International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 3 | ISSN : 2456-3307

Cross Domain Sentiment Classification Using Natural Language Processing

S. Vidya
Assistant Professor, Department of Computer Science and Engineering, Kalasalingam Institute of Technology, Krishnankoil, Tamil Nadu, India

ABSTRACT

The field of study that focuses on the interactions between human language and computers is called Natural Language Processing. It sits at the intersection of computer science, artificial intelligence and Natural Language Processing.NLP can be a approach for computers to analyze, understand, and derive meaning from human language in an exceedingly good and helpful way. By utilizing NLP, developers can organize and structure information to perform tasks like automatic summarization, translation, named entity recognition, relationship extraction, sentiment analysis, speech recognition, and topic segmentation.NLP systems have long crammed useful roles, like correcting descriptive linguistics, changing speech to text and automatically translating between languages.

Keywords: NLP, sentiment classification, domain thesaurus

I. INTRODUCTION

The ability to correctly identify the sentiment expressed in user-reviews about a particular product is an important task for several reasons. First, if there is a negative sentiment associated with a particular feature of a product, the manufacturer can take immediate actions to address the issue. Failing to detect a negative sentiment associated with a product might result in decreased sales.

From the users point-of-view, in online stores where one cannot physically touch and evaluate a product as in a real-world store, the user opinions are the only available subjective descriptors of the product. By automatically classifying the user-reviews according to the sentiment expressed in them, we can assist the potential buyers of a product to easily understand the overall opinion about that product.

Considering the numerous applications of sentiment classification such as opinion mining, opinion summarization, contextual advertising, and market analysis, it is not surprising that sentiment classification has received continuous attention.

Sentiment classification can be considered as an instance of text classification where a given review must be classified into a pre-defined set of sentiment classes. In binary sentiment classification, a review must be classified into two classes depending on whether it expresses a positive or a negative sentiment towards an entity.

Figure 1. Natural Language Processing

II. PROBLEM DEFINITION

- Considering the vast number of products sold online, it is both costly as well as infeasible to manually annotate reviews for each product type.
- Existing Systems fails to adapt a sentiment classifier that is trained using labeled reviews for one product to classify sentiment on a different product.

A drawback of this two-step approach that decouples the embedding learning and sentiment classifier training is that the embeddings learnt inside the primary step is agnostic to the sentiment of the documents, that is the final goal in cross-domain sentiment classification

III. SYSTEM DESIGN

3.1 POS TAGGING OF USER REVIEWS

Part of speech tagging additionally known as grammatical tagging or word category clarification, is that the method of marking up a word in a text (corpus) as corresponding to a selected part of speech, supported each its definition and its context its relationship with adjacent and connected words in a phrase, sentence, or paragraph. A simplified type of this is often taught to school-age children, in the identification of words as nouns, verbs, adjectives, adverbs, etc.

Part-of-speech tagging is tougher than just having a listing of words and their parts of speech, because some words will represent over one part of speech at different times, and since some parts of speech are complex or unspoken. This is not rare in natural languages, an outsized proportion of word-forms are ambiguous.

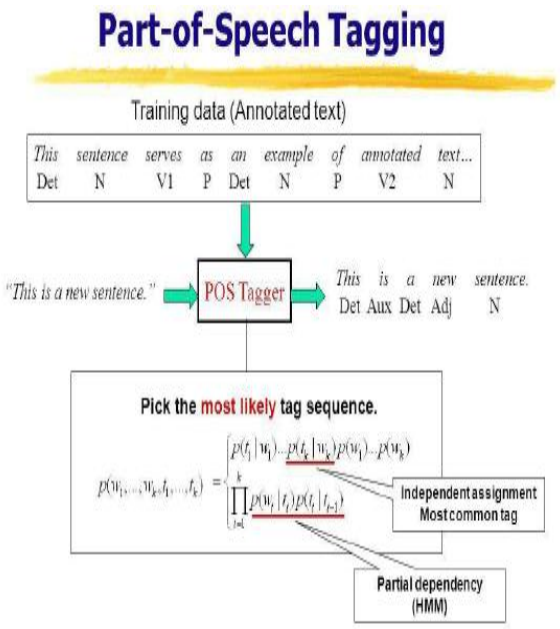

Figure 2. POS tagging of user reviews

3.2 CHUNKING THE REVIEWS

This section deals with the chunk tags. Not several of the problems mentioned above hold for outlining the chunk tags, based on the above definition of chunk, problems associated with various chunk types were mentioned. A chunk would contain a 'head' and its modifiers.

