**ASPECT BASED SENTIMENT ANALYSIS IN HINDI**

|  |  |  |
| --- | --- | --- |
| Name | USC Email | USC ID |
| Abijith Sankar Kottappurath Narayanan | [kottappu@usc.edu](mailto:kottappu@usc.edu) | 6701329807 |
| Nikhil Pachpande | [pachpand@usc.edu](mailto:pachpand@usc.edu) | 2957959351 |
| Nisargkumar Patel | [nisargkp@usc.edu](mailto:nisargkp@usc.edu) | 3713454420 |
| Atharva Kale | [atharvak@usc.edu](mailto:atharvak@usc.edu) | 9865707756 |

**Division of labor**

* Nisargkumar Patel: Aspect Term Extraction using CRF, Dataset preparation
* Nikhil Pachpande: Rule Based Aspect Term Extraction, Aspect Term Sentiment Analysis
* Abijith Sankar KN: Aspect Category Detection, Aspect Category Sentiment Analysis
* Atharva Kale: Results Evaluation

**Word count (excluding references)**

**Introduction**

In recent years, millions of people are into social media networking, blogs and review sites, where they express opinions of various entities including movies, restaurants, products etc. These websites provide myriad amount of information that are not only useful to the creators of these entities, but also to their customers and rivals. The field of Sentiment Analysis collects and analyze the opinions expressed on various of these entities to make decisive recommendations. Rather than considering the overall polarity of a review expressed by a customer, it would be more helpful to analyze feature specific opinions, since people tend to have mixed opinions on the various aspects of a single entity. Consider the sentence. *The bread is top notch as well*. In this sentence, the customer has expressed his opinion on the aspect, *bread*, with a *positive* sentiment and the review belongs to the *food* category having a *positive* sentiment. Hence, it is of utmost importance to extract these feature specific sentiments and their categories to provide a detailed level of recommendation. Internet has now gone beyond languages as people now have the privilege to express their opinion in their preferred language. With the advent of Unicode and the related developments in web and mobile technologies, the user-generated content in Hindi-the 3rd most spoken language in the world-on the internet is growing rapidly and provides an incentive for the sentiment analysis research of such texts. Our project aims at developing and analyzing Aspect Based Sentiment Analysis models on Hindi reviews, which is an area where very few research attempts have been made till date. In our work, we perform Aspect Based Sentiment Analysis with four subtasks:

* Aspect Term Extraction(ATE): Identifying the various aspects of the entity mentioned in the review
* Aspect Term Sentiment Analysis(ATS): Classifying the sentiment expressed towards each aspect
* Aspect Category Detection(ACD): Classification of the aspect terms into related review categories
* Aspect Category Sentiment Analysis(ACS): Classification of the category sentiments.

Multiple approaches are followed for each subtask and it enables us to make a comparative study on potential methods for the tasks. Some of the challenges pertaining to the implementation of this project is the non-availability of any benchmark setup that could provide a high-quality dataset, baseline model as well as the proper evaluation metrics for the system. We intend to improve upon the performance of the state-of-the-art models for this task.

**Method**

**Materials**

There are two main datasets used for the 4 subtasks in the project. Both the datasets were annotated by Indian Institute of Technology, Patna and is a publicly available research repository. Both the files have about 5400 review sentences distributed across 12 domains. The xml files are parsed into csv files with aspect term dataset having the following schema:

*domainID, review, termsFrom, termsTo, polarity, terms*

and aspect category dataset having the following schema:

*domainID, categoryPolarity, review, aspectCategory, aspectTermPolarity*

Every review can have more than one aspect terms with each aspect terms having their corresponding sentiments. Every review can fall into multiple aspect categories with each category having their corresponding sentiment. As a multi-label classification problem, the reviews fall into different combinations of categories like software, hardware, performance, pricing, scenery, place, ease\_of\_use etc. depending upon the review domain. There were three sentiment labels used for sentiment classification: positive, negative and neutral.

An instance of the aspect term dataset with its transliterated and translated versions are given below:

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Review | Aspect Term | Polarity |
| Devanagri | इसकी ऑडियो गुणवत्ता शानदार है | ऑडियो गुणवत्ता | Positive |
| Transliterated | isakee odiyo gunavatta shaanadaar hai | odiyo gunavatta | Positive |
| Translated | Its audio quality is superb | Audio quality | Positive |

An instance of the category detection dataset along with its transliterated and translated versions are given in the table.

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Review | Aspect Category | Polarity |
| Devanagri | इसकी ऑडियो गुणवत्ता शानदार है | Hardware | Positive |
| Transliterated | isakee odiyo gunavatta shaanadaar hai | Hardware | Positive |
| Translated | Its audio quality is superb | Hardware | Positive |

**Procedure**

Both datasets are divided into training and testing sets for the modeling and testing purposes with 80% of the original data used for training.

Aspect Term Extraction:

Aspect Term Sentiment Analysis:

* + Support Vector Machine Approach:
    - Training: To train the classifier, we used the following features:
      * Aspect Term’s Local Context:
        + We observe that the sentiment bearing words usually occur close to the aspect terms.
        + Hence, we look at the preceding and succeeding tokens of the aspect term and use this as the context window for training our classifier.
        + We used context window size of 10 for training.
      * Bigrams:
        + We trained our classifier on unigrams as well as bigrams from the context window.
    - Learning:
      * We used CountVectorizer with a low min\_df as the document vectorizer for building the model.
    - Classification:
      * We used linear kernel SVC as our classifier.
  + Lexicon Based Approach:
    - The lexicon consisted of Hindi words and their respective positive and negative polarity scores.
    - If the positive score of the defining word is greater than its negative score, then the polarity is determined to be positive or vice-versa.
    - If the scores are equal, then its assigned to be neutral.
    - If the word was absent in our lexicon, then we extracted the synonyms of the word and looked up the scores for any of the synonyms.

