# Sentiment Classification In Hindi

Sneha Mulatkar

**Abstract**: Sentiment Analysis is a Natural Language Processing and Information Extraction task that aims to obtain writer's feelings expressed in positive or negative comments, questions and requests, by analyzing a large numbers of documents. Generally speaking, sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall tonality of a document. In recent years, the exponential increase in the Internet usage and exchange of public opinion is the driving force behind Sentiment Analysis today.

**Index Terms**: Corpus, Gloss, Synset, WordNet, Disambiguation, Lexical analysis, Part-of-speech

## 1 INTRODUCTION

Sentiment Analysis (SA) is the task of prediction of opinion in text. Sentiment classification deals with tagging text as positive, negative or neutral from the perspective of the speaker/writer with respect to a topic. The first points deals with evaluating sense-based features against word-based features. The second issue that we address is in fact an opportunity to improve the performance of SA that opens up because of the choice of sense space. Since sense-based features prove to generate superior sentiment classifiers, we get an opportunity to mitigate unknown synsets in the test corpus by replacing them with known synsets in the training corpus. Note that such replacement is not possible if word-based representation were used as it is computationally not possible to make such large number of similarity comparisons. We use the corpus by Ye et al. (2009) that consists of travel domain reviews marked as positive or negative at the document level. Polarity Identification focuses on the classification of positive, negative or neutral expressions in texts. Polarity-related term feature interpretation, most of the proposed methods make use of manually annotated or automatically constructed lists of polarity terms. WordNet was designed to establish the connections between four types of Parts of Speech (POS) - noun, verb, adjective, and adverb. The smallest unit in a WordNet is synset, which represents a specific meaning of a word. It includes the word, its explanation, and its synonyms. The specific meaning of one word under one type of POS is called a sense. Each sense of a word is in a different synset. Synsets are equivalent to senses = structures containing sets of terms with synonymous meanings. Each synset has a gloss that defines the concept it represents. For example, the words night, nighttime, and dark constitute a single synset that has the following gloss: the time after sunset and before sunrise while it is dark outside. Synsets are connected to one another through explicit semantic relations. Some of these relations (hypernym, hyponym for nouns, and hypernym and troponym for verbs) constitute is-a-kind-of (holonymy) and is-a-part-of (meronymy for nouns) hierarchies. For example, tree is a kind of plant, tree is a hyponym of plant, and plant is a hypernym of tree.

## 2 MOTIVATION

The main motivation behind our approach is that users that are somehow 'connected' may be more likely to hold similar opinion therefore relationship information can complement what we can extract about a user's viewpoint. What other people think has always been part in decision making:

"Do you think I should buy this camera?"
"Why do you vote for X?"

Before the spread of the Internet people used to ask friends. Now a days Internet contains a huge amount of opinions in forums, blogs, review sites, tweet, comment.But it is now always easy to find and analyze the opinions needed for decision making. Sentiment analysis is classifying the polarity. Most of the research has been done in English language and work has been done. We propose to detect the sentiment of Hindi language. [10] The challenge is Hindi is rich and is free order language as compared to English. Also the scarcity of resource for Hindi language brings challenges from collection and generation of database.

## 3 LITERATURE SURVEY

As demonstrated in this document, the numbering for sections upper case Arabic numerals, then upper case Arabic numerals, separated by periods. Initial paragraphs after the section title are not indented. Only the initial, introductory paragraph has a drop cap. Web content mining is intended to help people discover valuable information from large amount of unstructured data on the web. Movie review mining classifies movie reviews into two polarities: positive and negative. [7] As a type of sentiment-based classification, movie review mining is different from other topic-based classifications. The main objective of this work is to classify a large number of opinions using web-mining techniques into bipolar orientation (i.e. either positive or negative opinion). Such kind of classification could help consumers in making their purchasing decisions. Research results along this line can lead to users' reducing the time on reading threads of text and focusing more on analyzing summarized information. Review mining can be potentially applied in constructing information presentation.

### Features for Sentiment Analysis

Feature engineering is an extremely basic and essential task for Sentiment Analysis. Converting a piece of text to a feature vector is the basic step in any data driven approach to SA. In the following section we will see some commonly used features used in Sentiment Analysis and their critiques.

### Term Presence vs. Term Frequency

Term frequency has always been considered essential in traditional Information Retrieval and Text Classification tasks. But Pang-Lee et al. (2002) found that term presence is more important to Sentiment analysis than term frequency. [1] That is, binary- valued feature vectors in which the entries merely indicate whether a term occurs (value 1) or not (value 0). This is not counter intuitive as in the numerous examples we saw before that the presence of even a single string sentiment bearing words can reverse the polarity of the entire sentence. It has also been seen that the occurrence of rare words

contain more information than frequently occurring words, a phenomenon called Hapax Legomena.

**Term Position**

Words appearing in certain positions in the text carry more sentiment or weightage than words appearing elsewhere. This is similar to IR where words appearing in topic Titles, Subtitles or Abstracts etc are given more weightage than those appearing in the body. Although the text contains positive words throughout, the presence of a negative sentiment at the end sentence plays the deciding role in determining the sentiment. Thus generally words appearing in the 1st few sentences and last few sentences in a text are given more weightage than those appearing elsewhere.