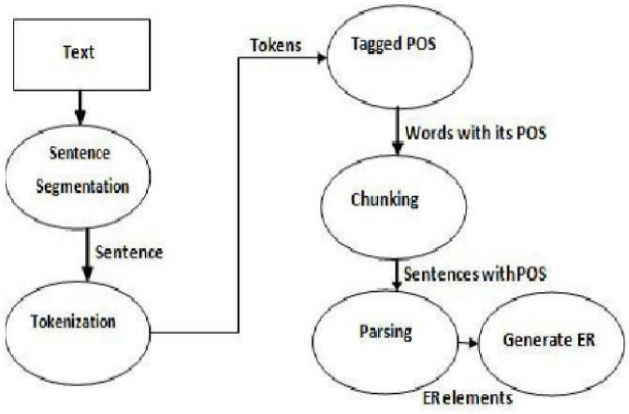

Figure 3. Chunking reviews and aspect extraction

3.3 BUILDING DOMAIN THESAURUS ON 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 groping based on the target domain.

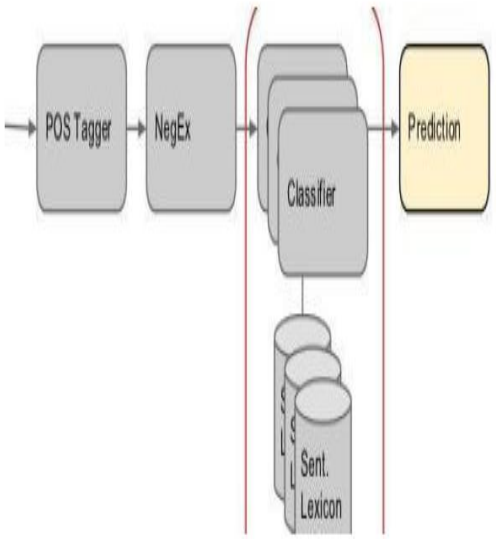

Figure 4. Building Domain Thesaurus

3.4 SENTIMENT CLASSIFICATION AND SERVICE RECOMMENDATION

The implementation uses Natural Language Processing Techniques for extracting aspects and uses the Domain Thesaurus to classify the Aspects based on the Target Domains. Valance and Arousal will be calculated to calculate rating for the particular aspects in the user Review. We use Product Reviews as well as Hotel Reviews for Implementation

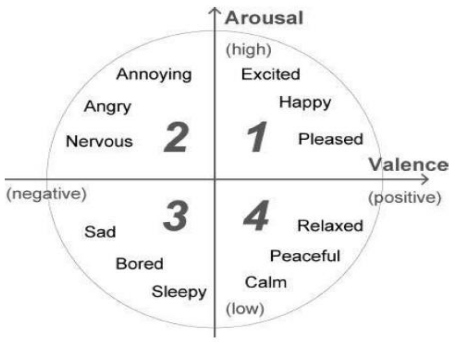

Figure 5. Sentiment Classification

IV. SYSTEM ARCHITECTURE

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.

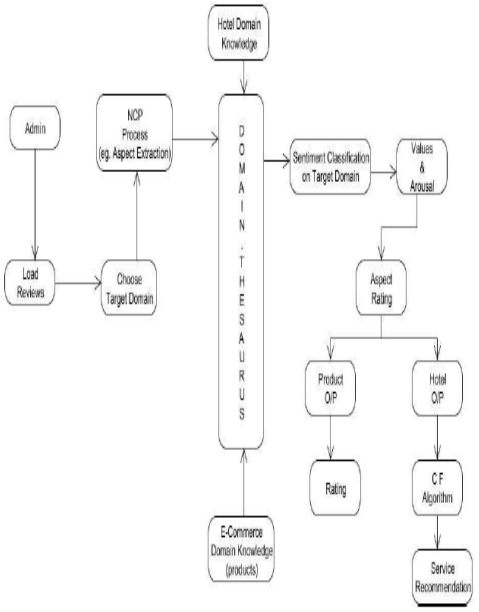

Figure 6. Architecture of NLP Process

In Hotel Domain we extend ranking to give Personalized Service recommendation to user based on requirements to user. Ranking is done for all hotels based on Ratings by similar users using CF (Collaborative Filtering) and 20 will be sorted based on Bubble Sort Algorithm to have the most appropriate personalized recommendation for the User.

Sentiment classification can be considered as an instance of text classification where a given review must be classified into a pre-defined set of sentiment classes. In binary sentiment classification, a review must be classified into two classes depending on whether it expresses a positive or a negative sentiment towards an entity.

V. MODULES DESCRIPTION

5.1 POS TAGGING OF 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 or TSV Format.

The CSV(Comma separated values) files were read and manipulated using Java API that itself developed by us 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.

5.2 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.

5.3 BUILDING DOMAIN THESAURUS ON TARGET DOMAIN

A Domain Thesaurus is made looking on the Keyword Candidate List and Candidate Services List. Keyword Candidate List and Candidate Services List are interdependent on the target domains and it will be ready before porting the classifier to target domain. Expert Knowledge should be given for preparing the domain Thesaurus.

