EasyChair Preprint

№ 5690

Sentiment Analysis on Hindi Content: a Survey

Priyanka Mishra and Shilpa Agnihotri

June 3, 2021

Sentiment Analysis on Hindi Content: A Survey

Prof. Priyanka Mishra
,Prof .Shilpa Agnihotri

Department of Computer Science and Engineering

SIRT

Bhopal, India

ABSTRACT

Sentiment analysis means analyzing the sentiment information in order to draw the conclusion and understand the overall situation. In order to understand data machine needs to understand what are sentiments from the input like product review, movie review, news review, comments from blogs or any social website and give output as positive or negative or neutral review. Various algorithms and classifiers are present for sentiment analysis. This paper gives a survey of various approaches used specifically for Hindi language.

Keywords

Sentiment analysis, Classifier, Sentiments, Natural Language Processing.

INTRODUCTION

Sentiment analysis is the process of determining the emotional tone behind a series of words which is used to gain and understand the attitudes, opinions and emotions expressed from language. It is highly useful in social media monitoring as it allows us to gain an overview of the wider public opinion behind certain topics. The applications of sentiment analysis are broad and powerful. The ability to extract insights from social data is a practice that is being widely adopted by organisations across the world. So we need a sentiment analysis system for Hindi language.

In Natural Language Processing (NLP), sentiment analysis is an automated task where machine learning is used to rapidly determine the sentiment of large amount of text or speech. Applications include tasks like determining how excited someone is about an upcoming movie, marathon participation in a particular city, reviews for any new product in market, converting written restaurant review into 5-star scale across various categories like food quality, service and value of money.

Majority of the existing work has been done for English language. Very little attention has been paid towards other languages. As the Internet is reaching to more people within the world, there is huge increase in web content as people feel comfortable

with their native language. Hindi is the fourth highest speaking language in the world. The increasing user-generated content on the Internet is the motivation behind the sentiment analysis research. So there is a need to pay attention in direction of sentiment analysis for Hindi Language.

In this paper, we are giving a survey of various classifiers and the approaches used for sentiment analysis specifically for Hindi language data.

SENTIMENT ANALYSIS

Sentiment Analysis is a natural language processing task that deals with the extraction of opinion from a piece of text with respect to a topic (Pang et al., 2008). A large number of advertising industries and recommendation systems work on understanding liking and disliking of the people from their reviews. Hindi is most commonly spoken language in the world. So the information content in Hindi is important to be analyzed for the use of industries and government(s).[7]

Sentiment analysis is very difficult for Hindi language due to numerous reasons as follows :[7]

(1) Unavailability of well annotated standard corpora, therefore supervised machine learning algorithms cannot be applied.

(2) Hindi is a resource scarce language; there are no efficient parser and tagger for this language.

(3) Limited resources available for this language like HindiSentiWordNet (HSWN). It consists of limited numbers of adjectives and adverbs. All the words are available in inflected forms. Even all the inflected forms of the word are not present. HSWN is created using the Hindi WordNet and English SentiWordNet (SWN). During the creation of this resource for Hindi language, it is assumed that all synonyms have the same polarity while all antonyms have the reverse polarity of a word. This assumption neglected word sense intensity in terms of polarity, however polarity intensity of their word is important in opinion mining.

(4) Even, Translation dictionaries may not account for all the words because of the language variations. Same words may be used in multiple contexts and context dependent word mapping is a difficult task, error prone and requires manual efforts. Using Translation method for generating subjective lexicon, there is a high possibility of losing the contextual information and sometimes may have translation errors.

LITERATURE SURVEY

In this section we cite the relevant past literature.

Many researchers have worked on various aspects of opinion analysis. (Pang et al., 2002), (Turney, 2002) worked on document level sentiment classification. (Wiebe et al., 1999), (Intelligent and Wilson, 2003), (Yu and Hatzivassiloglou, 2003), (min Kim, 2004), (Hu and Liu, 2004) worked on sentence level sentiment classification. More recently (Wilson, 2005), (Agarwal et al., 2009) worked at phrase level sentiment classification.

