RESEARCH PAPER | Mathematics | Volume : 4 | Issue : 8 | August 2014 | ISSN - 2249-555X

Sentiment Analysis for Hindi Text using Fuzzy Logic

KEYWORDS: Opinion Mining, Fuzzy logic, Part of speech, NLP.

Shweta Rana

Department Of Mathematics, Amity University, Haryana

ABSTRACT *With the Web, especially with the explosive growth of the user generated content on the Web, the world has changed. One can post reviews of products at merchant sites and express views on almost anything in Internet forums, discussion groups, and blogs, which are collectively called the user generated content. Now if one wants to buy a product, it is no longer necessary to ask one's friends and families because there are plentiful of product reviews on the Web which give the opinions of the existing users of the product. .Finding opinion sources and monitoring them on the Web, however, can still be a formidable task because a large number of diverse sources exist on the Web and each source also contains a huge volume of information. In many cases, opinions are hidden in long forum posts and blogs. It is very difficult for a human reader to find relevant sources, extract pertinent sentences, read them, summarize them and organize them into usable forms. An automated opinion mining and summarization system is thus needed. Opinion mining, also known as sentiment analysis, grows out of this need. Many approaches have been developed for sentiment mining for English , Chinese and Arabic texts. Here in my work I have proposed a method for sentiment mining for hindi text.*

I. Introduction

Textual information in the world can be broadly classified into two main categories, facts and opinions. Facts are objective statements about entities and events in the world. Opinions are subjective statements that reflect people's sentiments or perceptions about the entities and events. Much of the existing research on text information processing has been focused on mining and retrieval of factual information, e.g., information retrieval, Web search, and many other text mining and natural language processing tasks. One of the main reasons for the lack of study on opinions is that there was little opinionated text before the World Wide Web. Before the Web, when an individual needs to make a decision, he/she typically asks for opinions from friends and families. When an organization needs to find opinions of the general public about its products and services, it conducts surveys and focused groups. With the Web, especially with the explosive growth of the user generated content on the Web, the world has changed. One can post reviews of products at merchant sites and express views on almost anything in Internet forums, discussion groups, and blogs, which are collectively called the user generated content. Now if one wants to buy a product, it is no longer necessary to ask one's friends and families because there are plentiful of product reviews on the Web which give the opinions of the existing users of the product. For a company, it may no longer need to conduct surveys, to organize focused groups or to employ external consultants in order to find consumer opinions or sentiments about its products and those of its competitors. Finding opinion sources and monitoring them on the Web, however, can still be a formidable task because a large number of diverse sources exist on the Web and each source also contains a huge volume of information. In many cases, opinions are hidden in long forum posts and blogs. It is very difficult for a human reader to find relevant sources, extract pertinent sentences, read them, summarize them and organize them into usable forms. An automated opinion mining and summarization system is thus needed. Opinion mining, also known as sentiment analysis, grows out of this need. In this Dissertation we have introduced a method for extracting the opinion for Hindi text.

II. OPINION MINING

Opinion Mining is a new and exciting field of research concerned with extracting opinion related information from textual data sources. It has the potential for a number of interesting applications both in commerce and academic areas, and poses novel intellectual challenges, which continues to attract considerable research interest. In this section the research field of opinion mining is introduced, its motivations, key tasks and challenges are discussed in more details.

III. PROPOSED SOLUTION

The proposed method is based on the combinations of opinion words around each product feature in a review sentence. This methodology extracts the sentiment from hindi text and determines the strength of opinion orientation (very weak, weak, moderate, very strong and strong) on the product feature using fuzzy logic technique.

For example, consider these sentences:

1) We extremely enjoy this camera.
2) We like this camera.
3) The picture quality is very good.
4) The picture quality is good

By human interpretation it is obvious that the intensity of sentences l and 3 is more than sentence 2 and 4. As various methods classify these sentences into the positive, negative and natural classes while the proposed method aims to classify these sentences into the granularity levels (i.e. very weak, weak, moderate, very strong and strong by combining opinion words (i.e. Adverb, adjective and verb). Here first we gather reviews and then we find the opinion words and phrases. Opinion (sentiment) words and phrases are words and phrases that express positive or negative sentiment. Although words that express positive and negative orientation are usually adjective and adverb, verb and noun can be used to express opinion. Consider that sentence: "very" as adverb, "good" as adjective, "enjoy" and "like "as verb.

