JCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-11, Nov 2018 E-ISSN: 2347-2693

Aspect retrieval in Hindi language feedback using Rule based method

Deepali Mishra Tiwari

Information Science Engineering, MVJ College of Engineering, Bangalore

Corresponding Author: deepalimishra@mvjce.edu.in, Tel.: +91-8884020125

Available online at: www.ijcseonline.org

Accepted: 18/Nov/2018, Published: 30/Nov/2018

Abstract— The era of Web has resulted in generation of vast amount of user-generated content and analysis of all data by the viewer is time consuming process and viewers are interested towards the specific features of an entity, so the aspect based sentiment analysis became more important. Aspect term Extraction is the prime aim for the aspect based sentiment analysis. Aspect based sentiment Analysis have explored the rule based approach to extract the aspect term . By Experimented approach we have achieved average results for aspect term extraction.

Keywords**—**lexicon,aspect,lemma *etc.*

I. INTRODUCTION

User-generated content is an important source of information to mine the sentiment or opinion of people on different products and services. The developing technology with ease of reachability and better connectivity has led to wide spread use of blogs, forums, e-news, reviews channels and the social networking platforms such as Facebook, Twitter. These social networking platforms have exponentially increased the amount of information generated on daily basis. Thus mining the data and identifying user sentiments, wishes, likes and dislikes is one of an important task that has attracted the focus of research community from last decade. The World Wide Web plays a crucial role in gathering public opinion; these opinions play an important role in making business related decisions. To obtain the factual and subjective information on companies and products, analysts are turning towards web to gather information. Extracting public opinion from this information is a major task.

Due to increase in the technology the user generated data size increasing rapidly and Users are not interested to read all the reviews because it consumes more time, so user want a summarization of the review in the form of each feature involved in the particular domain review. The reason behind this summarization is that each people have different feature choice for the entity, so they only want to know the positive and negative opinion about feature. To get the opinion about aspect of an entity, Aspect level sentiment analysis is occurred in the picture.

The primary goal of the paper is to build a model to extract the Aspect term for Hindi reviews. In this project, we indicate a method for aspect term extraction because for the aspect level sentiment analysis most important thing is aspect term extraction after that extracting opinion oriented polarity for each aspect term in the text. For Aspect term extraction using rule based approach.

II. RELATED WORK

Research is been carried out in the field of Hindi sentimental analysis for over a few years. Sentiment Analysis has been carried out at three levels: document level, sentence level and aspect level. The Aspect level Sentiment analysis for Hindi Language has just now started. Many research work has been done for the Aspect level Sentiment Analysis for English language, so for this project work we considered the papers in which this kind of work done for English language .Brief description of the research done in Aspect level sentiment analysis work have been summarized below.

In [1] paper the work has been attempted is the feature based summarization of product review that means for one feature of a product in how many positive and negative sub-review (sentences) ,we are fetching from whole product review .The work has been done in three steps like identification of feature ,identification of opinion word and their semantic orientation(positive or negative) and the summarization with respect to each feature of the product. For the feature selection process they used NLP approach, in which POS tagging of each review has been done .They consider the noun and noun phrases as the feature of the product. They created a file for collecting the sentences in which the noun and noun phrase are present. They identified the frequent feature using association mining and also done the compactness and redundancy pruning. This approach they considered the adjective as the opinion word that are mostly used to describe the feature of the product. For the identification of the opinion orientation they used the WordNet based approach and according to the opinion word orientation with respect to the feature and classify as positive and negative .They compare the result of feature extraction is compared with the existing well known and publicly available term extraction and indexing system, FASTR and claimed that for feature extraction their system is giving good result and the average sentence classification accuracy is around 84%.

