Available online at www.sciencedirect.com

ScienceDirect

Materials Today: Proceedings XX (2016) XXX–XXX

materialstoday: PROCEEDINGS

www.materialstoday.com/proceedings

PMME 2016

A Hybrid Knowledge Mining Approach to Develop a System Framework for Odia Language Text Processing☆

Brojo Kishore Mishra, Rekhanjali Sahoo*

Associate Professor.C. V. Raman College of Engineering,Bhubaneswar,752054, Odisha, India.
Assistant Professor. Gurukula Institute of Technologt,Bhubaneswar,752056,Odisha, India.

Abstract

Words are evolved and use for expression of man's inner feelings. Any language-spoken or written by human beings, use many words in sentences of language to expresses his feelings and emotions through sentences. To supplement our own mother tongue, we borrow such words from other languages. Odisha, which is a state in the eastern part of India, has more than 33 million people speaking and writing this language. The culture and knowledge stored in many forms through Odia language text has a rich heritage. Odia is the mother language of the majority of the people of Odisha at present and also in the past. Various text forms such as reviews, news, and blogs are natural language processing task that mine information from opinion mining, and classify them on the basis of their polarity as positive, negative or neutral. The last few years, enormous increase has been seen in Odia language on the Web. This proposal gives an overview of the work that has been done in Odia language**.** The present work is a beginning to a higher goal of mining opinions or sentiments of people. The various phases of Natural Language Processing (NLP) included here are Lexical, Morphological and Syntactic-Semantic stages, to generate the Root word, Part of Speech, Suffix and Synonym of words in Odia text.

Selection and Peer-review under responsibility of International Conference on Processing of Materials, Minerals and Energy (July 29th – 30th) 2016, Ongole, Andhra Pradesh, India.

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000 .
E-mail address: author@institute.xxx

Selection and Peer-review under responsibility of International Conference on Processing of Materials, Minerals and Energy (July 29th – 30th) 2016, Ongole, Andhra Pradesh, India.

Keywords: Opinion Mining, Sentiment Analysis, Natural Language Processing, Odia Language, Classification.

1. Introduction

For Natural Language processing researchers, Opinion Mining is a recent area of interest. Opinions are subjective statements that reflect sentiments or perceptions about the entities and events. The use of machine learning techniques and data mining algorithms has taken a great role in the text classification and translation processes. The social media, blogs, forums, e-commerce web sites, etc. encourages citizens to share their opinion, emotions and feelings publically. It is an extension of data mining which utilizes natural language processing techniques to extract people's opinion from World Wide Web. The Opinion mining system analyze each text and see which part contain opinionated word, which is being opinionated and who has written the opinion. Sentiment analysis analyzes each opinionated word or phrase and determines its sentiment polarity orientation, whether it is positive or negative or neutral. It gives the summarized opinion of a writer or speaker. Sentient analysis can be done at word level, sentence level and document level. Opinion mining is a type of text mining which classify the text into several classes. Sentiment analysis which also known as Opinion mining use some algorithm techniques to categorize the user opinions into positive, negative and neutral classes .This categorization of text is called polarity of text. Opinion mining is to determine the attitude of a speaker or a writer with respect to the same topic or the overall contextual polarity (negativity or positivity) of a document. So in other words it can say that Opinion mining is to find out what other people think about a particular text, that text may be an odia text or any other language. The main objective of Sentiment analysis is classification of sentiment. It classifies the given text into three levels. Document level, sentence level, and entity / aspect level.

1.1. Data mining:

Data mining is the process of extracting knowledge or predicting previously unknown and useful trends from large quantities of data by using the knowledge of multidisciplinary fields such as statistics, mode identify, artificial intelligence, machine learning, database and so on. We have been collecting tremendous amounts of information. Initially, with the advent of computers and means for mass digital storage, we started collecting and storing all sorts of data, counting on the power of computers to help sort through this amalgam of information [6]. Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases [7].

1.2. Opinion mining:

Opinions are very important in the life of human beings. These Opinions helped the humans to carry out the decisions. Opinion Mining is a recent area of interest for Natural Language (NLP) researchers. Opinion mining (OM) is a recent discipline at the crossroads of information Extraction and computational linguistics which is concerned not with the topic a document is about, but with the opinion it expresses. We always get bothered during any decision making that what other people think. Searching for opinion poll is always a good alternative to take own decision. Generally people look forward among their acquainted circle for others views regarding any particular topic or matter [6]. 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 is of two types: facts and opinions [3]. Sentiment analysis in a multilingual world remains a challenging problem, because developing language-specific sentiment lexicons is an extremely resource intensive Process. Such lexicons remain a scarce resource for most languages. Opinion mining is a type of text mining which classify the text into several classes. Sentiment analysis which also known as Opinion mining use some algorithm techniques to categorize the user opinions into positive, negative and neutral classes .This categorization of text is called polarity of text. The main objective of Sentiment analysis is classification of sentiment. It classifies the given text into three levels.

