IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 1, January 2021

DOI 10.17148/IJARCCE.2021.10106

Lexicon Based Sentiment Analysis for Hindi Reviews

Kameshwar Singh

Linguistics and Language Technology Department

Mahatma Gandhi International Hindi University Wardha (Maharashtra) 442001

Abstract: Online shopping increasing rapidly Now a day, a huge amount of hindi reviews is present on the E- commerce website. In this paper we have proposed a strategy for classifying given Hindi texts in to different classes and then extract sentiments in terms of positive, negative and neutral for identified classes. Negation is also handled in the proposed system. There are mainly two approaches used for sentiment analysis- lexicon based and machine learning based approach. We emphasis on lexicon based approach which depends on an external dictionary. The system classifies the reviews as positive, negative and neutral and calculate the score for Hindi language. The Methodology used in proposed system is Hybrid approach and Modal is Statical based.

Keywords: Lexicon based approach, Hybrid approach, Statical based model.

I. INTRUDUCTION

A growing business in the E-markets nowadays is buying goods online. A major feature in this case is that people also show their reaction and attitude towards the product through their product reviews. This may cause other customers to think about whether or not to buy the product on the e-market based on previous customer experiences and reviews. These product reviews reflect the perception of individuals about the product [1]. Customers post reviews online By analyzing his Experience, product quality and deficiencies can be understood. Sentiment analysis is handled at many levels of granularity i.e. at the document level, sentence level, phrase level, word level. etc.

1. **Document level**- sentiment analysis is determined at whole document e.g. the document is given bellow-

इस किताब ने मेरे दिल को झकझोर के रख दिया बहुत ही अच्छा उपन्यास है इसको पढ़कर मज़ा आया लेकिन कुछ आंसू भी निकल आये इसको पढ़ने के बाद जैसे लग रहा था कि मैं भी मुखर्जी नगर में इन्हीं सब के साथ हु मैं लेखक को धन्यवाद देना चाहता हूँ मैं तो ये कहूंगा कि असली डार्क हॉर्स लेखक है जिसने मुझे इससे परिचित कराया कि भाग्य ही सब कुछ नहीं होता है असली चीज मेहनत होता है सही रास्ता और मेहनत से कुछ भी पाया जा सकता है इस दुनिया में ।।।।। बहुत बहुत धन्यवाद लेखक का और अमेज़न का जिसके माध्यम से मुझे डार्क हॉर्स पढ़ने का मौका मिला

At the document level classification this document classify Positive Openion.

Sentence Level- for classifying sentence level sentiment analysis there are tow types of sentence given below-

1. Simple sentence

शानदार उपन्यास...लेखक ने जैसा देखा वैसा उतार दिया (Positive sentence)

Phrase level- Ditermine the polarity of document/sentence at the level of Sarcasm and phrase identifying e.g. the sentence is given below-

1. इस उपन्यास की कहानि **<u>हृदय को छू गई।</u>** (Positive Phrase)
2. कुमार ने इस कहानि के रोचकता को **खालिया**। (Negative Sarcasm)

Word level- Word level sentiment analysis is determines the polarity of sentences/documents for each Morphological and Phonemic variation it contains. e.g. the sentences given below are classified at Morph level as-

1. इस पुस्तक के लेखक बहुत बड़े गुरु है। (Positive sentice)

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 1, January 2021

DOI 10.17148/IJARCCE.2021.10106

2. इस पुस्तक के लेखक बहुत बड़े गुरुउ.....हैं। (Negative sentence)

In above sentence the word is same at the Morph level but in second sentence the at word length level change the meaning of sentence.There are many NLP technique which detects the sentiments of E-commerce website like Stop word removing, Parts of Speech Tagging, Name Entity Recognition (NER), which is trailed by bags of words etc. I am using lexicon approach which is used to classify the text into three classes: Positive, Negative and Neutral with the help of dictionaries. The challenges that arise during extraction of the features and then doing classification of that text are given below but some of the challenges are removed by cleaning the text data set.