Paper[1] A Fall-Back Strategy for Sentiment Analysis in Hindi: a Case Study

Aditya Joshi, Balamurali A R, Pushpak Bhattacharya (2010) developed a sentiment annotated corpora in the Hindi movie review domain. There are 3 approaches:

a) First approach is training a classifier on this annotated Hindi corpus and using it to classify a new Hindi document.

b) In the second approach, they translated the given document into English and use a classifier trained on standard English movie reviews to classify the document & detect the polarity of the translated document using a classifier in English, assuming polarity is not lost in translation.

c)In the third approach, we develop a lexical resource called Hindi-SentiWordNet(H-SWN) and implement a majority score based strategy to classify the given document.

Observations

- Manually annotated corpus for Hindi was created.
- Hindi-SentiWordNet based on the equivalent for English was created.
- Their study suggested that machine learning-based approaches are better than resource-based approaches.
- To predict the sentiment of a document Naive approach is used in which prior priority of terms are present in it.
- SentiWordNet 1.1 an automatically generated WordNet-based lexical resource with polarity scores attached to the senses is used. The sum of the scores adds to 1. WordNet linking is just specified which can be used for mapping between synsets of WordNets of different languages.
- POS tags are found out.
- Corpus is standardized to UTF-8 format. With 250 Hindi movie reviews.
- Term frequency, TF-IDF, Term Presence are used for feature representation to see the effect on Hindi review documents. Google translate is used.
- Accuracy for in-language SA, MT-based SA & Resource based SA are 78.14%, 65.96% and 60.31% respectively.

Three approaches for sentiment analysis of the Hindi Documents are [2]:

1. In-language Sentiment Analysis

The approach relies on the availability of the resources needed to analyze the sentiment content in Hindi. RapidMiner5.0 for classification of document is used. Learner used for classification is LibSVM with SVM type as C-SVC with Vanilla properties.

Experiment Setup	Representation	Accuracy
Unigram + Stemmed	TF	67.83
Unigram + Stemmed	TP	66.23
Unigram + Stemmed	TF-IDF	68.65
Unigram + not stemmed	TF	74.57
Unigram + not stemmed	TP	72.57
Unigram + not stemmed	TF-IDF	78.14
Unigram + Stemmed	TF	61.2
Unigram + Stemmed	TP	57.2
Unigram + Stemmed	TF-IDF	62.4
Unigram + not stemmed	TF	70.02
Unigram + not stemmed	TP	71.72
Unigram + not stemmed	TF-IDF	71.72

Figure 1 : Results of in-language sentiment analysis(in %) for Hindi using different features [2]

Limitation : The proposed approach relies on the availability of the resources needed to analyze the sentiment content in Hindi.

2. Machine Translation(MT)-based Sentiment Analysis

In this approach, a translation module is used to translate documents in Hindi to English using Google translate assuming the sentiment of a document is preserved in translation. Translated document is given to the classifier which gives polarity as the output.

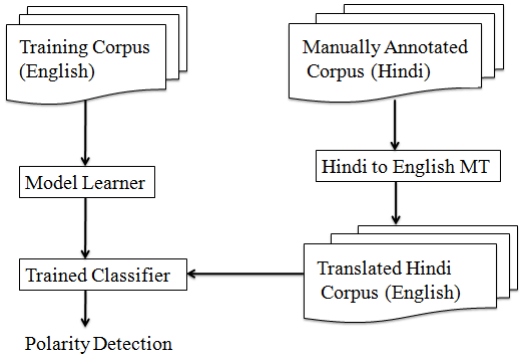

Figure 2 : Procedure for MT-based Sentiment Analysis [2]

Limitation : Online Google translate is used to translate Hindi to English words.

3. Resource based Sentiment Analysis

A classifier is implemented to use scores in H-SWN. Stemmer and stop word list are used.

Algorithm

- For each word in the document,
  - Apply stop word removal and stemming (depending on the variant of the experiment)
  - Look up the sentiment scores for each word in the H-SWN.
  - Assign a polarity to a word based on the maximum of the scores
- Assign to a document the polarity which majority of its words possess.