The work has been explored in paper [2] is an unsupervised information extraction system OPINE which mines reviews in order to build a model of important product features, their evaluation by reviewers, and their relative quality across products. This system is built on the top of the Know-it-all , a Web-based, domain-independent information extraction system. The system finds the syntactic relation between the parsed review by dependency parser ,syntactic relation means “mobile has “ (“mobile has good sound quality”) , mobile and sound quality relation is expressed by has word . This proposed OPINE system is finding the parts ,properties ,feature of parts, related concept and related concept feature in the each product for example if we talking about the scanner product scanner cover, scanner size ,battery life ,scanner image and scanner image size. The extracted part and properties are selected as the frequent features by computing the PMI scores between the phrase and *meronymy discriminators* associated with the product class (*e.g.*, “of scanner”, “scanner has”, “scanner comes with”, etc. for the Scanner class). OPINE distinguishes parts from properties using WordNet’s IS-A hierarchy (which enumerates different kinds of properties) and morphological cues (*e.g.*, “-iness”, “-ity” suffixes). For finding the opinion word they used rule based approach and made 10 rules which are based on the syntactic dependency parser output. The opinion words searching is based on the frequent feature i.e. if an explicit feature is present in the sentence than only the rules will be applied. They find the semantic orientation of each opinion word based on the feature, neighborhood feature, support feature. They identify the opinion phrases by the syntactic rules. They experimented the system with 7 different of product review , in which 5 product review data is same as paper[1] and rest two are hotel review(www.tripadvisor.com) and scanner review(www.amazon.com) are collected and the annotation of these reviews are done manually .They claimed 22% higher precision on the feature extraction task compared to previous task[1] with the 5 domain data and for 2 new domain hotel and scanner for opinion word extraction and identifying the polarity of opinion word they got higher precision as compared to the existing approach.

Paper [3] proposed a new method called opinion word expansion and double propagation. In paper , they discussed about the dependency between the words like direct and indirect dependency. They used the POS tag and parser dependency relation tag for making the rules for extracting the aspect and opinion words. In double propagation method, they start search of the aspect term with one opinion word and continue the search of aspect term and opinion word. The propagation then stops as no more features or opinion words can be extracted. For polarity assignment of the opinion word , paper describe some rule like heterogeneous , homogeneous and intra-review rules. They also mention some aspect pruning method. They evaluated their result with different model and claimed average precision and recall for extracting the aspect are 0.88 and 0.83 respectively.

In paper [4] , They proposed a novel idea to find opinion words or phrases for each feature from customer reviews in an efficient way. The focus of this paper is to get the patterns of opinion words/phrases about the feature of product from the review text through adjective, adverb, verb, and noun. The extracted features and opinions are useful for generating a meaningful summary that can provide significant informative resource to help the user as well as merchants to track the most suitable choice of product. They carried out the experiments using customer reviews of 5 electronic products: two digital cameras, one DVD player, one MP3 player, and one cellular phone.They had claimed 85.70% and 73.30% recall and precision for the used approach.

Paper [5] discussed about the explicit and implicit aspect extraction using the rule based approach. In this ,they described the method for explicit and implicit aspect extraction. for the implicit aspect extraction algorithm and lexicon, we use the corpus developed by Cruz-Garcia et al. (Cruz-Garcia et al., 2014), who manually labeled each IAC (implicit aspect clue) and their corresponding aspects in a well-known corpus for opinion mining .The IAC synonyms and antonyms are extracted through the WordNet based approach and polarity orientation of IAC extracted by SenticNet. They proposed the rule for the aspect extraction which is based on the presence of the subject noun and also given some additional rule for the two aspect present in one review. They used the dataset of paper [1] and semeval 2014 data (laptop and restaurant dataset) for experimenting their approach and got higher precision across all domain with respect to all previous approaches for the aspect extraction in the paper [1] dataset and the semeval 2014 dataset the precision are 82.15% and 85.21% for the laptop and restaurant domains.

In paper [6] they have used the priority based approach to aspect level sentiment analysis. They have segregated all the review based on selected aspect for this work and done sentiment classification for reviews of each aspect term and for this work they used naïve based and KNN based approach and based on the sentiment classification of review and count of positive and negative review they decided they orientation of the specific feature.

Paper[7] presents experimental work on a new kind of domain specific feature-based heuristic for aspect-level sentiment analysis of movie reviews. they have devised an aspect oriented scheme that analyses the textual reviews of a movie and assign it a sentiment label on each aspect. The scores on each aspect from multiple reviews are then aggregated and a net sentiment profile of the movie is generated on all parameters. They have used a SentiWordNet based scheme with two different linguistic feature selections comprising of adjectives, adverbs and verbs and n-gram feature extraction. Theyhave also used SentiWordNet scheme to compute the document-level sentiment for each movie reviewed and compared the results with results obtained using Alchemy API. The sentiment profile of a movie is also compared with the document-level sentiment result. They have collected 10 review for 100 movies . The results obtained show that used scheme produces a more accurate and focused sentiment profile than the simple document-level sentiment analysis.