- Handling the Morphological variations.
- Identifying context performing word sense.
- Polarity Shifting detection [4].
- Handling the big data which consist of the opinions given by the people. Informal languages, slang word/abbreviation or emoticons usage.
- Spelling mistakes/ typo mistakes.
- Disambiguation word Identification
- Handling Multiple speling.

Section I gives introduction of this topic. Section II Related word Section III Proposed system gives brief description of sentiment Analysis in English as well as the work done in Indian Regional language like Hindi etc. includes proposed architecture. Section IV describes methodology used in this paper and also describes the experimentation and evaluation in detail. At last Section V gives conclusion and future scope of this paper.

II. RELATED WORK

The Most of Research work done for sentiment Analysis on Hindi text. there is been a considerable progress in past few year. Some of them were:

Aditya Joshi, Balamurli [5] proposed cross lingual sentiment analysis for Indian Languages.

Joshi, B. A. R, and P. Bhattacharyya [7] proposed a fall back strategy for sentiment analysis of Hindi language. In this model three 5approaches were used- In-language translation, machine translation resource based sentiment analysis for Sentiment analysis of Hindi language .In the first approach a sentimental annoted corpora has been developed in the Hindi movie review domain and it involves a training classifier to classify a new document in Hindi language. In the second approach, a classifier trained on standard English movie reviews has been used to translate the given document into English. In the third approach a lexical resource called HindiSentiWordNet is developed and majority score based strategy is implemented to classify a given document.

N. Mittal, B. Agarwal, G. Chouhan [10] studied on the Hindi language content. In this paper, it is examined that by appropriate handling of negation and discourse relation it may improve results in comparison to other existing methods.

Richa Sharma,Shweta Nigam, Rekha Jain[1]" Polarity Detection Of Movies Reviews In Hindi Language, in this research paper the Polarity classifying Positive and Negative.

In [6] is a rule based approach for understanding sentiment from Malayalam reviews. In this paper, a sentence level sentiment extraction is used. The sentence level sentiment extraction is helpful in movie websites that user comments. For analyzing sentiments negation rules are used. It will reduce the errors. Here first collects the corpus from movie websites or blogs and newspapers. Using Sandhirules the sentence re divided in to various tokens. Then each word is compared with the pre-annotated list, and then the words are classified into positive, negative or neutral polarity. After this apply negation rule to identify the overall polarity. Accuracy of the paper is 85%.

III. PROPOSED SYSTEM

1. Hindi Text corpus build by train data.
2. Opinion words extraction and Seed list preparation
3. Polarity detection of reviews
4. Identify the sentiments for text corpus.
5.

Description of the system in short

Algorithm

1. Read text file from e-commerce website Product reviews.

2 Scrape Data from the Webpage like article text and heading.

3. Perform Pre-processing techniques such as Segmentation, Tokenization and Stop-words Removal for Text Summary. Every word is reduced to its root word.

4. Compare all the words with positive, negative and Neutral words list

If match found.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 10, Issue 1, January 2021
DOI 10.17148/IJARCCE.2021.10106

Increment the corresponding counters
Else
Continue
5. For remaining words in the sentence
Check for Negation property
If found
Allocate reverse polarity
Else
Continue
6. If other words found repeat step 4
Else
Go to Step 7
7. Compare Positive and Negative pointer
8. Display the polarity.
9. Then the probability of each word in that class is calculated.

Data Crawling

Manually collect product reviews from various E-commerce website. We have crawled more the 1 e-commence websites etc. Following this process we have collected a total of 2500 reviews.
List of few sources..
http://www.amazon.com
http://www.flipkart.com
http://www.patanjaliayurved.in

