European Journal of Academic Essays, Special Issue (1): 44-48, 2014
ISSN: 2183-1904
www.euroessays.org

Recent Trends in Analysis of Algorithms and Complexity Theory

Opinion Mining for online Oriya Text

< Jena Manoj Kumar >[1], < Balabantaray Rakesh Ch. >[2]

[1] MLE-MIS Officer
Orissa Community Tank management Project, District project unit,Angul
Dept. of Water Resources, Govt. of Odisha.
manojoctmp@gmail.com

[2] Associate Professor
Dept.of Computer Science & Engineering
IIIT, Bhubaneswar
rakeshbray@gmail.com

Abstract: The paper aims at discussing the opinion mining technique can be applicable for analyzing online oriya text. The Oriya language is a classical language belongs to Indo Aryan family. The opinions in oriya are subjective information which represents user's sentiments, feelings or appraisal related to the same. The concept of opinion is very broad. It can be positive, negative and neutral. In this paper we focus on the various opinion mining techniques that can facilitate the task of opinion mining. The precise method for predicting opinions enable us, to extract sentiments from the web and foretell the feeling of the person, which could prove valuable for research at various level. Much of the research work had been done on the processing of opinions or sentiments recently because opinions are so important that whenever we need to make a decision we want to know others' opinions. This opinion is not only important for a user but is also useful for an organization.

Keywords: Free word order, opinion holder polarity, sentiment.

---

1. Introduction

Opinion mining can be viewed as a kind of processing of natural language for tracking the attitudes, feelings or appraisal of the public about particular topic, product or services. All information available in web are of two types: facts and opinions. The facts are the objective expressions which describe the entities, events and properties whereas the opinion is the subjective expression which describes people's opinions, emotions and sentiments towards entities and their properties. The current search engine searches for facts because they assume the facts are true and can be expressed with keywords. But these search engines do not find the opinions because opinions or sentiments are very difficult to express by keywords and that is why there ranking strategy are not appropriate for opinion retrieval.

Now, the web has significantly changed its way that people comments their views and opinion on any product and services. User can post their comments on any internet forums, review sites, blogs and discussion group which are commonly known as user generated content which contains the important information. The opinion in oriya available in web on a particular topic is limited. So little emphasis is given to extract opinion from online text. The characteristic of oriya language is free word order means the arrangement of word less mater in organization of sentences. The art Opinion Mining is to recognize the subjectivity and objectivity of a text and further classify the opinion orientation of subjective text. In short we say that Opinion Mining is an automated extraction of subjective content from text and identifying the orientation such as positive or negative in that text. It aims to explore feelings of a person

Corresponding Author: Manoj Kumar Jena

who write the text. It used Natural Language Processing and Machine Learning ethics to determine opinion in the text. The evaluation of opinion can be done in two ways:

1. Direct opinion, gives positive or negative opinion about the object directly. For example, in oriya "Rama jane vala sasaka thile"(Ram was a good administrator) expresses a direct opinion.

2. Comparison means to compare the object with some other similar objects. For example, in oriya "Ramanka sasana Ravananka sasan sange tulana jogya nuhe. (Ram's administration can't be comparable with Ravan's administration.)

2. Characteristics of Oriya language:

Oriya is a Classical language which is of indo Aryan origin. It is free in terms of sentence constructionage which follows Paninian framework with even Vibhaktis (case relation) viz. karta, Karma,karan Sampradan, Apadaan, Sambandha and Adhikaran to assign case roles to noun entities. According to vibhaki and inflections the sentences is analysed as the oriya language is highly inflectious the sentences are divided into different part of speech tag. The part of speech tags can be considered as a unit which can be used for extracting opinion from text.

Structure of Oriya sentence:

Ram Ravanaku sita pain teera dwara lankare marile.

