Generate ratings and recommendations from E-Commerce and Boarding House reviews using sentiment classification

Divya V[1], Durgadevi K[2],
Sathish Saravanan P[3],
Assistant Professor
Department of Information Technology, Dhanalakshmi College of Engineering,
Chennai, Tamil Nadu, India.

Abstract: To adapt a sentiment classifier trained on a particular domain (source domain), to a different domain (target domain), we develop a sentiment classifier which extracts aspects from reviews and analyze the sentiment sensitive embedding and rate it based on the arousal. We demonstrate the sentiment classification for two different domains with a single classifier without any training on target domains. The Domain Thesaurus on particular target domain is build and the classifier is able to give the needed results hassle free. The implementation uses Natural Language Processing(NLP) techniques for extracting aspects and classifies the aspects based on the Target Domain using Domain Thesaurus. Valence and Arousal will be calculated to calculate rating for the particular aspects in the user review.

Index Terms: Cross Domain, Domain Thesaurus, NLP, POS tagging.

I. INTRODUCTION

Sentiments expressed in user-reviews about a particular product must be correctly identified for many purposes. The main purpose is when a particular feature of a product is associated with a negative sentiment then the manufacturer can immediately work on to rectifying that particular issue. Failing to identifying or detecting a negative sentiment of a product feature may result in decreased sales. Considering online stores where one cannot physically touch a product as in real-world store, the customer review are the only means of evaluating a product.

By automatically classifying the user-reviews based on the sentiment expressed in them, we can assist potential buyers of a product to easily understand the overall opinion about the product[1]. Considering the numerous applications of sentiment classification such as opinion mining[2], opinion summarization[3], contextual advertising[4], and market analysis[5], it is not surprising that sentiment classification has received continuous attention.

Sentiment classification is an instance of text classification where a document is classified into to a predefined set of sentiment classes [6]. Here document refers to various types of user reviews. A review is classified into two classes depending on whether it expresses positive or a negative sentiment towards a product. A review can be assigned a discrete sentiment score (e.g. from one to five stars) that indicates the degree of the positivity or negativity of the sentiment. Aspect is extracted from a review which has been identified as sentiment bearing by performing further analysis. As vast number of products is sold online, it is costly and infeasible to manually annotate reviews for every product.

The Cross-Domain Sentiment Classification is an attractive way of adapting a sentiment classifier that is trained using labeled reviews for one product to classify sentiment on a different product. Consider the following situation where we have trained a sentiment classifier using

labeled reviews for electronic products and would like to apply it to classify sentiment on Boarding House domain. Here the term domain refers to a collection of reviews written on a particular product. Source domain is the domain from which we train our sentiment classifier and the target domain is the domain to which we apply the trained classifier. Words such as *reliable, durable, working* properly are used to express sentiment about electronic products, whereas words such as *clean, cheap, comfortable* are used to express sentiment about boarding houses. Unfortunately, this mismatch of features between the source and target domains causes a sentiment classifier trained on electronic products to perform poorly when applied to boarding houses.

II. RELATED WORK

Cross-Domain Sentiment Classification is classified as unsupervised and supervised methods. Structural Correspondence Learning (SCL) [7] selects a set of pivots (features that represent to the same sentiment in both source and target domains are referred to as pivots in literature), using some criteria. By first predicting the pivots as additional features, SCL attempts to reduce the mismatch between features in source and target domains.

Spectral Feature Alignment (SFA) [8] splits feature space into two mutually exclusive groups: domain independent features and domain specific features. Another model, Sentiment Sensitive Thesaurus (SST) [9][10] that lists words that express similar sentiments in the source and target domains. Consider the example, where SST created for the two domains electronic products and boarding house lists *working properly* as a related word for *comfortable*. The thesaurus is automatically created using a sentiment sensitive asymmetric similarity measure that uses sentiment labels in the source domain documents. Analogous to the thesauri-based query expansion in information retrieval, SST is used to expand the source domain feature vectors by appending related features in the target domain. A binary logistic regression classifier is trained using the expanded feature vectors corresponding to the source domain labeled documents. Unlike, SCL or SFA, SST does not create lower-dimensional embedding.

III. METHODS AND MATERIALS

We propose a cross domain sentiment classifier which can adapt to different variety of domains without need to do much training on target domain. We build domain thesaurus which can easily classify the aspects and sentiments associated with it. Our proposed method is different from SCL and SFA in that, we consider not only the unlabeled data but also labeled data for the source domain when constructing the representation.