Researchers have compiled set of words and phrases for

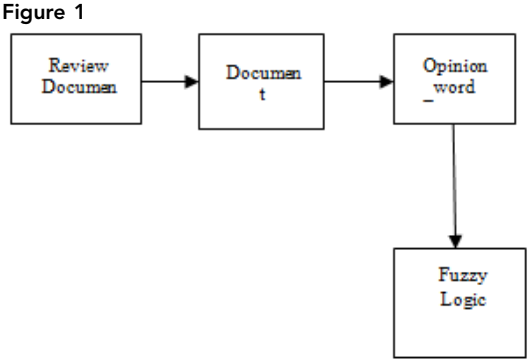
Volume : 4 | Issue : 8 | August 2014 | ISSN - 2249-555X

adjective, adverb and verb and noun respectively. Such lists are collectively called the opinion lexicon. Each set usually obtained through a bootstrapping process. However, the lists only have opinion words that are adjectives and adverbs, we added verb and noun lists identified in the same way. In order to utilize different lists, we need to perform part-of-speech (POS) tagging as many words can have multiple POS tags depending on their usages. The part-of-speech of a word is a linguistic category that is defmed by its syntactic or morphological behavior. Common P O S categories in Hindi are: noun, verb, adjective, adverb, pronoun, proposition, conjunction and inteljection. We use NLP parser for P O S tagging.

IV. BLOCK DIAGRAM

The block diagram of our proposed method is given below. It consists of four blocks namely Review Documemt, Documenting Preprocessing,, Opinion Word Extractor and Fuzzy Logic. Document preprocessing is to preprocess the review collection for further proceedings. Feature Decider is to identify and extract frequent features of the product, which is commented on. Opinion Word/Phrase Extractor is to collect the opinionated phrases about the product features. Adjective/Adverb Intensity Map is the ontology of the list of weighted opinion words. Intensity Finder is to measure the weight of the opinion phrases using fuzzy approximation and to rank products accordingly.

Figure 1



V. DOCUMENT PROCESSING

We processed only the description part of each review, here processing means removing all the logos, images, and other graphics from the reviews and split review into sentences to create a plain text file of reviews.

Part-of-Speech Tagging (POS Tagging)

Not all the words in review sentences are useful for identifying product features and orientations of the discussed product. The nouns and noun phrases in the sentences are likely to be the features that customers comment on, while adjectives are often used to express opinions and feelings. The following review sentences are typical examples from reviews of a digital camera: "The picture quality is nice and the pictures are fantastic"; "The buttons are easy to use and nice". We collect frequent nouns and adjectives, adverbs from the review file after part-of-speech (POS) tagging using a POS tagger. Applying the POS tagger, above review sentence tagged as "The/DT picture/NN quality/NN is/VBZ nice/JJ and/CC the/DT pictures/NNS are/VBP fantastic/JJ ./.". The slashes with the followed up capital letters in the tagged sentence indicate the type of words1 before them. For example, "nice" is tagged with "/JJ", indicating it's an adjective in the sentence. POS taggers are not always perfect and review sentences may be quite complex or irregular. Therefore, tagging errors cannot be avoided. We make a list of adjectives, adverbs that are generally used to qualify the features of a product and nouns, which are usually used as features of a product. It's a one-time effort to create the file for a specific domain of product like Digital Cameras, Mobile Phones, Color TV etc.

Tag Part of speech

Table 1.

CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun
PRPS	Possesive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	To
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Wh-determiner
WP	Wh-pronoun
WPS	Possesive wh-pronoun
WRB	Wh-adverb

Example

संसद और वधिानसभा में नहलिाओं के लएि ३३ प्रतशित आरक्षण की मांग ।

Output

संसद <NO_TAG> और <CONJ> वधिानसभा <NOUN> में <POST_PREPOSITION> नहलिाओं <NOUN> के <POST_

Volume : 4 | Issue : 8 | August 2014 | ISSN - 2249-555X

PREPOSITION> लिए <NO_TAG> ३३<NUM> प्रतिशत <NOUN> आरक्षण <NO_TAG> की <POST_PREPOSITION> मांग <NOUN>! <PUNC>

सुप्रीम कोर्ट ने बुधवार को भूमि घोटाले के सिलसिले ।

NOUN+PP+NOUN+PP+NOUN+PP+NOUN+PUNC

Stop Words Removal

We remove words like digits, prepositions, articles, and proper nouns like brand name of product (Cannon, Kodak) etc from the POS tagged review file, as they are redundant in our system. To create a list of words, which are to remove from the review, file is a one-time job and helps better extraction of opinion phrases/words from the POS tagged review file.