5.4 SENTIMENT CLASSIFICATION AND SERVICE RECOMMENDATION

The Chunked Reviews of the user is retrieved and the Keywords (Aspects) corresponding to the user is analyzed for its Valence and Arousal. Valence Means whether the Keywords means a positive or Negative thing and arousal answers, how much it is?. Ratings are given for each domain in Target based on the Valence and Arousal for each User of each review.

In Hotel Domain we extend ranking to give Personalized Service recommendation to user based on requirements to user. Ranking is done for all hotels based on Ratings by similar users using CF (Collaborative Filtering) and will be sorted based on Bubble Sort Algorithm to have the most appropriate personalized Recommendation for the User.

VI. SYSTEM IMPLEMENTATION

6.1 NATURAL LANGUAGE PROCESSING ALGORITHM

NLP algorithms are generally based on machine learning algorithms. Rather than hand-coding huge sets of rules, Natural Language Processing will support on machine learning to automatically learn these rules by analyzing a collection of examples and creating a statically inference. In general, a lot of data is analyzed

6.2 COLLABORATIVE FILTERING ALGORITHM

CF algorithm is adopted to get acceptable recommendations. It aims at calculating a personalized rating of every candidate service for a user, and then presenting a personalized service recommendation list and recommending the foremost acceptable services to him/her.

Collaborative filtering is methodology of creating automatic predictions concerning the interests of a user by collecting preferences or taste information from several users (collaborating).

VII. COMPARISON GRAPH

A comparison graph is drawn between the patterns and natural language processing.

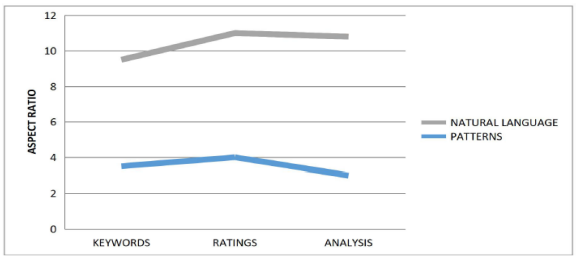

Figure 7. Patterns VS Natural Language Processing

DATA	PATTERNS	NATURAL LANGUAGE
Keywords	3.5	6
Ratings	4	7
Analysis	3	7.8

Figure 8.Patterns and Natural Language Processing

The proposed system of Natural language is described in the form of graph as Keywords, Ratings and Analysis are high when compared to the existing system of Patterns.

VIII. CONCLUSION

We considered three constraints that must be satisfied by an embedding that can be used to train a cross domain sentiment classification method. We evaluated the performance of the individual constraints as well as their combinations using a benchmark dataset for cross domain sentiment classification. Our experimental results how that some of the combinations of the proposed constrains obtain results that are statistically comparable to the current state-of-the-art methods for cross-domain sentiment classification. Unlike previously proposed embedding earning approaches for cross-domain sentiment classification, our proposed method uses the label information available for the source domain reviews, thereby learning embeddings that are sensitive to the final task of application, which is sentiment classification. In our future work, we will do further research in how to deal with the case where term appears in different categories of a domain thesaurus from context and how to distinguish the positive and negative preferences of the users from their reviews to make the predictions more accurate.

IX. FUTURE ENHANCEMENT

The Natural Language Processing is implemented to analyze the reviews of the previous user. The NLP process comprises 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.Valence and Arousal is implemented for calculating Ratings of Aspects of a Hotel. The BigData manipulations from CSV through Our Own JAVA API enforce developer friendly access.

X. REFERENCES

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

[2]. 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.

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

[4]. 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.

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

[6]. H. Daume III, A. Kumar, and A. Saha, "Co-regularization based semi-supervised domain adaptation," in Proc. Adv. Neural Inf. Process. Syst. 23, 2010. pp. 478-486.

[7]. D. Lopez-Paz, J. M. Hernandez-Lobato, and B. Scholkopf, "Semisupervised domain adaptation with non-parametric copulas," in Proc. Adv. Neural Inf. Process. Syst. 25, 2012, pp. 674-682.

[8]. H. Daume III, "Frustratingly easy domain adaptation," in Proc.45th Annu. Meeting Assoc. Comput. Linguistics, 2007, pp. 256-263.

[9]. J. Blitzer, R. McDonald, and F. Pereira, "Domain adaptation with structural correspondence learning," in Proc. Conf. Methods Natural Language Process., 2006, pp. 120-128.

[10]. J. Blitzer, M. Dredze, and F. Pereira, "Biographies, bollywood,boom-boxes and blenders: Domain adaptation for sentiment classification