WORKING PROCEDURE

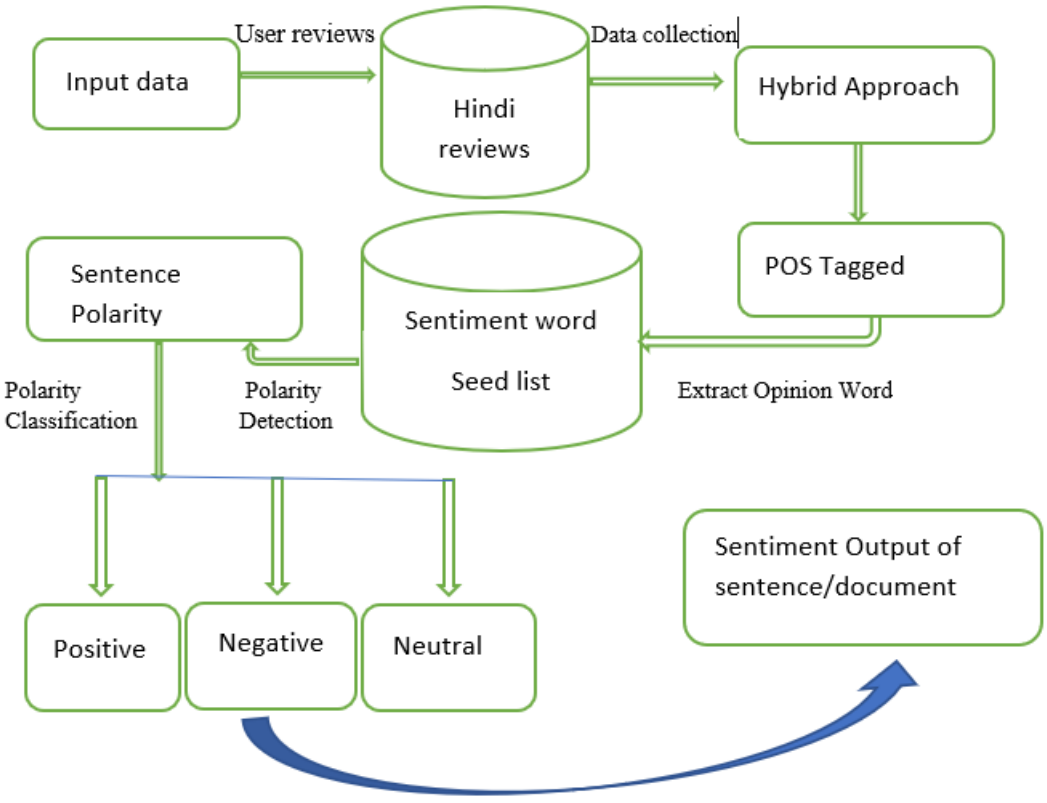

Figure 1. Sentiment Analysis Hindi Product Reviews System

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 1, January 2021

DOI 10.17148/IJARCCE.2021.10106

PROPOSED MODULE DESCRIPTION

1. **Validation**: We will validate whether the text is Hindi or Non-Hindi. For validation purpose we will be using sentiment word list. Along with that special characters like? / @ $ % will be removed.
2. **Tokenization**: Sentence is divided into number of tokens with the help of dot net string tokenizer.
3. **Stop Word Removal**: Stop words are the words which do not convey any meaning they are just used to complete the sentence. For example, की,के,में,भी Here all the stop words will be removed.
4. **Morphological analysis:** Here we will find root words.

For example, अच्छाई ~~अच्छा~~ →

5.Negation Handling

Since Hindi language is known for its nature of being unstructured. Negation handling for this language can be quite difficult. This stage involves treating negation in text. The negation operator (ना,नहीं,मत etc.) present in the text mostly inverts the meaning of the text which affects the sentiment score or polarity in a critical way. To handle situation first we consider a couple of words of size (4 to 7). Mostly the system looks for the occurrences of "नही". On encountering a negation operator assign reverse polarity to all the words that appears in the given example.

1. यह किताब अच्छी नहीं है।

Week Negative Polarity with single negation.

2. नहीं जी यह किताब अच्छा नहीं है।

Strong Negative Polarity with Double Negation.

3. लेखक की पकड़ कही भी कमजोर नहीं है।

Strong Positive Polarity with single negation.

4. ना मौलिकता ना ही काम अच्छा है।

Week Negative Polarity with Double negation.

POS Tagging: Part of speech will be assigned to each word. Whether it is Noun, Pronoun, Verb or Adjective, Adverb, etc.