Sense Consideration	Stemming	Stop Word Removal	Accuracy
Most Common Sense	No	No	56.35
	Yes	Yes	53.96
All Senses	No	No	60.31
	No	Yes	57.53
	Yes	Yes	55.95

Figure 3 : Results for resource-based sentiment analysis[2]

Algorithm for creation of H-SWN

1. For each synset in the SWN, repeat 2 to 3
2. Find the corresponding synset in Hindi WordNet
3. Project the scores of a sysnset in SWN to the corresponding sysnset in Hindi WordNet.

Experiment	Accuracy %
In-language sentiment analysis	78.14
MT-based sentiment analysis	65.96
Resource-based sentiment analysis	60.31

Figure 4 : Comparison of approaches[2]

Limitation : Even though the word is present in the database the correct sentence with meaning is not accurate i.e. the meaning of the original language changes. This can be overcome by word sense disambiguation.

Paper[2] Hindi Subjective Lexicon : A Lexical Resource for Hindi Polarity Classification

Bakliwal et al. (2012) created lexicon using a graph based method. They explored how the synonym and antonym relations can be exploited using sample graph traversal to generate the subjectivity lexicon. Their proposed algorithm achieved approximately 79% accuracy.[3]

Observation

- Developed lexicon of adjectives and adverbs with polarity scores using Hindi WordNet
- Developed an annotated corpora of Hindi Product Reviews.
- Simple scoring method classification is used.
- Shallow parser is used to identify adjectives & adverbs in a sentence.
- Negation handling was performed.
- Adjective scoring-Baseline+stem+NH gave maximum accuracy of 78.41 and Adjective & adverb scoring for baseline+stem+NH gave 79% accuracy.

Limitation : Work can be extended to incorporate Word Sense Disambiguation (WSD) and morphological variants which could result in better accuracy for words which have dual nature. We experimented with adjectives and adverbs, this work can be extended for other parts of speech (verbs and nouns).

Paper[3] Sentiment Analysis in Hindi

Naman Bansal and Umair Z. Ahmed. Advisor : Amitabha Mukherjee used Semi-Supervised approach to train Deep Belief Network on small percentage of labelled data (150) and assign polarity to unlabelled data. 300 sentences dataset stored in xml. Sentence lever sentiment is adopted. Data preprocessing stage, Deep belief architecture divided in 2 stages: pre-training model & fine tuning step. They report 71% accuracy using DBN on English language, 76% using active deep learning and 64% on Hindi language using Deep Belief Network.[4]

Observations : Movie reviews were obtained from IITB for research purpose and content from jagran.com website was also collected and were stored in XML format so that it is easily to parse. The semi-supervised learning method based on DBN architecture is divided into two steps: first is pre-training model is constructed by greedy layer-wise unsuperwised learning using Restricted Boltzman Machine (RBM) in which all the labeled data together with L labeled data are utilized to find the parameter space W with N layers. Second: fine log likelihood using gradient descent method is used. The parameter space W is retained by a negative log likelihood cost function using L labeled data to fine the parameters space only according to labeled data. Experiments carried out for multiple configurations of number of neurons in the hidden layers. The best configuration for the deep belief network founded was Five layer network, One Input, Three Hidden and One Output Layer.

Limitation : The sentences are marked incorrectly as negative or positive even though they are not. Negation handling is missing.

Paper[4] A Hybrid Approach for Twitter Sentiment Analysis

Namita Mittal et al. (2013) introduced an approach for automatically classifying the sentiment of Twitter messages. These messages are classified as either positive or negative. A 3 stage hierarchical model is proposed by the author; first : labeling with emoticans; second : using predefined list of words with strong positive or negative sentiments; third : tokens are weighted based on subjectivity lexicon.[5]

Observations : The author used 60,000 tweets. Rules for assigning weight to the tokens which are not present in the positive as well as negative list of words. Accuracy obtained without discourse by SentiWordNet was 64.694, proposed probability based method 71.116,SentiWordNet then probability based method 69.511, probability based method then SentiWordNet 71.83, Hybrid approach (SWN+Probability) 72.563. With discourse results obtained for the accuracy were; for SentiWordNet 66.052, proposed probability based method 71.625, SentiWordNet then probability based method 70.193, probability based method then SentiWordNet 73.35, Hybrid approach (SWN+Probability) 73.72.The results shows that incorporation of discourse makers improves the sentiment classification accuracy.

Limitation : Rules for handling sarcasm is missing.