We demonstrate the sentiment classification for two different domains with a single classifier without any training on target domains. We extend the SST model proposed earlier to build the domain thesaurus on particular target domains and out classifier is able to give the needed results hassle free. The implementation uses Natural Language Processing t e c h n i q u e s for extracting aspects and uses the domain thesaurus to classify the aspects based on the target domains. Valence and Arousal will be calculated to calculate rating for the particular aspects in the user review. We use product reviews as well as boarding house reviews for implementation of which boarding house domain sentiment classification can be extended to give service recommendation to users based on their requirements.

A user-based Collaborative Filtering (CF) algorithm is adopted to generate appropriate recommendations and sorted based on Bubble Sort Algorithm. It aims at computing a personalized rating of each candidate service for a user, and then providing a personalized boarding house service recommendation list and

recommending the most appropriate boarding house services to him/her.

1. POS Tagging for User Reviews

Huge collection of data is retrieved from open source datasets that are publicly available from web applications like Trip Advisor and Amazon. The data's are in CSV (Comma separated values) or TSV (Tab Separated Values) format. The CSV files were read and manipulated using Java API which is developer friendly, light weighted and easily modifiable.

The user review for two different domains were loaded as a CSV or TSV file, parsed using API and then each review by each customer is processed sequentially. The reviews were given one by one to POS Tagger which splits each word in the review and tags it based on the Parts of Speech the word belongs.

2. Chunking the Reviews and Aspect Extraction

Chunker Process is done on each and every review of all and the products. The Chunker Process will take POS tagged output as input for grouping the words based on meaning of the review. Chunker Process is done so that we can easily extract the sentiment embedding associated with the aspects of the particular review. The meaningful words that should be read continuously for proper understanding of the review are marked with square bracket. Now the Aspects in each review are extracted from the POS Tagger result.

The Noun and Phrasal Verbs are the key attributes in any sentence. So those things were extracted from the tagged reviews and marked as aspects of the particular review by a user. Now mappings are done to properly annotate the user review and associated aspects with the Chunks in it.

3. Building Domain Thesaurus on Target Domains

A domain thesaurus is built depending on the Keyword Candidate List and Candidate Services List. Keyword Candidate List and Candidate Services List are interdependent on the target domains and it can be prepared before porting the classifier to target domain. Expert knowledge should be given for preparing the domain thesaurus. The Domain thesaurus can be updated regularly to get accurate results of the recommendation system. Now the aspects extracted are subjected to domain grouping based on the target domain.

4. Sentiment C l a s s i f i c a t i o n a n d S e r v i c e Recommendation

The chunked reviews of the user are retrieved and the keywords (Aspects) corresponding to the user is analyzed for its Valence and Arousal. Valence can be defined as whether the keywords refers a *Positive* or *Negative* aspect and Arousal answers, how much *P o s i t i v e* or *Negative* it is. Ratings are given for each domain in Target based on the Valence and Arousal for each user of each review.

For product reviews the overall rating is now manipulated by averaging values of each rating of several users of a particular product. In boarding house domain, we extend rating to give personalized service recommendation to user based on requirements to user. Ranking is done for all boarding houses based on ratings by similar users using CF(Collaborative Filtering) and will be sorted based on Bubble Sorting Algorithm to have the most appropriate personalized recommendation for the user.