1. ना_/NEG ही /RP किताब_/NN का/PREP लेखन/NN_नया_/JJ _है_/VFM और/CC ना/NEG ही/RP किताब_/NN में/PREP ज्यादा/JJ मजेदार/NN स्टोरी_/VFMहै।

Opinion words extraction and Seed list preparation

Seed list is prepared first in which most frequently used Hindi words along with their polarity are stored. All the opinion words which were extracted after the POS tagging are first matched with the stored words in the seed list if it is matched with the words stored in the seed list then there is no need to determine the synonyms of the word. But if the word is not found in the seed list then the synonyms of that word are determined with the help of Hindi dictionary that is also built by us. Each synonym is matched with the words in the seed list, if any synonym is matched the opinion word along with its synonyms is stored in the seed list with same polarity. It grows every time whenever synonyms words found in Hindi dictionary are matched with seed list.

Polarity detection of reviews

In the last phase, the polarity of the collected reviews is determined with the help of seed list and Hindi dictionary. The polarity of the reviews is determined on the basis of majority of opinion words, if positive words are more in the review than the polarity of the review is positive otherwise it is negative. If positive and negative words are equal in a review the polarity is neutral. As negation is also handled in this approach, so if the opinion word is followed by not then the polarity of review is reversed. e.g. the sentence.

Id	Doman	Token	Sentence	Sentiment word		
				Net	Pos	Neu
	Books	690	234	123	234	123
2	Ayurvedic Medicine	934	345	323	657	234
3	Mobile	234	267	134	234	65
4	Leptop	765	234	67	423	45
5	Home Appliance	885	348	345	311	22

Table1.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 1, January 2021

DOI 10.17148/IJARCCE.2021.10106

Dataset Statistics

After pre-processing, our dataset contains 1500 review sentences across 3 domains. There are a total of 678 positive, 456 negative, 134 neutral and reviews (sentence-level). Overall it contains 2456 and 4,509 tokens and sentiment word, respectively. Polarity classification of these sentiment word count to 1,986 positive, 569 negative, 1,914 neutral and . Overall and domain-wise details of this dataset are reported in Table.

EXPERIMENTS & RESULTS

Experiment is conducted on Product's reviews. Product's reviews were collected from several E-commerce websites contain Hindi reviews. Reviews were applied as input to the system which classifies these reviews and determine the polarity of these reviews and present the summarized positive
and negative and Neutral results which prove to be helpful for the users. Input reviews were also classified by us to determine how well the system classified the reviews as compared to human judgement. Three evaluation measures are used on the basis of which system performance is computed, these are:

- Precision
- Recall
- Accuracy

The common way for computing these measures is based on the confusion matrix shown in
Table 2.

Instances Predicted	Instances Predicted	Instances Predicted
Actual positive instances	# of True positive instances (TP)	# of false negative instances (FN)
Actual negative instances	# of false positive instances (FP)	# of True Negative instances (TN)

Table2

Precision: It is the ratio of true positive predicted instances against all positive predicted instances.
Precision= tp/(tp+fp) ….eq 1.
Recall: It is the ratio of true positive predicted instances against all actual positive instances.
Recall =tp/(tp+fn) ….eq 2.
Accuracy: It is the ratio of true predicted instances against all predicted instances.
Accuracy=(tp+tn)/(tp+fp+tn+fn) …eq3.

Method	Precision	Recall	Accuracy
With Negation	63.13%	44.15%	86.16%
Without Negation	67.23%	47.06%	88.14%

CONCLUTION

In this paper we propose a Sentiment Analysis of Product's Reviews in Hindi Language. We have collect Product's review various online sources, performed pre-processing to clean the data, and annotated the dataset with morphological based and polarity classes. The dataset comprises of Hindi product reviews collected from various online sources across 3 domains. Based on this dataset we build seed list for extraction and sentiment classification. Evaluation results on 3-fold cross-validation show the overall precision, recal values of 63.96%, 44.72% , respectively for aspect term extraction and an accuracy of 64.05% for sentiment classification. We also make the dataset available to the community for the advancement of further research involving Indian languages. In future, we would like to investigate Hybrid Language on E-Commerce Website. We would also like to explore deep learning methods for dynamic aspect based sentiment analysis.

