. Wiebe, “Effects of Adjective Orientation and Gradability on Sentence Subjectivity,” *Proc. 18th Conf. Computational Linguistics*, pp. 299-305, 2000.

[20] R. Mcdonald, K. Hannan, T. Neylon, M. Wells, and J. Reynar, “Structured Models for Fine-to-Coarse Sentiment Analysis,” *Proc. 45th Ann. Meeting of the Assoc. of Computational Linguistics*, pp. 432- 439, 2007.

[21] D. Bollegala, D. Weir, and J. Carroll, “Cross-Domain Sentiment Classification Using a Sentiment Sensitive Thesaurus,” *IEEE Trans. Knowledge and Data Eng*., vol. 25, no. 8, pp. 1719-1731, Aug. 2013.

[22] S.J. Pan and Q. Yang, “A Survey on Transfer Learning,” *IEEE Trans. Knowledge and Data Eng.*, vol. 22, no. 10, pp. 1345-1359, Oct. 2010.

[23] T. Dunning, “Accurate Methods for the Statistics of Surprise and Coincidence,” *Computational Linguistics*, vol. 19, no. 1, pp. 61-74, Mar. 1993.

[24]] W. Jin and H.H. Ho, “A Novel Lexicalized HMM-Based Learning Framework for Web Opinion Mining,” *Proc. 26th Ann. Int’l Conf. Machine Learning*, pp. 465-472, 2009.

[25] Landauer, T.K., &Dumais, S.T. “A solution to Plato’s problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge,” *Psychological Review*, 104, 211-240. 1997.

[26] Santorini, B. “Part-of-Speech Tagging Guidelines for the Penn Treebank Project” (3rd revision, 2nd printing). *Technical Report, Department of Computer and Information Science, University of Pennsylvania.* 1995.

[27] Spertus, E. “Smokey: Automatic recognition of hostile messages,” *Proceedings of the Conference on Innovative Applications of Artificial Intelligence*(pp. 1058-1065). Menlo Park, CA: AAAI Press. 1997.

[28] Tong, R.M. “An operational system for detecting and tracking opinions in on-line discussions,” *Working Notes of the ACM SIGIR 2001 Workshop on Operational Text Classification* (pp. 1-6). New York, NY: ACM. 2001.

[29] Turney, P.D. “Mining the Web for synonyms: PMI-IR versus LSA on TOEFL,” *Proceedings of the Twelfth European Conference on Machine Learning*(pp. 491-502). 2001.

[30] John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and JennWortman. “Learning bounds for domain adaptation,” *In Annual Conference on Neural Information Processing Systems 20*, pages 129–136. MIT Press, 2008.

[31] John Blitzer. “Domain Adaptation of Natural Language Processing Systems,” *PhD thesis, The University of Pennsylvania*, 2007.

[32] Steffen Bickel, ChristophSawade, and Tobias Scheffer. “Transfer learning by distribution matching for targeted advertising,” *In Advances in Neural Information Processing Systems 21*, pages 145–152. 2009.

[33] Tong, R.M. “An operational system for detecting and tracking opinions in on-line discussions,” *Working Notes of the ACM SIGIR 2001 Workshop on Operational Text Classification*(pp. 1-6). New York, NY: ACM. 2001.