The goal of this paper [8] is to develop a system which can rate the features of restaurant/product along a scaled range from zero to five.They have aimed to generate significant attributes of a restaurant in the 4 categories: Food, Price, Ambience and Service and according to the aggregated review (positive and negative review) about these features they had given rating to selected aspect. They used wordnet based approach to find the feature and for the opinion word extraction used dependency relation based approach .The aspect level sentiment classification done by using sentiwordnet .Each aspects containing positive and negative review are considered for the rating of each aspect

Paper [9] work has been contributed to make a benchmark dataset for the Aspect level sentiment analysis for Hindi language. They collected 12 domains data from different websites and manually annotated the review ,in which they annotated the aspect term ,aspect term category ,aspect term polarity , each review sentence splitting ,classified the sentences into category like positive, negative ,neutral and conflict. They BIO method to describe the aspect term and also define the position of the aspect term. To check the goodness of annotations by different annotators we calculate inter-rater agreement. Cohen's Kappa coefficient is a statistical measure to analyses the inter-rater agreement and an average agreement of 95.18% was obtained. They used the conditional random field model using different features like Word & local context, POS information, chunk information and the suffix and prefix for the aspect term extraction and the SVM model with the features for the sentiment analysis.

Paper [10] work has been explored towards the creation of Hindi dependency parser and the method and models used to build it. In this paper they have given the brief detail about the parser output like dependency tag and dependency relation generated by parser between words. This paper is used to under the Hindi dependency parser and this parser is used for the future work in the Hindi language towards sentiment analysis work.

Paper [11] work has been contributed towards aspect based sentiment classification for Hindi dataset [9] .In this paper they have built a LSTM RNN for aspect based sentiment analysis. Due to the scarcity of Hindi data set they have used English and French data set are translated by using Standard Machine translations Tools.

As per literature survey different approaches like machine learning, pattern based and rule based approaches are used for the aspect extraction task for English language. Recently this year, we found that the aspect level sentiment analysis work has been explored for Hindi language. For this task they have created a bench mark data which is available online and used the machine learning approach for aspect extraction task .Till now as per our survey we did not find any other approach for aspect extraction for Hindi language. So we have explored the aspect extraction approaches used for English language and we came to know that rule based approach is giving good result in the case of English language and we decided to use rule based approach for aspect extraction work for Hindi language.

After literature survey, the many approaches like pattern and rule based are experiment for aspect extraction task in English language .The aspect extraction task is more attempted in English language due to richness of annotated data and the resources .In the explored approaches as per survey rule based approach giving good result .In Hindi language aspect level sentiment classification work recently this year is experimented by the machine learning approach. Only machine learning approach is till now explored for the aspect extraction work, so we need to explore more approaches for the aspect extraction in Hindi language.

III. METHODOLOGY

The main focus of the paper is extraction of the aspect term using rule based approach. The proposed model has certain steps of process like preprocessing, creation of rule set, adaption of rule set to extract aspect for review data and result and analysis. The flow of process is shown in figure.1.

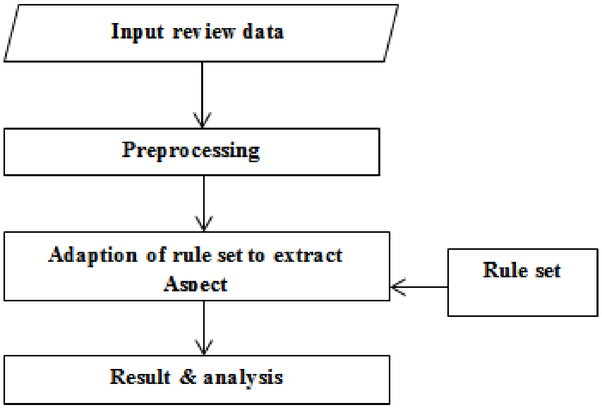


Figure. 1 flow diagram for proposed model

3.1 Preprocessing

The input data is a collection of Hindi reviews and English reviews from various domains. First we have translated English reviews into Hindi language and the preprocessing of the review both data Hindi dependency parser is used. By preprocessing we can find the word id, lemma, POS tag, parent id and dependency label of words in a sentence. In next section gives more detail about Hindi dependency parser.