REFRENCES

1. Richa Sharma,Shweta Nigam, Rekha Jain (2014),” Polarity Detection Of Movies Reviews In Hindi Language” International Journal on Computational Sciences & Applications (IJCSA) Vol.4, No.4, August 2014.
2. Sumedha Ubale, Ankita Sarang, Kajol Wadye, Prof. Nita Patil,” Hindi Sentiment Analysis” nternational Journal on Future Revolution in Computer Science & Communication Engineering ISSN: 2454-4248 Volume: 4 Issue: 4 536 – 540.
3. Prof. Omprakash Yadav, Rahul Patel, Yash Shah, Saneesha Talim,” Sentiment Analysis on Hindi News Articles” International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056.Volume: 07 Issue: 05 | May 2020 www.irjet.net p-ISSN: 2395-0072.

DOI 10.17148/IJARCCE.2021.10106

4. R. Xia, F. Xu, C. Zong, Q. Li, Y. Qi and T. Li, \Dual sentiment analysis: Considering two sides of one review," IEEE transactions on knowledge and data engineering,vol. 27, pp. 2120{2133, 2015.
5. Balamurali A R, Aditya Joshi, Pushpak Bhattacharyya," Cross-Lingual Sentiment Analysis for Indian Languages using Linked WordNets", COLING 2012, pp. 73-8, December 2012.
6. Nair , D.S.,Jayan , J.P., Rajeev , R.R.,Sherly,E. " SentiMa- Sentiment extraction for Malayalam." Advances in computing, Communication and Informatics (ICACCI) 2014 International Conference on IEEE, 2014
7. Joshi, B. A. R, and P. Bhattacharyya.(2010)," A fall-back strategy for sentiment analysis in Hindi: a case study" In International Conference On Natural Language Processing (ICON).
8. Akshat Bakliwal, Piyush Arora, Vasudeva Varma,(2012)"Hindi Subjective Lexicon : A Lexical Resource For Hindi Polarity Classification".In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC).
9. Akshat Bakliwal, Piyush Arora, Ankit Patil, Vasudeva Varma,(2011),"Towards Enhanced Opinion Classification using NLP Techniques" In Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP), IJCNLP, pages 101–107, 2011.
10. Namita Mittal,Basant Agarwal,Garvit Chouhan,Nitin Bania,Prateek Pareek,(2013), "Sentiment Analysis of Hindi Review based on Negation and Discourse Relation"in proceedings of International Joint Conference on Natural Language Processing, pages 45–50,Nagoya, Japan, 14-18 .
11. Piyush Arora, Akshat Bakliwal and Vasudeva Varma,(2012) "Hindi Subjective Lexicon Generation using WordNet Graph Traversal" In the proceedings of 13th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing), New Delhi, India
12. Subhabrata Mukherjee, Pushpak Bhattacharyya,(2012)"Sentiment Analysis in Twitter with Lightweight Discourse Analysis", In Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012.
13. Namita Mittal, Basant Agarwal, Garvit Chouhan, Prateek Pareek, and Nitin Bania (2013) "Discourse Based Sentiment Analysis for Hindi Reviews" P. Maji et al. (Eds.): PReMI 2013, LNCS 8251, pp. 720–725, 2013.
14. Piyush Arora, Sentiment Analysis For Hindi Language, *MS Thesis IIITH,* 2013.

15. Rinku T S, Merlin Rajan and Varunakshi Bhojane , "Various Approaches Used for Tagging and Chunking in Malayalam " , International Journal of Scientific and Engineering Research,Volume:5 ,may 2014.

BIOGRAPHY

Kameshwar Singh is currently pursuing Ph.D (Language Technology) at Mahatma Gandhi Antarrastriya Hindi Vishwavidyalaya Wardha Maharashtra. His research interests area is Language Technology and Natural Language Processing, Sentiment Analysis.