(Ram killed Ravan with an arrow for Sita in Lanka)

Rama-karta/Ravan-Karma/teera-karan/sita-sampradan/ lanka-Adhikaran

The chunks can be identified as:

Rama-karta Ravanku- karma[1] Sita –sampradan -pain[2] terra-karan –dwara[3] lankare- Adhikaran- marile[4]

3. Classification of Opinion Mining at different level:

Generally when analyzing opinion mining the available online text are analyzed and that available on line text is treated as a document. That document tells about a particular topic. That topic is mined at three levels. Those are discussed below:

3.1 Task of Opinion Mining at Document level

Document level opinion mining is about classifying the overall opinion presented by the authors in the entire document as positive, negative or neutral about a certain object. The assumption is taken at document level is that each document focus on single object and contains opinion from a single opinion holder. Turney [4] present a work based on distance measure of adjectives found in whole document with known polarity i.e. excellent or poor. The author presents a three step algorithm i.e. in the first step; the adjectives are extracted along with a word that provides appropriate information. Second step, the semantic orientation is captured by measuring the distance from words of known polarity. Third step, the algorithm counts the average semantic orientation for all word pairs and classifies a review as recommended or not. In contrast, Pang et al. [5] present a work based on classic topic classification techniques. The proposed approach aims to test whether a selected group of machine learning algorithms can produce good result when opinion mining is perceived as document level, associated with two topics: positive and negative. He present the results using nave bayes, maximum entropy and support vector machine algorithms and shown the good results as comparable to other ranging from 71 to 85% depending on the method and test data sets. Apart from the document-level opinion mining, the next sub-section discusses the classification at the sentence-level, which classify each sentence as a subjective or objective sentence and determine the positive or negative

3.2 Task of opinion mining at Sentence level

I. The sentence level opinion mining is associated with two tasks [1] [2] [3]. First one is to identify whether the given sentence is subjective (opinionated) or objective. The second one is to find opinion of an opinionated sentence as positive, negative or neutral. The assumption is taken at sentence level is that a sentence contain only one opinion for e.g., "The picture quality of this camera is good." However, it is not true in many cases like if we consider compound sentence for e.g., "The picture quality of this camera is amazing and so is the battery life, but the viewfinder is too small for such a great camera", expresses both positive and negative opinions and we say it is a mixed opinion. For "picture quality" and "battery life", the sentence is positive, but for "viewfinder", it is negative. It is also positive for the camera as a whole. Riloff and Wiebe [6] use a method called bootstrap approach to identify the subjective sentences and achieve the result around 90% accuracy during their tests. In contrast, Yu and Hatzivassiloglou [7] talk about sentence classification (subjective/objective) and orientation (positive/negative/neutral). For the sentence classification, author's present three different algorithms: (1) sentence similarity detection, (2) naïve Bayens classification and (3) multiple naïve Bayens classification. For opinion orientation authors use a technique similar to the one used by Turney [4] for document level. Wilson et al. [8] pointed out that not only a single sentence may contain multiple opinions, but they also have both subjective and factual clauses. It is useful to pinpoint such clauses. It is also important to identify the strength of opinions. Like the document-level opinion mining, the sentence-level opinion mining does not consider about object features that have been commented in a sentence. For this the feature level opinion mining is discuss in the next sub-section.

3.3 Task of Opinion mining at Feature level

The task of opinion mining at feature level is to extracting the features of the commented object and after that determine the opinion of the object i.e. positive or negative and then group the feature synonyms and produce the summary report. Liu [9] used supervised pattern learning method to extract the object features for identification of opinion orientation. To identify the orientation of opinion he used lexicon based approach. This approach basically uses opinion words and phrase in a sentence to determine the opinion. The working ofof lexicon based approach [10] is described in following steps.

