**IJCTA,** 9(22), 2016, pp. 249-256

# Information Retrieval in Web for an Indian Language: An Odia Language Sentimental Analysis Context

**Sanjib Kumar Sahu* Priyanka Behera** D.P.Mohapatra*** Rakesh Chandra Balabantaray******

***Abstract :*** Today internet has reached to every people hand, so the user generated data has increased tremendously. Now a day's people post everything comes to their mind on social networking sites, blogs, discussion forums and review sites. People are openly giving views on any universal topics, products, politics, policies and many more. People are showing likeness or hatred for anything by writing their perception. So with add to in surfer generated data or information can provide one of the key aspect to ongoing researches, and industrial development and to our government in view to extract the valuable information to decide things like marketing campaign strategies, product preferences in market, political parties agendas, company strategies, travel & tourism and social events.. This generated content by users has become one of the key sources to organizations belonging to different fields for knowing or learning or to seek any general intent or sentiment of public and this can be done by Sentiment Analysis. Sentiment Analysis has been a major research platform from last few decades. Basically, Sentiment Analysis is a process of analyzing the thinking, emotions, feeling and attitude of the person from his piece of text. Sentiment analysis is known for its one of the key features i.e in clustering the polarity of text as whether it shows positivity, negativity or neutrality.

So far whatever the work is done in this field was always for English language. Country like India where more than100 million people are using internet and where more than 1600 languages are spoken but sentiment analysis is applied only for languages like Hindi, Bengali, Telugu, Malyalam, and Punjabi . This paper attempts to describe application of sentiment analysis on Odia language. The system classifies the polarity of Odia language in positive and negative sentiments. The experimental result shows that the system on multi-sized datasets yielding > 90% accuracy in general.

***Keywords :*** Information Retrieval, Sentiment Analysis, Indian Languages, Multi-Sized dataset.

## 1. INTRODUCTION

In today's world of globalization, the most developing countries like India where more than 100 million internet users are present, the analysis of opinion of the people is essential. Since internet is no longer monolingual anymore and with utf-8 standard for Indian language introduced, it is strongly accepted and used by the Indian people. So content in Indian languages like Hindi, Marathi, Bengali, Punjabi has increased has grown rapidly. All major news papers, Government officials, political parties campaign, *e*-commerce Company have set up their websites in local Indian languages to increase its reach to the people. Micro-blogging websites also features of many Indian languages to chat. Bloggers are writing in their own languages to reach their own native people. The rise of user-generated content of Indian languages on web world led to analyze the sentiment of Indian people.

* Dept. of Computer Science and Application, Utkal University, Bhubaneshwar, Odisha, India

** USICT, Guru Gobind Singh Indraprastha University, Delhi, India

*** Dept. of Computer Science and Engineering, NIT, Rourkela, Odisha, India

**** Dept. of Computer Science and Engineering, IIIT, Bhubaneswar, Odisha, India

As far as development in sentiment analysis with respect to Indian languages is concerned, most of the work done in this domain is for English language and European languages. Work has also been done based on sentiment analysis in few languages of India like Hindi, Bengali, Tamil, Malyalam, Panjabi and Gujrati.

This paper is presents the sentiment analysis of, one of the Indian language which is Odia or Oriya language. Sentiment Analysis using supervised learning was never applied before for the language Odia as this was confirmed through the literature survey and attempts to show the sentiments behind the Odia documents.

Odia (ଓଡ଼ିଆ) is also known as Indo-Aryan language which is generally spoken by more than 40 million people from the state of Odisha and its adjoining neighbor state. Odisha state is the eastern part of India and that's where Odia language is the 10th largest language stated by the constitution part-VIII of India [3].

Odia language is spoken by 80% population of Odisha[1], and odia is also spoken in few parts of West Bengal, Jharkhand, Chhattisgarh and Andhra Pradesh[2]. Odia has been a language known for its classical language touch in India on basis of having long literary history [4, 5]. The other languages close to it are Bengali and Assamese. There closeness led us to work towards Odia language and working for this language.

Sentiment Analysis is a phenomenon of extracting sentiment or opinions expressed by user over particular product, policies, person or movie. It classifies the sentiments of the given data into positive and negative polarity.