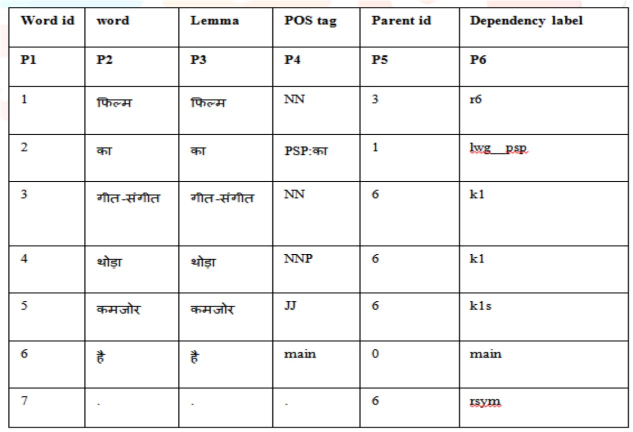
3.2 Hindi Dependency Parser

It is a freely available API on web used to know the relationship between words according to Hindi grammar sentence structuring rules. It tokenized sentence into words and gives the information about each word's word id, word, lemma, POS tag , parent id and dependency label.

3.3 Hindi Dependency Parser

It is a freely available API on web used to know the relationship between words according to Hindi grammar sentence structuring rules. It tokenized sentence into words and gives the information about each word's word id, word, lemma, POS tag , parent id and dependency label . Table 1 shows the sample output of dependency parser.

Table 1. sample output of dependency parser



Word id	word	Lemma	POS tag	Parent id	Dependency label
P1	P2	P3	P4	P5	P6
1	फिल्म	फिल्म	NN	3	r6
2	का	का	PSP:का	1	lwg__psp
3	गीत-संगीत	गीत-संगीत	NN	6	k1
4	थोड़ा	थोड़ा	NNP	6	k1
5	कमजोर	कमजोर	JJ	6	k1s
6	है	है	main	0	main
7	.	.	.	6	rsym

3.3 Rule Set

The rule set is a collection of rules which is used for extracting the aspect terms in the review data. For creation of set the rule based approach is used.

3.3.1 Rule based approach

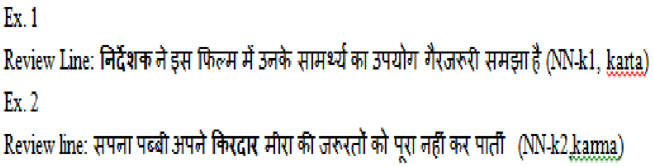
The rule based approach based on the precondition like IF than based. For example if student got 80% marks than grade is A. We

had explained this approach in a simple way but in case of the aspect extraction in the review data is little bit different. In this case, we are using the dependency relationship between words and make rule to extract the aspect. For extraction of words relationship, we used the Hindi dependency parser for rule set creation. We create rules based on the dependency between words .By using parser 'output, we build the rules. Rules are basically in the form of IF-than based structure.

3.3.2 Rule set creation detail

For making rule for aspect extraction, we have used the translated review. As we know, In Hindi language , the sentence structuring is little bit complicated in nature like one sentences can be written in different way. For extracting aspect in the sentence, we have to understand, how aspect term can be expressed in the sentence. In Hindi language , as per provided database, we found that like other language aspect term most of the noun, but in Hindi language according to the Hindi grammar like **Karak Rachna,** it can be karta , karma , karan etc with the noun part of speech tag(NN POS tag). According to karak of Hindi language, we have developed rules. Each karak have relation between the other karak and Part of speech (POS) and exploring that relation, rules are generated. Each karak have its specific role in the

sentence. We used this knowledge because each language have different way of the sentence making and the position of each POS tag in the sentence.
In Hindi language also the nouns are modified by adjective, noun and verb based on these kind of concepts the rules are generated .In Hindi language noun can be with different dependency label. Here we are showing some example to give idea about aspect term as a noun with different form of noun.



Ex. 1

Review Line: निर्देशक ने इस फिल्म में उनके सामर्थ्य का उपयोग गैरजरूरी समझा है (NN-k1, karta)

Ex. 2

Review line: सपना पब्बी अपने किरदार मीरा की जरूरतों को पूरा नहीं कर पातीं (NN-k2,karma)

3.3.3 Rule set and its details

As we know that the aspect term are mostly noun, so the generated rules are based on the relation of the noun with the other Part-of–speech (POS) with the help of dependency tags. We had made rules based on the POS tag and the dependency label like NN- k1, NN-k2, NN-k5 etc. In the rule based approach one type of aspect term can be found by different rules based on the sentence structure. Rules are created on the basis of frequent coming relation between two words with their POS tag and dependency tag.