1. Identification of opinion words
2. Role of Negation words
3. But-clauses

4. Applying different techniques to opinion mining

4.1 Finding Frequent Nouns and Noun Phrases

This method finds *explicit aspect expressions* that are nouns and noun phrases from a large number of reviews in a given domain. Nouns and noun phrases (or groups) identified using the frequency approach and is a discriminator. Web search was used to find the number of hits of I did a lot of research last year before I bought this camera... It kind a hurt to leave behind my beloved nikon 35mm SLR, but I was going to Italy, and I needed something smaller, and digital.The pictures coming out of this camera are amazing. The 'auto' feature takes great pictures most of the time. And with digital, you're not were identified by a part-of-speech (POS) tagger. Their occurrence frequencies are counted, and only the frequent ones are kept. A frequency threshold can be decided experimentally. The reason that this approach works is that when people comment on different aspects of an entity, the vocabulary that they use usually converges. Thus, those nouns that are frequently talked about are usually genuine and important aspects. Irrelevant contents in reviews are often diverse, i.e., they are quite different in different reviews. Hence, those infrequent nouns are likely to be non-aspects or less important aspects. Although this method is very simple, it is actually quite effective. Some commercial companies are using this method with several improvements.

4.2 Using Opinion and Target Relations

Since opinions have targets, they are obviously related. Their relationships can be exploited to extract aspects which are opinion targets because sentiment words are often known. This method was used in (Hu and Liu, 2004) for extracting infrequent aspects. The idea is as follows: The same sentiment word can be used to describe or modify different aspects. If a sentence does not have a frequent aspect but has some sentiment words, the nearest noun or noun phrase to each sentiment word is extracted. Since no parser was used the "nearest" function approximates the dependency relation between sentiment word and noun or noun phrase that it modifies, which usually works quite well. For example, in the following sentence, "The software is amazing."If we know that "amazing" is a sentiment word, then "software" is extracted as an aspect. This idea turns out to be quite useful in practice even when it is applied alone. The sentiment patterns method in uses a similar idea. Additionally, this relation-based method is also a useful method for discovering important or key aspects (or topics) in opinion documents

because an aspect or topic is unlikely to be important if nobody expresses any opinion or sentiment about it.

4.3 Using Supervised Learning

Opinion extraction can be seen as a special case of the general information extraction problem. Many algorithms based on supervised learning have been proposed in the past for information extraction The most dominant methods are based on *sequential learning* (or *sequential labeling*). Since these are supervised techniques, they need manually labeled data for training. That is, one needs to manually annotate aspects and non-aspects in a corpus. The current state-of-the-art sequential learning methods are *Hidden Markov Models* (HMM) and c*onditional Random Fields* (CRF). The application of lexicalized HMM model to learn patterns to extract aspects and opinion expressions. They trained CRF on review sentences from different domains for a more domain independent extraction. A set of domain independent features were also used, e.g. tokens, POS tags, syntactic dependency, word distance, and opinion sentences. Integrated two CRF variations, i.e., Skip- CRF and Tree-CRF, to extract aspects and also opinions. Unlike the original CRF, which can only use word sequences in learning, Skip-CRF and Tree- CRF enable CRF to exploit structure features. The rules are mined based on sequential pattern mining considering labels (or classes). One can also use other supervised methods. First candidate aspect and opinion word pairs using a dependency tree, and then employs a tree structured classification method to learn and to classify the candidate pairs as being an aspect and evaluation relation or not. Aspects are extracted from the highest scored pairs. The features used in learning include contextual clues, statistical co-occurrence clues, among others. It is using a partially supervised learning method called one-class SVM) to extract opinion. Using one-class SVM, one only needs to label some positive examples, which are aspects, but not non-aspects. They also clustered those synonym aspects and ranked aspects based on their frequency and their contributions to the overall review rating of reviews. The traditional supervised learning and semi-supervised learning for opinion extraction. used a supervised method.