Our paper presents a model based on supervised sentiment analysis which uses Naive Bayes algorithm as its base. This algorithm is simple in nature which classifies test and to does that it needs a small amount of trained data and then finds parameters important for such classification and this is done in very small amount of time as to train as compared to other available models. Its going to present that the extent of correctness is achieved using Naive Bayes algorithm. We use here feature extractor algorithm for sentiment analysis. Feature extraction is most important task in sentiment analysis. In extraction feature, space is the base where it is converted into reduced new space without deleting any features and by replacing actual features with a less representative set. In other words, the purpose of feature extraction is to derive new feature set that are combination of original feature set and are unrelated [6]. Feature extraction is used to score the feature and these score is use to determine positive negative polarity.

A model for doing the sentiment analysis over Odia language is the main feature of this paper. Other part of this paper contains related work or literature review is discussed in section II. Methodology and system architecture are presented in section III. Experiment results are presented in section IV. Section V presents the experimental analysis. Conclusion and future work are presented in VI section.

## 2. RELATED WORK

**This part of the paper discusses all the research done in past of sentiment analysis for Indian languages. As to make it simpler and easy to interpret we have divided the related work in two sections :**

### (A) Work done in field of sentiment analysis of Indian languages

As far as Indian languages are concerned, most of the work done in Hindi language and few work done in languages like Bengali, Marathi, Telgu, Malyalam, Punjabi and Gujrati.

In [7] authors proposed "Hindi subjective lexicon" in which they generated subjective lexicon using graph traversal based method. Their proposed algorithm achieved 79% of accuracy.

In [8] author proposed "Hindi Sentiment orientation system" is an unsupervised learning approach which use dictionary to determine the polarity of reviews. It also handles negation of words in Hindi language. Their proposed system achieved accuracy up to 65%.

In [9] proposed "Hindi senti word net" to identify sentiment associated with Hindi words. Language specific challenges including negation and discourse are handled to improve the accuracy of system which achieved around 80%.

In [10, 11] performed a task of emotion tagging of Bengali words. They classified words in six emotion classes that are anger, happy, surprise, disgust and fear along with three type of intensities high, general and low for sentence level annotation.

"Sentiment analysis of political reviews in Punjabi language" proposed an approach to determine sentiment orientation i.e. polarity by scoring method [15].

In this paper [16] sentiment analysis for Malayalam movie review is carried out using machine learning technique CRF combined with a rule based approach and the system achieved 82% of accuracy.

### (B) Work done in field of sentiment analysis of Odia languages

In field of sentiment analysis for Odia language not much work is done but still lot of work is done on processes like wordnet development, tokenization, part of speech tagging, stemming, morphological analysis, stop word removal, syntactic-semantic analysis, discourse integration, pragmatic analysis, opinion mining.

The paper [17] describes the OriNet in which odia words with meanings, synonym, antonym, usage and part-of-speech is mention. They have used java programming to represent the system.

The paper [18] proposed and describes the stemmer for odia language. The applied affix, prefixes, postfix and infix stripping algorithm to find the stem or root word.

"Developing odia Morphological analyzer using Lt-toolbox" discusses the research done in view to develop a Morphological analyzer. Here the paradigm of words is created by the use of XML based morphological dictionary from Lt-toolbox packages [19].

The paper [20, 21, 22] describe the Paninian framework for odia language. It applied to identify structurally correct odia sentence. It describe the relation between surface form (vibhakti) and semantic (karaka) roles.

## 3. METHODOLOGY

The need of implementation comes from the intent of having the ability to classify the Odia text as in terms of positivity or negativity of sentiments. Any model based on Naive Bayes was made capable of creating a simplified training corpus. The corpus is made easy to read and to work on as it was based on positive and negative data. The output of this is a basic binary (or boolean) classifier capable of identifying +*ve* and –*ve* data was trained. This classifier is well suited in specifying the positive and negative polarities, which leads to a precise output near to 90% in the corpus with these two categories. The process showing in figure is described below in detail.