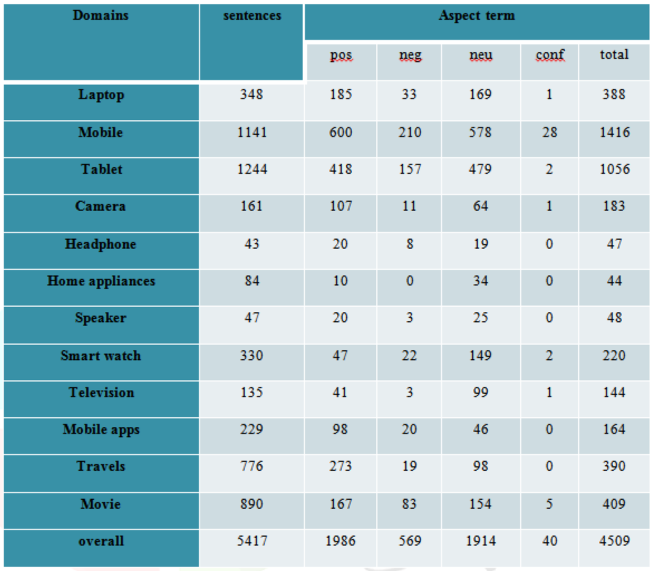
After giving the details of the used field details, proceed towards the details of rules. The rules are made for each category of POS tag and dependency label .Each rule extract one type of aspect term. We have generated more rules based on the Hindi grammar and frequent coming pattern.

As we know, the Hindi dependency parser is used for this experiment and according to the output of the parser, rules are generated. We found that according to Hindi parser noun aspect term expressed by different dependency tag. Hindi language the dependency of aspect term can be with different Part of speech like verb, adjective and adverb. The generated rules are in the rule set are adapted for Hindi review data .

IV. DATA SET DETAILS

We are adapting the data set from paper [9] ,which is a benchmark data set publicly available annotated dataset for 12 domain , but for this experiment 10 domain data is used and paper [5] semval 2014 data set . The proposed approach is experiment in Hindi review data with respect to domain and the data details is shown in table 2 .

Table 2. Data details



Domains	sentences	Aspect term				
		pos	neg	neu	conf	total
Laptop	348	185	33	169	1	388
Mobile	1141	600	210	578	28	1416
Tablet	1244	418	157	479	2	1056
Camera	161	107	11	64	1	183
Headphone	43	20	8	19	0	47
Home appliances	84	10	0	34	0	44
Speaker	47	20	3	25	0	48
Smart watch	330	47	22	149	2	220
Television	135	41	3	99	1	144
Mobile apps	229	98	20	46	0	164
Travels	776	273	19	98	0	390
Movie	890	167	83	154	5	409
overall	5417	1986	569	1914	40	4509

V. RESULT & ANALYSIS

This section, we have discussed about the result of the proposed approach for extraction of aspect term and the error analysis. The result is given by the standard performance metrics.

5.1 Performance metrics

The performance of the proposed approach for aspect extraction measured using the metrics precision and recall. The result is evaluated across the previous build model for aspect term extraction.

Precision: Precision can be thought of as a measure of *exactness* i.e., what percentage of tuples labeled as positive are actually positive.

Precision = TP / (TP+FP)

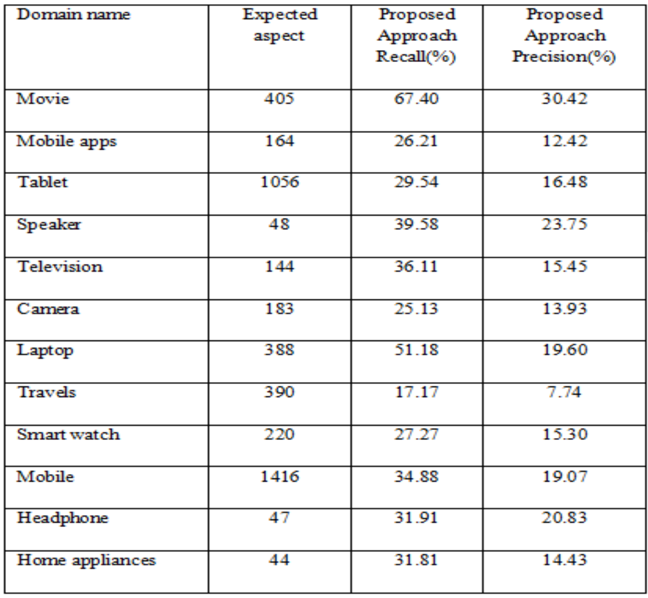
Recall: recall is a measure of *completeness i.e* what percentage of tuples are classified as positive in total positive tuples.