VI. OPINION WORD EXTRACTER

Opinionated words are mainly adjectives/adverbs and they are used to qualify nouns, we extract two or three consecutive words from the POS tagged review if their tags1 conform to any of the patterns in the table below. Here JJ tags indicate adjectives, NN tags are nouns, RB tags are adverbs and VB tags are verbs and so on. For example, Pattern 2 indicates that three consecutive words are extracted if the first word is an adverb and the second word is an adjective/adverb, and the third word is a noun.

Extracted word pattern
Table 2.

Pattern	First Word	Second Word	Third Word
Pattern 1	JJ	NN/NNS	NN/NNS
Pattern 2	RB/RBR/RBS	JJ/RB/RBR/RBS	NN/NNS
Pattern 3	RB/RBR/RBS	VBN/VBD	-
Pattern 4	VBN/VBD	NN/NNS	-

We collect all opinionated phrases (mostly 2/3 words phrases) like (Adjective, Noun), (Verb, Noun), (Adverb, Adjective), (Adverb, Adjective, Noun) (Adjective, Noun, Noun) etc. from the processed POS Tagged Review File and accepted only those Opinion Phrases in which the words in the opinion oriented phrases are either in Weight File (containing Adjectives/Adverbs/Verbs) or in Feature List (containing Nouns).

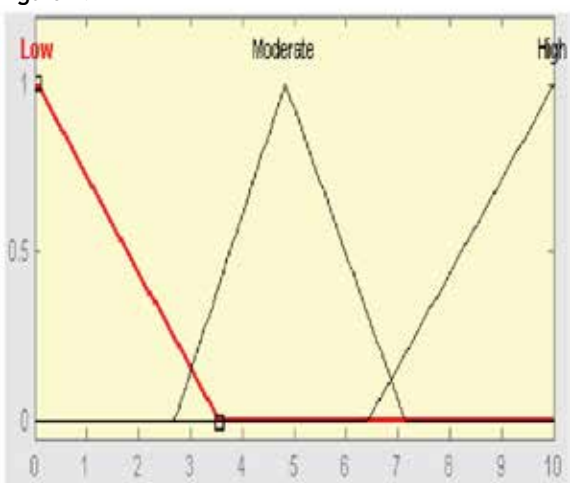
VII. FUZZY LOGIC

At first the inputs should become as fuzzy data; in our method we have four inputs following as: adjective, verb, adverbs and noun which are known as opinion words. Special degree for each of these words are associated by human expert, for example: like: 4 love: 5, good: 3, excellent: 6, really: 5, extremely: 9, enjoy:8, very: 5.

VII. MEMBERSHIP FUNCTION

The membership function is a graphical representation of the magnitude of participation of each input. It associates a weighting with each of the inputs that are processed, define functional overlap between inputs, and ultimately determines an output response. The rules use the input membership values as weighting factors to determine their influence on the fuzzy output sets of the final output conclusion. It is defined for finding membership value for each of the inputs. In general, there are three types of MF, namely triangular, trapezoidal, and generalized bell-shape. In proposed technique triangular Membership Function is used. Rank of MF is decelerated by human experts; the linguistic variable used to represent them as divided into three levels: low, moderate and high. Below figure shows membership function that we use in our method.

Figure 2.



Following is the triangular membership function (MF) to obtain the value:

$$\mu(a,b,c,d) = \max(\min((a-b/c-b) , (d-a/d-c)) , 0)$$

After calculating we get a value between 0 and 1. We mark 0 as very strongly negative, 0.5 as neutral and 1 as very strongly positive. Suppose we get the value for the sentence s as $\mu(s) = 0.8$ so we can say it is strongly positive.