1.3. Classification techniques:

Odia is a Classical language 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 are divided into different part of speech [3] tag. The part of speech tags can be considered as a unit which can be used for extracting opinion. To analyzing opinion mining, the online text is treated as a document. That document tells about a particular topic. That topic is mined at three levels. Those are discussed below:

- *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. Here we use a three step algorithm i.e.,
    - Adjectives are extracted along with a word that provides appropriate information.
    - The semantic orientation is captured by measuring the distance from words of known polarity.
    - The algorithm counts the average semantic orientation for all word pairs and classifies a review as recommended or not.

  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. Here we use the results using naive bayes, maximum entropy and support vector machine algorithms.
- *Task of opinion mining at Sentence level:*
  The sentence level opinion mining is associated with two tasks. 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 [3] one opinion. For example “The sound quality of this system is good.” However, it is not true in many cases. For the sentence classification, author’s present three different algorithms: (1) sentence similarity detection, (2) naïve Bayens classification and (3) multiple naïve Bayens classification. Here 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
- *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. Used supervised pattern learning method to extract the object features for identification of opinion orientation. Here we will use opinion mining task with a majority of the implementation [3] of Bayesian classifiers, neural networks, and SVMs (Support Vector Machines).

2. Odia language text mining:

Odia Language is the official language of Odisha, a state in the eastern part of India having more than 4.2 billion readers and writers. It has a rich heritage and culture and knowledge is stored in many forms through Odia language text. It used Natural Language Processing and Machine Learning ethics to determine opinion in the text. The evaluation of opinion can [3] be done in two ways:

- Direct opinion, gives positive or negative opinion about the object directly. For example, in Odia “Rama jane vala sasaka thile”(Ram was a good administrator) expresses a direct opinion.
- Comparison means to compare the object with some other similar objects. For example, in Odia “Ramanka sasana Ravananka sasan sange tulana jogya nuhe. (Ram’s administration can’t be comparable with Ravan’s administration.)

  The conversion of text in language study is not a new idea. Natural language-understanding systems convert samples of human language into more formal representations that are easier for computer programs to manipulate. And that is going to be fed to the further process of language processing in the field of Opinion Mining. Extraction of opinion expression from text, eventually including relations with the rest of content. It develops an in-depth understanding of both the algorithms available for the processing of

linguistic information and the [6] underlying computational properties of natural languages. Computational linguists dealing with syntax and semantics of languages have long dealt with the problem of making sense of the message conveyed in a narrative. The syntax, in general, is relatively easy to understand and interpret, but the semantics always posed a comparatively complex problem. The problem is compounded by the fact that word usage in any language is full of ambiguity, where the same word may have many senses depending on the context of the narrative.

2.1. Existing Techniques:

Based on the above initial experiments we believe that reordering of the target language phrases improve substantially by tapping the available resources for Odia [6]. In theory, natural-language processing is a very attractive method of human-computer interaction Natural-language understanding is sometimes referred to as an AI-complete problem, because natural-language recognition seems to require extensive knowledge about the outside world and the ability to manipulate it. The definition of understanding is one of the major problems in natural-language processing. These areas being the key focus of most research done in NLP and will continue to increase in complexity in the future analysis. Sentiment analysis is an important area of NLP with a large and growing literature. Excellent surveys of the field include, establishing that rich online resources have greatly expanded opportunities for opinion mining and sentiment. Sentiment analysis systems can be divided according to the scope of the input; therefore we have document-level (where the classification of opinions depend on the whole document), sentence-level, or phrasal-level which analyzes part of the sentence

2.2. Translation Support System:

Another way to model the construction of the dependency tree is using finite state machines or transition systems [6]. A transition system for dependency parsing is a quadruple S = (C; T; cs;Ct) where,

- C is a set of configurations, each of which contains a bu_er of (remaining) nodes and a set A of dependency arcs,
- T is a set of transitions, each of which is partial function T:C -> C
- cs is an initialization function, mapping a sentence x = w_0;w_1; : : :w_n to a configuration
- Ct _ C is a set of terminal configurations. The translation system from a foreign to a regional language consists of many problems. Any natural language is a free language, i.e. its structure is not fixed. Especially, for a language like English which has syntactic parsers of high quality, it is always desirable to tap these existing resources. The structure can keep changing as the user wishes. Hence a good translation system will have to handle as many grammar constructs as possible.

2.3. Detecting Emotion in text:

Subjective language is language used to express opinions, emotions. Both types of language are useful in text analysis: Subjective language is useful for automatic [3] subjectivity analysis and objective language is useful for information extraction. Emotions are expressed in subjective language so it would appear that subjectivity analysis is the only area beneficial in emotion detection. The process in which humans manually label a text is called annotation.

2.4. 2.4. System Architecture Design:

The diagrammatic representation of the system architecture is presented below.