Recall = TP/ (TP+FN)

5.2 Results

The build rules are made based on Translated reviews and applied to Hindi review data. As we know that the data, which we have used for the rule based approach is already experimented by machine learning approach and the results for aspect extraction they have claimed are shown in table 3.

Table 3 Result of proposed approach for aspect extraction



Domain name	Expected aspect	Proposed Approach Recall(%)	Proposed Approach Precision(%)
Movie	405	67.40	30.42
Mobile apps	164	26.21	12.42
Tablet	1056	29.54	16.48
Speaker	48	39.58	23.75
Television	144	36.11	15.45
Camera	183	25.13	13.93
Laptop	388	51.18	19.60
Travels	390	17.17	7.74
Smart watch	220	27.27	15.30
Mobile	1416	34.88	19.07
Headphone	47	31.91	20.83
Home appliances	44	31.81	14.43

VI. Conclusion And Future work

We have used a rule based approach to fetch the aspect term. The result of the proposed approach for the aspect word extraction in terms of recall and precision are 67.40% and 30.42% for movie domain and other domain are giving average results. In some domains, we got good recall value as compare to existing approach but precision of used approach is less, due the used Dependency Parser , annotated data and unable to find multiword aspect term.

As we know that for English language so much good resources are freely available with good quality. But for Hindi language less annotated data and resources are available. Even though with available data and resources, we tried to attempt the proposed approach. As the work is totally dependent on the data and dependency parser, so may recall are average and precision is low. The rule based approach is time consuming approach means tuning of the rule is must so that the performance of rule will be good.

We found that only rule based approach is not enough for aspect extraction work, in future we have to explore hybrid approach so that we can extract more aspect because rule build by us are not generic they are domain specific. Our future work should more towards in extraction of multiword aspect term and creation of domain specific lexicon resources for other domain also, so that we will achieve more precision for aspect term extraction and aspect level classification accuracy improvement.

References

[1] Mining and summarizing customer reviews; Minqing Hu and Bing Liu; In Proceedings of the ACMSIGKDD International Conference on Knowledge Discovery & Data Mining; 2004 ; pages 168–177, Aug.

[2] Extracting product features and opinions from reviews ; Ana-Maria Popescu and Oren Etzioni; In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-2005); 2005; pages 3–28.

[3] Opinion Word Expansion and Target Extraction through Double Propagation ; Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen; Computational linguistics; 2011 ; 37(1):9–27.

[4] Extracting Product Features and Opinion Words Using Pattern Knowledge in Customer Reviews;Su Su Htay and Khin Thidar Lynn; The ScientificWorld Journal Volume 2013, Article ID 394758, 5 pages

[5] A Rule-Based Approach to Aspect Extraction from Product Review; Soujanya Poria, Erik Cambria, Lun-Wei Ku, Chen Gui, Alexander Gelbukh ;ACM ;2014.

[6] An Approach to Perform Aspect level Sentiment Analysis on Customer Reviews using SentiscoreAlgorithm and Priority Based Classification ;Aishwarya Mohan, Manisha.R, Vijayaa.B, Naren.J ; International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 4145-4148

[7] Sentiment Analysis of Movie Reviews -A new Feature-based Heuristic for Aspect-level Sentiment Classification;,V.K. Singh, R. Piryani, A. Uddin,P. Waila; Researchgate;2015

[8] Aspect Based Analysis for Rating Prediction of the Restaurant Reviews;Namita Mittal, Basant Agarwal, Shalini Laddha, Manish Sharma;International Journal of Computer System (ISSN: 2394-1065), Volume 02– Issue 03, March, 2015

[9] Aspect based Sentiment Analysis in Hindi: Resource Creation and Evaluation; Md Shad Akhtar, Asif Ekbal and Pushpak Bhattacharyya; LRTC ; 2015

[10] Hindi Dependency Parsing and Treebank Validation ; Bharat Ram Ambati ;LRTC;2005

[11] Solving Data Sparsity for Aspect based Sentiment Analysis using Cross-linguality and Multi-linguality ;Md Shad Akhtar , Palaash Sawant, Sukanta Sen ,Asif Ekbal and Pushpak Bhattacharyya;Proceedings of NAACL-HLT 2018, pages 572–582 New Orleans, Louisiana, June 1 - 6, 2018. c 2018 Association for Computational Linguistics

[12] https://bitbucket.org/sivareddyg/hindi-dependency-parser