VIII. CONCLUSION AND FUTURE WORK

We have proposed a fuzzy logic method for identifying semantic orientation of opinions for Hindi text. The method is able to classify reviews into different classes: very strong negative, strong negative, neutral, strong positive and very strong positive. As a future work, we can implement this method intent to implement dataset. We expect the accuracy of classification will be increased by combining opinion words (i.e. Adverb, A adjective and verb).We believe that there is rich potential for future research. We believe that this method will become increasingly important as more people can express their opinions/experiences on the Web in Hindi. Our experimental results indicate that the proposed method are effective in decision making about a review. In our future work, we plan to group features according to the strength of the opinions that have been expressed on them, e.g., to determine which features strongly like and dislike. This will further improve the feature extraction and the subsequent summarization.

Volume : 4 | Issue : 8 | August 2014 | ISSN - 2249-555X

REFERENCE

[1] Bo Pang, Lillian Lee "Opinion Mining and Sentiment Analysis", Foundations and Trends in IR, Vol-2pp1-135, 2008 | | [2] B. Liu, M. Hu, J Cheng "Opinion Observer: Analyzing and Comparing Opinions on the Web", IW3C2, Chiba, Japan, May 10-14, 2005 | | [3] Hu, M., and Liu, B. Mining and summarizing customer reviews. In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004) (2004), pp. 168–177. | | [4] M. Hu and B. Liu, "Mining and summarizing customer reviews " in Proceedings of the 2004 SIGKDD, 2004 pp. 168-177 | | [5] A. Andreevskaia and S. Bergler. Mining WordNet for Fuzzy Sentiment: Sentiment Tag Extraction from WordNet Glosses. In EACL'06, pp. 209-215,2006. | | [6] P. Beineke, T. Hastie, C. Manning, and S. Vaithyanathan. An Exploration of Sentiment Summarization. In Proc. Of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications, 2003. | | [7] R. Al-Shalabi and R. Obeidat , "Improving KNN Arabic Text Classification with N-Grams Based Document Indexing", in Proceedings of the Sixth International Conference on Informatics and Systems, Cairo, Egypt, 2008. | | [8] Kevin Duh and Katarin Kirchoff. 2004. Pos tagging of dialectal arabic: A minimally supervised approach. | | [9] Smriti Singh, Kuhoo Gupta, Manish Shrivastava, and Pushpak Bhattacharyya. 2006. Morphological richness offsets resource demand – experiences in constructing a pos tagger for hindi. In Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pages 779–786, Sydney, Australia, July. Association for Computational Linguistics. | | [10] Akshar Bharati, Vineet Chaitanya, and Rajeev Sangal. 1995. Natural Language Processing – A Paninian Perspective. Prentice-Hall India. | | [11] John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the 18th International Conf. on Machine Learning, pages 282–289. Morgan Kaufmann, San Francisco, CA. | | [12] Jurafsky D and Martin J. Speech and Language Processing. Pearson Edu. 2000. | | [13] Dermatas E. and Kokkinakis G. 1995. Automatic stochastic tagging of natural language texts. Computational Linguistics, 21(2), 137-164. | | [14] N. Kaji and M. Kitsuregawa. Automatic Construction of PolarityTagged Corpus from HTML Documents. COLING/ACL'06, 2006. | | [15] A-M. Popescu and O. Etzioni. Extracting Product Features and Opinions from Reviews. EMNLP-05, 2005. | | [16] A. Andreevskaia and S. Bergler. Mining WordNet for Fuzzy Sentiment: Sentiment Tag Extraction from WordNet Glosses. In EACL'06, pp. 209-215,2006. | | [17] S. Kim and E. Hovy. Determining the Sentiment of Opinions. COLING'04, 2004. | | [18] L. Zhuang, F. Jing, X.-Yan Zhu, and L. Zhang. Movie Review Mining and Summarization. CIKM-06, 2006. | | [19] Jeonghee Yi, Tetsuya Nasukawa. 2004. "Sentiment Analysis: Capturing Favorability Using Natural Language Processing" In K-CAP 03, October 23-25, Florida, USA. | | [20] Jeonghee Yi, Nasukawa, Tetsuya, Bunescu, Razvan, Niblack, Wayne, (2003) "Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques," icdm, , Third IEEE International Conference on Data Mining (ICDM'03). Florida, pp.427 | | [21] Narayanan, Ramanthan, Liu, Bing & Choudhary, Alok (2009) "Sentiment Analysis of Conditional Sentences". Proceedings of Conference on Empirical Methods in Natural Language Processing, Singapore |