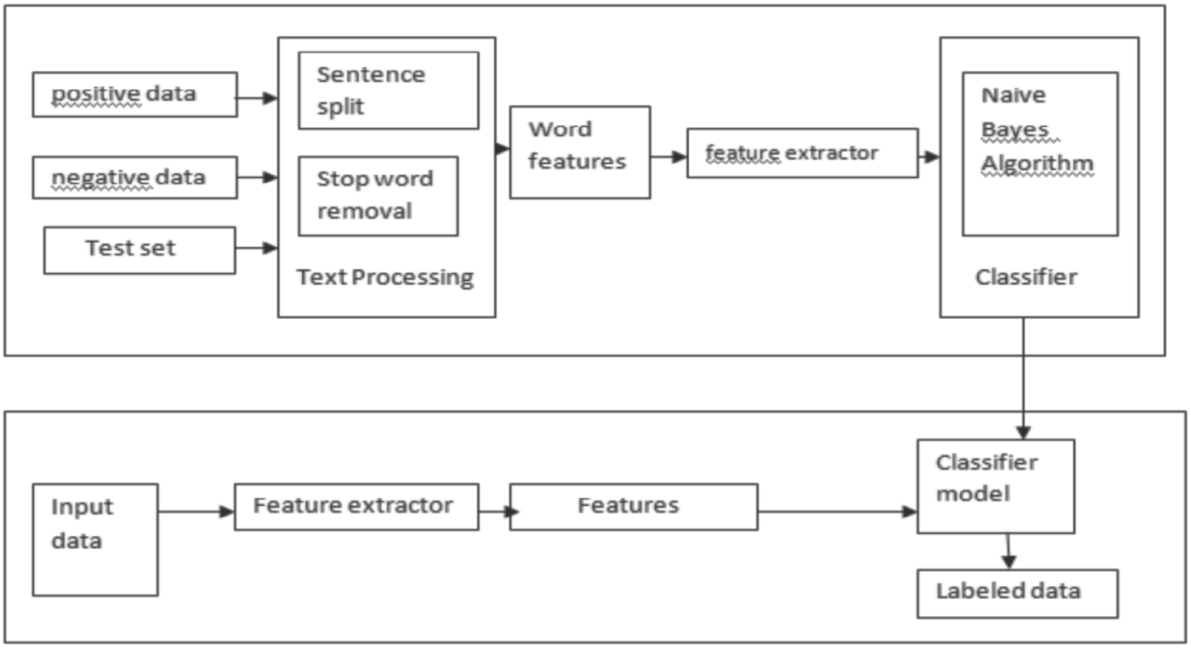


**Fig. 1. Architecture of the proposed system.**

Odia is a scarce resource language, for performing sentiment analysis for Odia language we don't have sufficient annotated datasets and other resources and even not much work done in field of sentiment analysis for Odia language. We built a dataset of 1000 sentences with equal numbers of positive and negative sentences. This dataset are built by taking few Odia movie reviews and around some Hindi movie reviews and translated in Odia language using Google translation. These reviews are manually annotated and corrected to select top 500 positive and 500 negative sentences. We used 500 datasets for training and other 500 datasets for testing. Both the sets hold equal numbers of positive and negative sentences. We prefer movie reviews as our datasets because it contains ample of opinionated words and it captures the majority of adjectives relevant for classification.

In order to achieve higher accuracy text preprocessing is applied to the input datasets. First tokenization is applied to text so that text is split into words or tokens and then noise like special character, punctuation mark, symbols, URLS and stop words are removed as not all words in the sentence are useful for sentiment analysis. So removing these words improve accuracy.

Now the feature extractor extracts the list of word feature. The word feature is the list of every distinct word order by frequency appearance. Feature extractor decide which feature is relevant and discard the irrelevant features and to do that it uses dictionary indicating what words are contained in the input passed. We use here the word feature list and input to create the dictionary. With the help of apply features method we can apply the features to the classifier. The variable training set contains labeled feature sets. It contains the list of feature word and the sentiment string for that feature word.

With this training set we train the Naive Bayes classifier. Naive Bayes classifier is supervised classification method based on training corpora containing correct label for each input. The classifier is based on Baye's theorem which provides us the freedom to assume features. This assumption is made to make the computation involved in classification simpler and because of this assumption; it is considered as "naive". This assumption has no impact on accuracy in classification of sentiment by much; rather it makes the classification algorithms fast to applied for large datasets [12]. Naïve Bayes classifier says that the effect of value of a feature($x$) on given class ($c$) is independent of the values of other features.