5. Design:

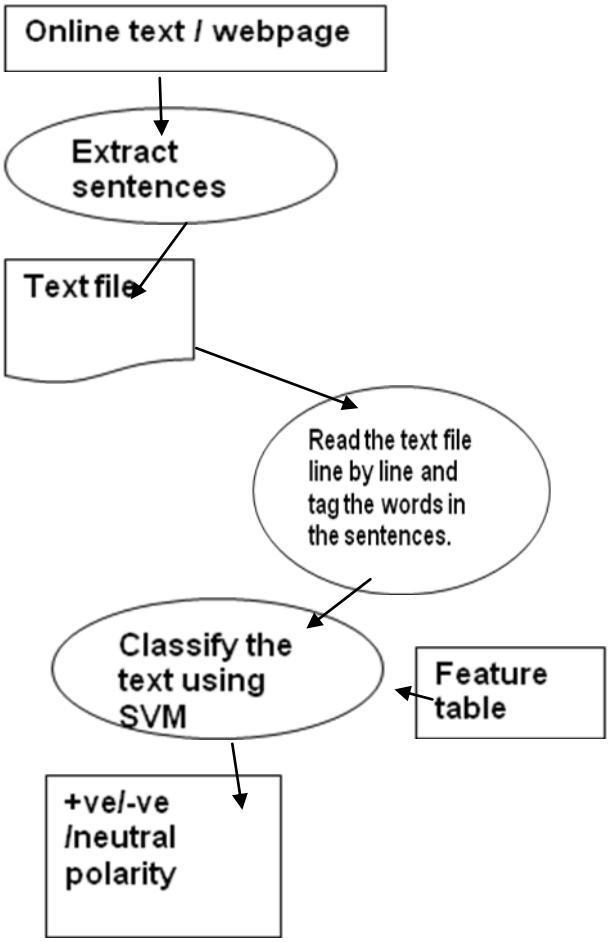
The online text is extracted from web site and it is put in a text file. The text is preprocessed by annotation of the sentences and SVM is applied for classification of text.

Algorithm

Input: HTML file, Text file storing features.
Outpt polarity i.e Positive/Negative/Neutral
Method:

1. Read the HTML file body part and extract sentences.
2. Number the extracted sentences.
3. Identify the feature in the sentences to extract opinion on those sentences
4. Words close proximity to the features are identified and compare with the word file and its corresponding score.
5. Apply SVM to categorize the text
6. **Conclusion:**

This paper presents the application of Opinion Mining Techniques for online Oriya Text. Opinion mining aims at recognizing, classifying and determining opinion orientations of the opinionated text. In this paper we first presented opinion mining techniques at various level, which determines whether a document or sentence carries a positive or negative opinion.



7. REFERENCES

[1] B. Liu. 2010. Sentiment Analysis: A Multifaceted Problem., Invited paper, IEEE Intelligent Systems.

[2] B. Liu. 2010 Sentiment Analysis and Subjectivity Second Edition, The Handbook of Natural Language Processing.

[3] B. Pang, L. Lee, and S. Vaithyanathan, 2002. Thumbs up? Sentiment classification using machine learning techniques," Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 79–86

[4] P.Turney 2002. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In: Proceeding of Association for Computational Linguistics, pp. 417--424.

[5] B. Liu 2008. Opinion Mining and Summarization, World Wide Web Conference, Beijing, China.

[6] E. Riloff, and J. Wiebe, 2003. Learning Extraction Patterns for Subjective Expressions, Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Japan, Sapporo.

[7] H. Yu, and V. Hatzivassiloglou, 2003. Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences, Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Japan, Sapporo.

[8] T.Wilson,., J. Wiebe,., R. Hwa,. 2004. Just how mad are you? Finding strong and weak opinion clauses. In: the Association for the Advancement of Artificial Intelligence, pp. 761--769.

[9] B. Liu, and J. Cheng, 2005. Opinion observer: Analyzing and comparing opinions on the web Proceedings

[10] X. Ding, B. Liu, and P. S. Yu, 2008. A holistic lexicon-based approach to opinion mining, Proceedings of the Conference on Web Search and Web Data Mining (WSDM).