## 4 THE SYSTEM

Sense Disambiguation (WSD) is defined as the task of finding *the correct* sense of the word in a context. The task needs large amounts of word and word knowledge. Let us consider the word संबंध in the following Hindi sentence: In this particular case, sense 1 is the most appropriate one, though sense 5 and 6 too are relevant.

| Relation | Meaning |
|---|---|
| Hypernymy/Hyponymy | Is-A(Kind-Of) |
| Entailment/Troponymy | Manner-Of(for verbs) |
| Meronymy/Holonymy | Has-A(Part-Whole) |

WSD Algorithm: Finding the word's Correct Sense. For a polysemous word w needing diambiguation, a set of context words in its surrounding window is collected. Let this collection be C, the context bag. For each sense s of w, do the following
Let B be the bag of words obtained from the Synonyms
Glosses

Example Sentences
Hypernyms
Glosses of Hypernyms
Example Sentences of Hypernyms
Hyponyms
Glosses of Hypernyms
Example Sentences of Hypernyms
Meronyms
Glosses of Meronyms
Example Sentences of Meronyms
Measure the overlap between C and B using the intersection similarity measure.

**3. Output that the sense s as the most probable sense which has the maximum overlap. [9]**

Figure gives the pictorial description of the basic idea of the strategy. The idea behind using the intersection similarity measure is to capture the belief that there will be high overlap between the words in the context and the *related words* found from the wordnet lexical and semantic relations and glosses. The belief that there will be high overlap between the words in the context and the related words found from the wordnet lexical and semantic relations and glosses[9].

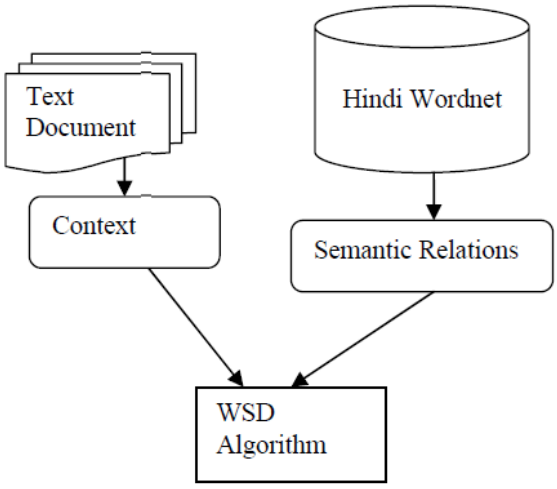


**Figure:** Extracting semantic relations from Wordnet and building context from the text for WSD

**SVM (Support Vector Machine)**

It classify the sentiment is positive or negative. It is commonly used package and is freely available

It analyze data and recognize pattern used for classification and regression analysis

Set of training examples each marked as belonging to one of two categories

SVM separate the 2 categories and build a wide gap.
सदस्य, बल्लेबाज, दृश्य, आश्रम are the root words
यों, ओं, अन, अनीय, are some of affixes

The steps of algorithm are as follows:

1. Stop words removal: Words in Hindi word list with length less than 3 are considered as stop words and are removed from the list at time of preprocessing.
2. Sorting single column word file: Bubble sorting algorithm is applied for sorting words present in single column word file with N number of words[10].
3. Stemming performed on sorted list:
   - **I.** Each word is compared with next 10 words assuming that maximum of 10 morphological variants is present in list.
   - **II.** If word is present as substring in next word then the word is broken as substring + remaining characters of word.
   - **III.** Substring is treated as root/base form and remaining characters are treated as affix[10].

| Word from Hindi word list | Base form |
|---|---|
| बालिकाएं | बालिका |
| सरकारिया | सरकार |
| बुराईयों | बुराई |
| मनुष्यता | मनुष्य |
| अदालती | अदालत |
| अदाकारा | अदाकार |
| सम्वेदनाओं | सम्वेदना |

## ACKNOWLEDGMENT

I owe a great many thanks to a many people who helped and supported me. My deepest thanks to the Guide of the project for guiding and correcting various documents of mine with attention and care..I express my thanks to the Principal of, Pillai Institute of Information Technology, New Panvel for extending his support.

## REFERENCES

[1] "HarnessingWordNet Senses for Supervised Sentiment Classification",Balamurali A,Aditya Joshi,Pushpak Bhattacharyya IITB-Monash Research Academy, IIT Bombay Dept. of Computer Science and Engineering, IIT Bombay.

[2] "A systematic Approach towards the Solution of the Polysemy Problem in Natural Language Processing" ,Abed Alhakim Freihat April 2011.

[3] "Sentiment Classification of Reviews Using SentiWordNet", Ohana, B., Tierney, B.: Sentiment classification of reviews using SentiWordNet. 9th. IT&T Conference, Dublin Institute of Technology, Dublin, Ireland.

[4] "Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches",Pimwadee Chaovalit *Department of Information Systems University of Maryland, Baltimore County* Lina Zhou *Department of Information Systems University of Maryland, Baltimore County.*

[5] "Sentiment Classification in Movie Reviews",An Approach Using Subjectivity Filtering Daniel Pomerantz, McGill University.

[6] "Sentiment Analysis, Indian Institute of Technology",Subhabrata Mukherjee, Bombay Department of Computer Science and Engineering, June 29, 2012.

[7] "Approach to Sentiment Analysis: Analytical Categories and Issues of Automation", Repindex.

[8] Peter D. Turney, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews," presented at the Association for Computational Linguistics 40th Anniversary Meeting, New Brunswick, N.J., 2002.

[9] Hindi Word Sense Disambiguation,Manish Sinha Mahesh Reddy Department of computer Science and Engineering Indian Institute of Technology Bombay,Mumbai.

[10] Hindi Stemmer Anubha Jain and Sujoy Das ,Research scholar, Department of Computer Applications, MANIT, Bhopal ,Associate Professor, Department of Computer Applications, MANIT, Bhopa