**This assumption called conditional independence.**

$$P(C \mid X) = \frac{P(X|C)*P(C)}{P(X)} \tag{1}$$

Where,

$P(C \mid X)$ : stands for posterior probability for gives class of features

$P(C)$ : gives the probability of class

$P(X \mid C)$ : states likelihood *i.e* is probability of predictor given class

$P(X)$ : prior probability of feature

In other form, $P(C \mid X) = p(x1|c)* p(x2|c)*p(x3|c)........*p(xn|c)$ (2)

To explain the concept more precisely, lets us explain by taking a word "ଭଲ (or bhala)" which occurs in 20 % of positive datasets and 2 % of negative datasets, then the likelihood score for the positive datasets is multiplied by 0.16, the likelihood score for the negative datasets is multiplied by 0.02.

The likelihood of word "ଭଲ"(or bhala or fine) falling into either of classes is equal since we have two classes. So likelihood is 0.5.

- Posterior probability of word "ଭଲ"(or bhala or fine) is being positive = Prior probability of being positive × Likelihood of being positive

  = 0.20 × 0.5 = 0.10 = 10 % chances that word "ଭଲ" being positive
- Posterior probability of word "ଭଲ"(or bhala or fine) is being negative = Prior probability of being negative × Likelihood of being negative.

  = 0.02 × 0.5 = 0.01 = 1 % chances that word "ଭଲ"(or bhala or fine) being negative.

Therefore, the naive Bayesian classifier predicts the word "ଭଲ"(or bhala or fine) in positive class.

Test set is used by the evaluation metric technique to generate score for the system by comparing the labels that it generate for the inputs in test set with correct labels for those inputs.

So when the input is passed to the system the feature extractor extract the features and apply it to trained naive bayes classifier model which then classify the input into the correct label.

## 4. EXPERIMENTAL RESULTS

This analysis is conducted on movie review. Reviews were applied as input to the system which classifies these reviews and determine the polarity of these reviews and present the summarized positive and negative results which prove to be helpful for the users. Input reviews were also classified by us to determine how well the system classified the reviews as compared to human judgments. Three evaluation measures are used, on the basis of which system performance is computed, these are:

• Accuracy • Precision • Recall

(*a*) **Accuracy :** The accuracy is given by percentage of test set that are correctly classified by classifier. That is,

$$\text{Accuracy} = \frac{\text{TP + TN}}{\text{TP + TN + FP + FN}} \tag{3}$$

(*b*) **Precision :** Gives measure of uniqueness *i.e.*, what percentage of elements that are labeled as positive are actually positive.

$$\text{Precision} = \frac{\text{TP}}{\text{TP + FP}} \tag{4}$$

(*c*) **Recall :** Gives the amount of wholeness *i.e.*, what percentage of positive the elements that are labeled as positive.

$$\text{Recall} = \frac{\text{TP}}{\text{TP + FN}} \tag{5}$$

where,

**TP :** These refer to number of **true positive** elements correctly predicted by the classifier.

**TN :** There are the numbers of **true negative** elements correctly predicted by the classifier.

**FP :** These are the number **false positive** elements incorrectly tag by classifier.

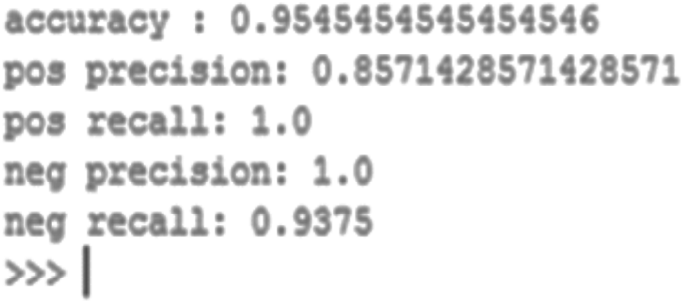
**FN :** These are the number of **false negative** elements incorrectly tag as negative by classifier.