Paper[5] Sentiment Analysis of Hindi Review based on Negation and Discourse Relation

Namita Mittal et al. (2013) proposed method for improving HSWN.770 reviews were taken as input. N-gram & POS-tagged N-gram approaches were also used. Fleiss kappa coefficient value 0.8092 &dataset has 104 words.

Observations : An algorithm for negation handling and discourse relation is given. Accuracy of 82.89% for positive reviews, 76.59% for negative and overall 80.21%.for improved HSWN + negation + Discourse.[6]

Limitations : The size of the dataset is small.

Paper[6] A Survey on Sentiment Analysis and Opinion Mining Techniques

Amandeep Kaur and Vishal Gupta (2013) gave the process of Sentiment Analysis is divided into 5 steps : process of Sentiment Analysis for Text(Lexicon generation), Subjectivity detection, sentiment polarity detection using Network Overlap Technique, sentiment structurization, sentiment summarization-visualization-tracking.[7]

Observations : There are four approaches to predict the polarity of a word. In the First strategy; an interactive game is provided which identify the polarity of the words. In the Second strategy, a bi-lingual dictionary is developed for English and Indian Languages. In the third strategy, word net expansion is done using antonym and synonym relations. In the fourth approach, a pre-annotated corpus is used for learning.

Limitation : subjective lexicon can be developed for the unexplored languages which does not have a

word net. The basic resources like parsers, named entity recognizers, morphological analyzers, and part of speech tagger need to be improved to reach the state of accuracy.

Paper[7] Sentiment Classification in Hindi

Sneha Mulatkar (May2014) used WSD algorithm (Sense disambiguation) is given and is used to find the correct sense of the word on a context. SVM is used. Cxv Term presence, vs term frequency, term position are described.[8]

Observations : SVM separate the 2 categories and build a wide gap which are root words and some are some of the affixes. Steps of algorithm contains: stop words removal, sorting single column word line and stemming performed on sorted list in which each word is compared with next 10 words assuming that minimum that 10 morphological variants is present in list. If word is present as substring in next word then the word is broken as substring + remaining characters of word. Substring is treated as root/base form and remaining characters are treated as affix.

Limitation : Term position is important. Words appearing in the 1st few sentences and last few sentences in a text are given more weightage than those appearing elsewhere.

The following table presents the summary of literature.

Paper[8] A Framework for Sentiment Analysis in Hindi using HSWN

Pooja Pandey & Sharvari Govilkar (2015) used HindiSentiWordNet (HSWN) for Hindi movie review to find the overall sentiment associated with the document. Polarity of the words in the review are extracted from HSWN and then final aggregated polarity is calculated which can sum either positive, negative or neutral. Synset replacement algorithm is used to find polarity of those words which don't have polarity associated with it in HSWN. Negation and discourse relations which are mostly present in Hindi movie review are also handled to improve the performance of the system.[9]

Observation : Existing HSWN is improved by adding missing sentimental words related to Hindi movie domain. The system comprises two stages:

1. Improving HindiSentiWordNet (HSWN)

2. Sentiment extraction.

During the first stage they are improving the existing HSWN with the help of English SentiWordNet, where sentimental words which are not present in the HSWN were translated to English and then searched in English SentiWordNet to retrieve their polarity. In the second stage, sentiment is extracted by finding the overall polarity of the document; which can be positive, negative or neutral. Here during pre-processing tokens are extracted from sentence and spell check is performed. Rules are devised for handling negation and discourse relation which highly influence the sentiments expressed in the document. Finally, overall sentiment orientation of the document is determined by aggregating the polarity values of all the sentimental words in the document.[9]

APPLICATIONS[10]

These are the applications of sentiment analysis.