IV. RESULTS AND DISCUSSIONS

The Natural Language Processing is

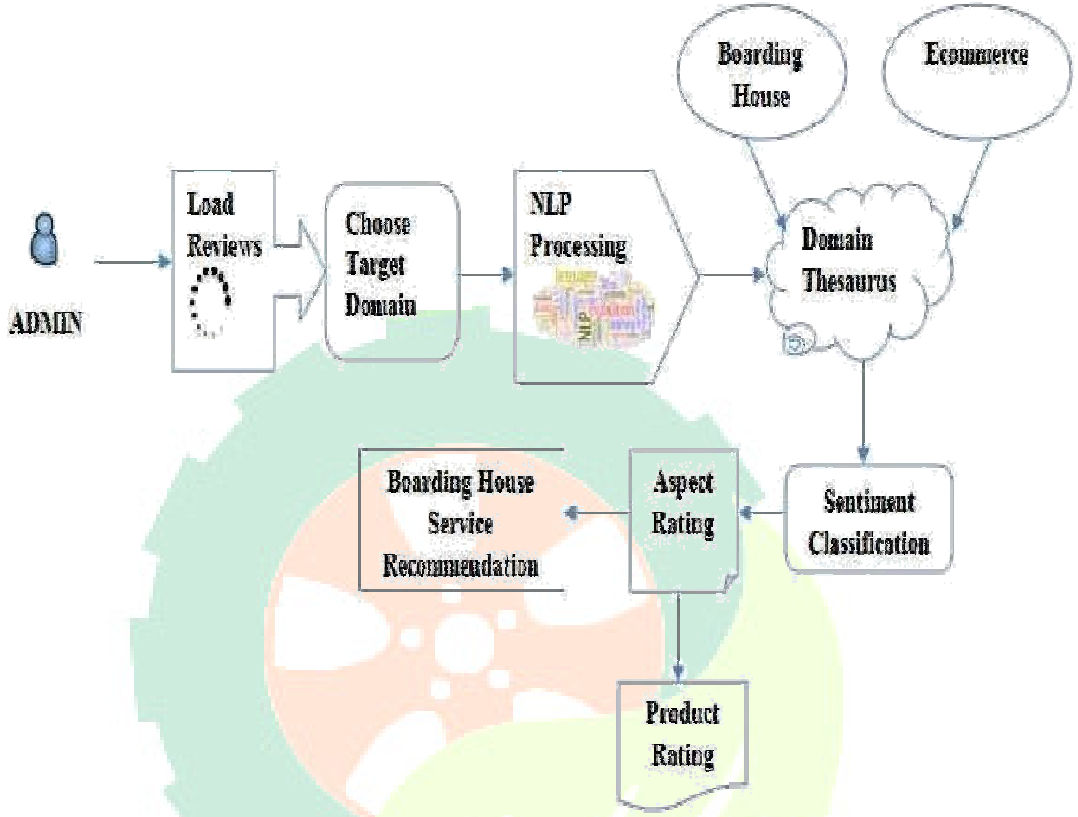

Figure 1: Architecture Diagram

implemented to analyze the reviews of the previous user. The Natural Language Process comprises of tokenizing a sentence or a word, POS (Parts of Speech) Tagging, extraction of Nouns and Verbs, synonym retrieval and spell check of extracted keywords using WordNet dictionary. The aspects extracted are subjected to domain grouping based on the target domain. Valence and Arousal is implemented for calculating the ratings of E-Commerce and Boarding House. Based on ratings service recommendation is provided for boarding house. The BigData manipulations from Comma Separated Value(CSV) through our own JAVA API enforce developer friendly access.

V. CONCLUSION

Sentiments expressed in user-reviews about a particular product in online stores are the only means of evaluating a product. This paper presents a cross domain sentiment classifier which can adapt to different variety of domains without need to do much training on target domain. The Domain Thesaurus on particular target domain is build and the classifier is able to give the needed results hassle free. The implementation uses Natural Language Processing(NLP) techniques for extracting aspects and classifies the aspects based on the target domain using domain thesaurus. Valence and Arousal will

be evaluated to calculate rating for the particular aspects in the user review.

REFERENCES

[1]Danushka Bollegala, Tingting Mu, John Yannis Goulerma, "Cross-Domain Sentiment Classification Using Sentiment Sensitive Embeddings", IEEE Transactions on Knowledge and Data Engineering, VOL. 28, NO. 2, February 2016.

[2]B.Pang and L.Lee, "Opinion mining and sentiment analysis," found. Trends Inf. Retrieval, vol 2, nos.1/2, pp. 1-135,2008.

[3] Y.Lu , C. Zhai, and N.Sundaresan, "Rated aspect summarization of short comments," in Proc.18th Int. Conf. World Wide Web,2009,pp. 131–140.

[4] T.-K. Fan and C.-H. Chang, "Sentiment-oriented contextual advertising," Knowl.Inf. Syst., vol. 23, no.3, pp. 321–344, 2010.

[5] M. Hu and B. Liu , "Mining and summarizing customer reviews," in Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2004, pp. 168-177.

[6] C. D. Manning and H. Sch€utze, Foundations of Statistical Natural Language Processing. Cambridge, MA, USA : MIT Press ,2002.

[7] J. Blitzer, M. Dredze, and F. Pereira, " Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification," in Proc. 45th Annu. Meeting Assoc. Comput. Linguistics, 2007, pp.440-447.

[8] S. J. Pan, X. Ni, J-T. Sun, Q. Yang, and Z.Chen,"Cross-domain sentiment classification via spectral feature alignment ," in Proc. 19th Int. Conf. World Wide Web,2010, pp.751-760.

[9] Danushka Bollegala, D. Weir, and J. Carroll, "Using multiple sources to construct a sentiment sensitive thesaurus for cross-domain sentiment classification," in Proc. 49th Annu. Meet. Assoc. Comput. Linguistics: Human Language Technol., 2011, pp. 132- 141.

[10] Danushka Bollegala, D. Weir, and J. Carroll, "Cross-domain sentiment classification using sentiment sensitive thesaurus," IEEE Transaction on Knowledge and Data Engineering, vol. 25, no. 8, pp. 1719-1731, Aug. 2013.