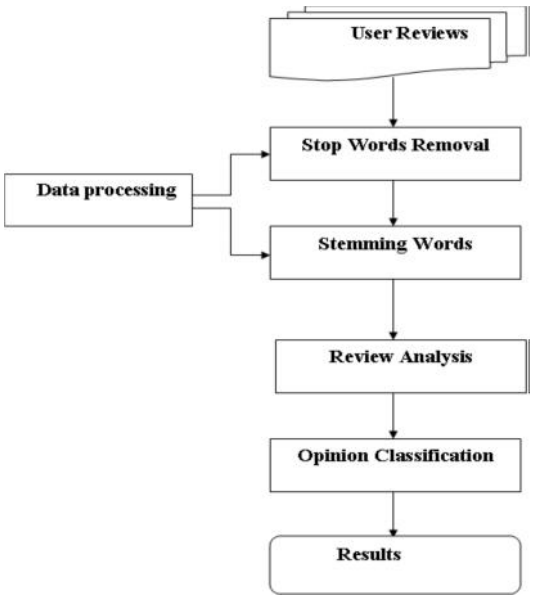

Fig 2. System Architecture

The collected reviews are pre-processed and the words are extracted. The stop words and stemming words are [3] removed at the stage of pre-processing. The extracted words are analyzed by reviewer and classify the sentence into positive and negative the product.

3. Road Map:

The roadmap of our proposed work will be as follows

Odiya text document → Information extraction module → Translation module → Preprocessing module → Classifier → Decision assistance module → Result.

4. Methodology and contributory work for Odia language

The different phases of NLP that can be used for opinion mining in Odia are:

- Lexical Analysis (Tokenization)
- Morphological Analysis
  i. Stemming
  ii. Tagging (POS)
- Syntactic-Semantic Analysis
- Discourse Integration- Anaphora Resolution
- Pragmatic Analysis

The block diagram of these phases is given below.

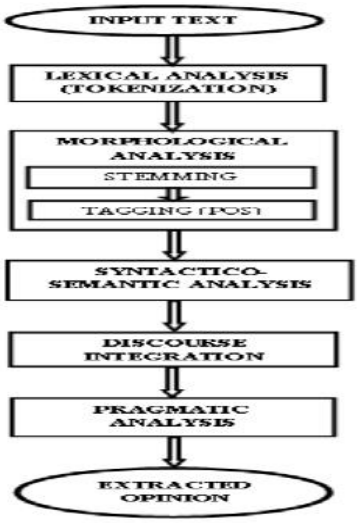

Fig.4 Different phases of Odia Language Understanding

5. Conclusion:

Opinion Mining which enhances traditional Natural Language Processing techniques by exploiting valuable information extracted from news texts. In this proposal, we proposed a number of techniques for mining opinion features from subjectivity detection & polarity detection product reviews based on data mining and natural language processing methods. This will further improve the feature extraction and the subsequent summarization. The sentiment analysis for sentence level is performed by naïve Bayesian classifier and aspect level opinion mining is for support vector machine. The user review is analyzed and rank for a particular product. The reviews are pre-processed to eliminate noise like stop words and stemming words are removed. The extracted words are classified into positive and negative in unigram using machine learning naive Bayesian classifier. Opinion mining aims at recognizing, classifying and determining opinion orientations of the opinionated text. In this proposal, we first presented opinion mining techniques at various levels, which determine whether a document or sentence carries a positive or negative opinion. In this work, we have analyzed Odia language text up to the Syntactic-Semantic phase, using the Paninian framework, which was found to be useful for Indian language processing, and not seen in literature to have been applied to Odia language processing.

6. References:

[1.] Sohag Sunder Nanda, Soumya Mishra, Sanghamitra Mohanty, “Oriya Language Text Mining Using C5.0 Algorithm”, International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 2 (1) , 2011, 551-554’

[2.] A.Shameem Fathima, D.Manimegalai, and Nisar Hundewale, “A Review of Data Mining Classification Techniques Applied for Diagnosis and Prognosis of the Arbovirus-Dengue”, IJCSI.

[3.] Poobana S, Sashi Rekha k , “Opinion Mining From Text Reviews Using Machine Learning Algorithm”, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 3, Issue 3, March 2015, pp 1567-1570.

[4.] Debnath Bhattacharyya, Susmita Biswas, Tai-hoon Kim, “A Review on Natural Language Processing in Opinion Mining”, International Journal of Smart Home, Vol.4, No.2, April, 2010, pp 31-38

[5.] Gayatri Dey, Hima Bindu Maringanti, “Paninian Framework for Odia Language Processing”, http://csidl.org/bitstream/handle/123456789/665/1.PDF?sequence=1

[6.] Jena Manoj Kumar , Balabantaray Rakesh Ch, “Opinion Mining for online Oriya Text”, European Journal of Academic Essays, Special Issue (1): 44-48, 2014.

[7.] Anita Kumari Nanda, Brojo Kishore Mishra, "Application of Fuzzy Data Mining in E-Government", 1st International Conference on Computing, Communication and Sensor Network, CCSN-2012, pp 188-193, 2012.