- In social media monitoring
  - VOC to track customer reviews, survey responses, competitors, it is also practical for use in business analytics and situations in which text needs to be analyzed.
  - Computing customer satisfaction metrics :We can get an idea of how happy customers are with your products from the ratio of positive to negative tweets about them.
- Identifying detractors and promoters
  - It can be used for customer service, by spotting dissatisfaction or problems with products.
  - It can also be used to find people who are happy with your products or services and their experiences can be used to promote your products.
- In finance firms/markets
  - To forecast market movement based on news, blogs and social media sentiment.
  - To identify the clients with negative sentiment in social media or news and to increase the margin for transactions with them for default protection.
  - There are numerous news items, articles, blogs, and tweets about each public company. A sentiment analysis system can use these various sources to find articles that discuss the companies

and aggregate the sentiment about them as a single score that can be used by an automated trading system. One such system is The Stock Sonar. This system (developed by Digital Trowel) shows graphically the daily positive and negative sentiment about each stock alongside the graph of the price of the stock.

- Reviews of consumer products and services :There are many websites that provide automated summaries of reviews about products and about their specific aspects. A notable example of that is "Google Product Search."
- Monitoring the reputation of a specific brand on Twitter and/or Facebook : One application that performs real-time analysis of tweets that contain a given term is tweet feel.
- Enables campaign managers to track how voters feel about different issues and how they relate to the speeches and actions of the candidates.
- Applications in business domain; Consider a question : "why aren't customers buying our products?" or "why aren't customers visiting our website?"We know the concrete data: price, specs, competition, etc.
- In politics/political science; Evaluation of public/voters opinions. Views/discussions of policy.
- Law/policy making.
- Sociology;
- Psychology : investigations or experiments with data extracted from NL text.

Table 1 : Summary of Literature Survey for Sentiment Analysis for Hindi Language

Sr. No.	Title of the paper	Author & Year of publication	Observations/Remarks
1	A Fall-back Strategy for Sentiment Analysis in Hindi: a Case study	Aditya Joshi, Balamurali A R, Pushpak Bhattacharya, 2010	They used 3 approaches: training classifier, translate given document to English & develop lexical resource called Hindi-SentiWordNet (H-SWN).Naive approach, SentiWordNet 1.1 is used. WordNet linking is used to map synsets of WordNets of different languages.POS tagging is done. 250 Hindi movie reviews. RapidMiner5.0 for document classification. LibSVM type-C learner for classification. TF-IDF gave highest accuracy of 78.19 & 60.31for resource based SA.
2	Hindi Subjective Lexicon : A Lexical Resource for Hindi Polarity Classification	Bakliwal et al. (2012)	Created lexicon using a graph based method. They explored how the synonym and antonym relations can be exploited using sample graph traversal to generate the subjectivity lexicon. Their proposed algorithm achieved approximately 79% accuracy.
3	Sentiment Analysis in Hindi	Naman Bansal and Umair Z. Ahmed. Advisor: Amitabha Mukherjee, IIT Kanpur	They used Semi-Supervised approach to train Deep Belief Network on small percentage of labelled data (150) and assign polarity to unlabelled data. 300 sentences dataset stored in xml. Sentence lever sentiment is adopted. Data preprocessing stage, Deep belief architecture divided in 2 stages: pre-training model & fine tuning step. They report 71% accuracy using DBN on English language, 76% using active deep learning and 64% on Hindi language using DBN.
4	A Hybrid Approach for Twitter	Namita Mittal et al. (2013)	A 3 stage hierarchical model for automatically classifying twitter messages is proposed by the author; 1) labeling with emoticans; 2) using predefined list of words with strong positive or negative

	Sentiment Analysis		sentiments; 3) tokens are weighted based on subjectivity lexicon. Maximum Accuracy obtained by hybrid approach with discourse was 73.72%.
5	Sentiment Analysis of Hindi Review based on Negation and Discourse Relation	Namita Mittal, Basant Agarwal, Garvit Chauhann, Nitin Bania, Prateek Pareek	Method for improving HSWN is proposed.770 reviews. N-gram & POS-tagged N-gram approaches used. Fleiss kappa coefficient value 0.8092 & dataset has 104 words. Accuracy of 82.89% for positive reviews, 76.59% for negative and overall 80.21%. for improved HSWN + negation + Discourse.
6	A Survey on Sentiment Analysis and Opinion Mining Techniques	Amandeep Kaur, Vishal Gupta, 2013	The process of Sentiment Analysis is divided into 5 steps : process of Sentiment Analysis for Text (Lexicon generation), Subjectivity detection, sentiment polarity detection using Network Overlap Technique, sentiment structurization, sentiment summarization-visualization-tracking.
7	Sentiment Classification in Hindi	Sneha Mulatkar, May 2014	WSD algorithm (Sense disambiguation) is given and is used to find the correct sense of the word on a context. SVM is used. Term presence, vs term frequency, term position are described.
8	A Framework for Sentiment Analysis in Hindi using HSWN	Pooja Pandey, Sharvari Govilkar, June 2015	Existing HSWN is improved with the help of English SentiWordNet, where sentimental words which are not present in the HSWN are translated to English and then searched in English SentiWordNet to retrieve their polarity. Sentiment is extracted by finding the overall polarity of the document; which can be positive, negative or neutral. During pre-processing tokens are extracted from sentence and spell check is performed. Rules are devised for handling negation and discourse relation which highly influence the sentiments expressed in the document. Finally, overall sentiment orientation of the document is determined by aggregating the polarity values of all the sentimental words in the document.