On the basis of these measures of evaluation, our experiment results show that 'Sentiment Analysis for Odia language' is performed well in the movie review domain.

Experiment result performed using 1000 sentences of movie review and the result is shown in table 1 below. Figure 2 shows the result of sentiment analysis using 500 datasets

**Table 1. Experimental Result.**

| ***Measures*** | ***Result*** |
|---|---|
| Accuracy | 0.918781726 |
| Precision (pos) | 0.857142857 |
| Recall (pos) | 1.0 |
| Precision (neg) | 1.0 |
| Recall (neg) | 0.9375 |

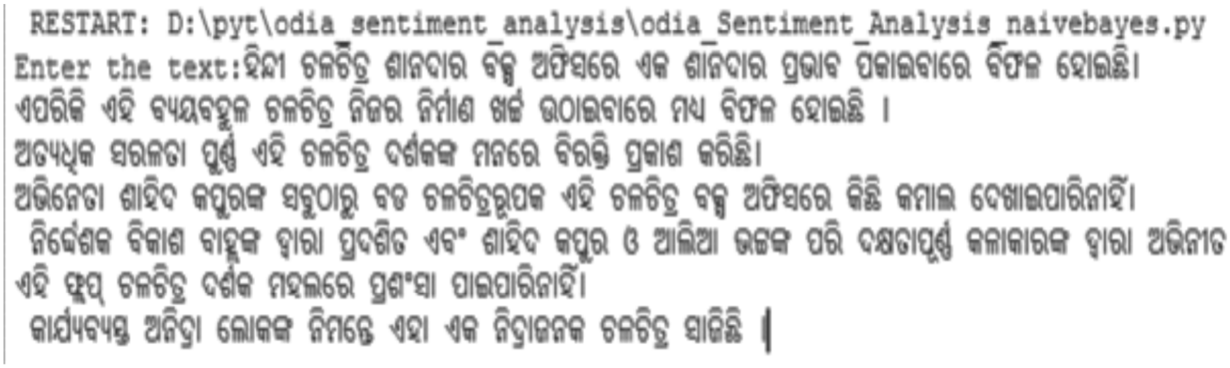


```
accuracy : 0.9545454545454546
pos precision: 0.8571428571428571
pos recall: 1.0
neg precision: 1.0
neg recall: 0.9375
>>> |
```