Aspect Category Detection:

* Machine Learning Based Approach
  + Pre-processing: Reviews with missing label data were tagged as ‘missing’
  + Tool Used: Python data analytics libraries Pandas, Sci-kit Learn.
  + Training: Tf-Idf Vectorizer of the review text with min\_df=0.5 and max\_df=0.8, along with category label was used for the training phase.
  + The features were passed to rbf svm classifier, linear svm classifier, liblinear svm classifier, random forest classifier and decision tree classifier.
  + Testing: The test data was classified using each of these models and results were compared.
* Lexicon Based Approach
* The aspect defining terms obtained from the hindi dependency parser from dependency based approach of aspect term extraction was used to create a lexicon of aspect categories with each category having their corresponding aspect defining terms.
* Hindi wordnet was used to find the synonyms and antonyms of the defining terms and added to the lexicon.
* While testing, we extract the defining terms in the review and perform a lookup in the lexicon and if found, assign the category to that review, else its assigned ‘misc’ category.

Aspect Category Sentiment Analysis

* Machine Learning Approach
* Training: The same vectorizer of ACD, along with sentiment labels are passed to rbf svm classifier, linear svm classifier and liblinear svm classifier
* Testing: The test data vectorizer is passed to the above mentioned classifier predictors and the results are compared.
* Lexicon Based Approach
* The sentiment score of each category is calculated from the sentiment score of each defining terms.
* The score is added up to check if it is >0,<0 or =0 to determine positive, negative or neutral sentiment.

**Evaluation**

We evaluate the performance of our system by using evaluation methods mentioned by SemEval 2014 task organizers. As a baseline system,  we will be comparing our results with research conducted by Akhtar et al.,2016. We will be using the same annotated dataset used in the baseline system  and divide dataset into 80:20 train and test data respectively. Precision, Recall and F-score will determine the performance of aspect term extraction(ATE). Aspect Term Sentiment(ATS) will be evaluated by Accuracy.  Aspect Category Detection (ACD) and Category Detection(CDS) will be measured by F-score metric from Scikit Learn.

**Aspect Term Extraction**:

Input: Predicted Aspect Term Dataset (Format:  {review\_id: [aspect term]} ),

          Correct Aspect Term Dataset (Format: domainID#review#from#to#polarity#aspect term)

Output: Precision, Recall and F-score

**Aspect Term Sentiment**:

Input: Aspect Term with predicted sentiment,

           Aspect Term with correct sentiment (Format: domainID#review#from#to#polarity#aspect term)

Output: Accuracy

**Aspect Category Detection**:

Input: TF-IDF vectorizer of predicted data, category label

Output:  F-score

**Aspect Category Sentiment**:

Input: TF-IDF vectorizer of predicted data, sentiment label

Output: F-score

**Results**

**Aspect Term Extraction**:

|  |  |  |  |
| --- | --- | --- | --- |
| Method | Precision | Recall | F-score |
| CRF | 0.613 | 0.412 | 0.493 |
| Rule based | 0.099 | 0.414 | 0.161 |

**Aspect Term Sentiment**:

|  |  |
| --- | --- |
| Method | Accuracy |
| SVM Based Approach | 0.688 |

**Aspect Category Detection**:

|  |  |
| --- | --- |
| Method | FScore |
| Rule Based Lexicon Approach | 0.164 |
| RBF SVM | 0.281 |
| Linear SVM | 0.513 |
| Lib Linear SVM | 0.515 |
| Decision Tree Classification | 0.300 |
| Random Forest Classification | 0.468 |

**Aspect Category Sentiment:**

|  |  |
| --- | --- |
| Method | FScore |
| Rule Based Lexicon Approach | 0.153 |
| RBF SVM | 0.433 |
| Linear SVM | 0.631 |
| Lib Linear SVM | 0.613 |

**Discussion**

**References**

Akhtar,M.S., Ekbal, A., Bhattacharya, P. (2016), “Aspect Based Sentiment Analysis: Resource Creation and Evaluation”,  *In Proceedings of Language Resources Evaluation Conference,* Portoroz, Slovenia

Akhtar,M.S., Ekbal, A., Bhattacharya, P. (2016), “Aspect Based Sentiment Analysis: Category Detection and Sentiment Classification for Hindi”,  *In Proceedings of Language Resources Evaluation Conference,* Portoroz, Slovenia

Balamurali, A.R., Joshi, A., Bhattacharyya, P. (2012),” Cross-lingual Sentiment Analysis for Indian Languages using Linked Wordnets”, In *COLING 2012, 24th International Conference on Computational Linguistics*. 73–82

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., Manandhar, S(2014) “Semeval-2014 task 4: Aspect based sentiment analysis”, In *Proceedings of the 8th International Workshop on Semantic Evaluation* , Dublin, Ireland

Toh, Z., Wang, W. (2014), “Aspect term extraction and term polarity classification system”, In *Proceedings of the 8th International Workshop on Semantic Evaluation*, 235–240

“Scikit Learn”. [Online].

Available: <http://scikit-learn.org/stable/modules/svm.html>

“HindiSentiWN”. [Online].

Available:<http://www.cfilt.iitb.ac.in/resources/senti/HSWN_downloaderInfo.php>

“Hindi Wordnet”. [Online].

Available: <https://bitbucket.org/sivareddyg/python-hindi-wordnet>

S. Reddy, “ Hindi Dependency Parser”. [Online].

Available: <https://bitbucket.org/sivareddyg/hindi-dependency-parser>

“Hindi Stopwords”. [Online].

Available: <http://www.ranks.nl/stopwords/hindi>