CONCLUSION

Sentiment analysis analyzes the sentiment information from the input like product review, movie review, news review, comments from blogs or any social website in order to draw the conclusion and understand the overall situation. In order to understand data machine needs to understand what are sentiments from the input and give output as positive or negative or neutral review.

Large amount of work in sentiment analysis has been done in English language, as English is a global language, but there is a need to perform sentiment analysis in other languages also. Large amount of other languages contents are available on the Web which needs to be mined to determine the sentiment.

Various algorithms and classifiers are present for sentiment analysis which depends on the Hindi data. Hence, we conclude with a survey of various approaches used specifically for Hindi language.

ACKNOWLEDGEMENTS

I am using this opportunity to express my gratitude to thank all the people who contributed in some way to the work described in this paper. My deepest thanks to my project guide for giving timely inputs and giving me intellectual freedom of work. I express my thanks to the head of computer department and to the principal of Pillai Institute of Information Technology, New Panvel for extending his support.

REFERENCES

[1] Daniel Jurafsky and James H. Martin, Speech and Language Processing, "An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition", University of Colorado, Boulder, Pearson Education series,2011

[2] Aditya Joshi,Balamurali AR, Pushpak Bhattacharya, Dept of Computer and Science Engineering, IIT Bombay, IITB-Monash Research Academy, IIT Bombay, Mumbai, India-400076, "A Fall-back Strategy for Sentiment Analysis in Hindi: a Case study ", Proceedings of ICON 2010:8th International Conference on Natural Language Processing.

[3] Akshat Bakliwal, Piyush Arora, Vasudeva Varma, "Hindi Subjective Lexicon : A Lexical Resource for Hindi Polarity Classification", Proceedings of the Eight International Conference on Language Resources & Evaluation(LREC), 2012

[4] Naman Bansal and Umair Z Ahmed, Advisor: Amitabha Mukherjee, Department of Computer Science and Engineering , Indian Institute of Technology, Kanpur, India, "Sentiment Analysis in Hindi"

[5] Namita Mittal et al., "A Hybrid approach for Twitter Sentiment analysis", 2013

[6] Namita Mittal, Bansat Agarwal, Garvit Chauhan, Nitin Bania, Prateek Pareek, "Sentiment Analysis of Hindi review based on Negation and Discourse Relation", International Joint Conference on Natural Language Processing, pg.45-50,Nagoya, Japan,14-18 Oct 2013.

[7] Amandeep Kaur and Vishal Gupta, "A Survey on Sentiment Analysis and Opinion Mining Techniques", Journal of Emerging Technologies in Web intelligence, Vol. 5,no. 4,Nov 2013

[8] Sneha Mulatkar, "Sentiment Classification in Hindi", International Journal of Scientific & Technology research(IJSTR 2014), Vol.3 , Issue 5, May 2014

[9] Pooja Pandey, Sharvari Govilkar, "A Framework for Sentiment Analysis in Hindi using HSWN", International Journal of Computer Applications(IJCA) (0975-8887), Vol.119-No.19, June 2015

[10] Mukesh Yadav, Varunakshi Bhojane, "Data Analysis & Sentiment Analysis for Unstructured D ata", International Journal of Engineering Technology, management and Applied Sciences (IJETMAS), Vol.2,Issue 7,Dec 2014 ISSN 2349-4476