**Fig. 2. Snapshot of result of sentiment analysis using 500 datasets.**

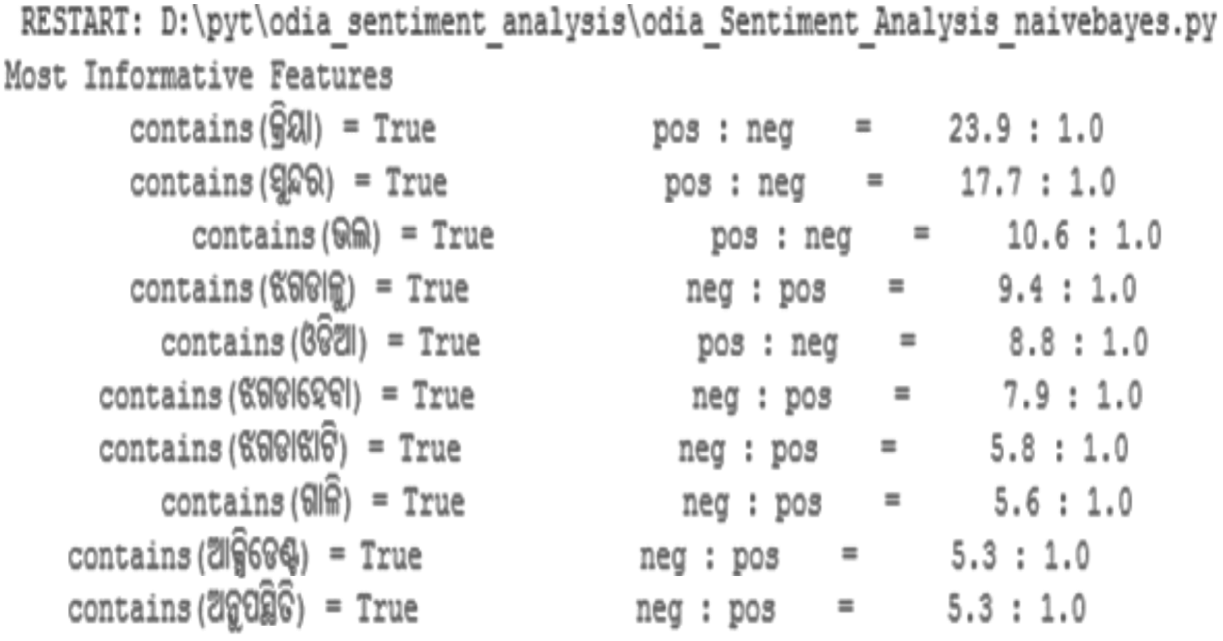
## 5. EXPERIMENTALANALYSIS

The Naive Bayes classifier after training with the trained datasets can show the most informative features using show_most_informative_feature ( ) function .We ask it, show the 10 features which are classified as positive or negative. The ratio (neg:pos) OR (pos:neg), implies that the particular word has been used more often as a positive or negative. For instance, the word “ସୁନ୍‌ଦର” (or beautiful) has been used 17.7 times as a positive sentiment than a negative sentiment in the text.  The following figure 3 and figure 4 will give most informative features.



```
 RESTART: D:\pyt\odia_sentiment_analysis\odia_Sentiment_Analysis_naivebayes.py
Enter the text:ହିନ୍ଦୀ ଚଳଚିତ୍ର ଶାନଦାର ବକ୍ସ ଅଫିସରେ ଏକ ଶାନଦାର ପ୍ରଭାବ ପକାଇବାରେ ବିଫଳ ହୋଇଛି।
ଏପରିକି ଏହି ବ୍ୟୟବହୁଳ ଚଳଚିତ୍ର ନିଜର ନିର୍ମାଣ ଖର୍ଚ୍ଚ ଉଠାଇବାରେ ମଧ୍ୟ ବିଫଳ ହୋଇଛି ।
ଅତ୍ୟଧିକ ସରଳତା ପୂର୍ଣ୍ଣ ଏହି ଚଳଚିତ୍ର ଦର୍ଶକଙ୍କ ମନରେ ବିରକ୍ତି ପ୍ରକାଶ କରିଛି।
ଅଭିନେତା ଶାହିଦ କପୁରଙ୍କ ସବୁଠାରୁ ବଡ ଚଳଚିତ୍ରରୂପକ ଏହି ଚଳଚିତ୍ର ବକ୍ସ ଅଫିସରେ କିଛି କମାଲ ଦେଖାଇପାରିନାହିଁ।
 ନିର୍ଦ୍ଦେଶକ ବିକାଶ ବାହ୍ଲଙ୍କ ଦ୍ୱାରା ପ୍ରଦର୍ଶିତ ଏବଂ ଶାହିଦ କପୁର ଓ ଆଲିଆ ଭଟ୍ଟଙ୍କ ପରି ଦକ୍ଷତାପୂର୍ଣ୍ଣ କଳାକାରଙ୍କ ଦ୍ୱାରା ଅଭିନୀତ
ଏହି ଫ୍ଲପ୍ ଚଳଚିତ୍ର ଦର୍ଶକ ମହଲରେ ପ୍ରଶଂସା ପାଇପାରିନାହିଁ।
 କାର୍ଯ୍ୟବ୍ୟସ୍ତ ଅନିଦ୍ରା ଲୋକଙ୍କ ନିମନ୍ତେ ଏହା ଏକ ନିଦ୍ରାଜନକ ଚଳଚିତ୍ର ସାଜିଛି |
```

**Fig. 3. Snapshot of Input Parameter.**



```
 RESTART: D:\pyt\odia_sentiment_analysis\odia_Sentiment_Analysis_naivebayes.py
Most Informative Features
        contains(କ୍ରିୟା) = True              pos : neg    =     23.9 : 1.0
        contains(ସୁନ୍ଦର) = True              pos : neg    =     17.7 : 1.0
            contains(ଭଲ) = True              pos : neg    =     10.6 : 1.0
        contains(ଝଗଡାଳୁ) = True              neg : pos    =      9.4 : 1.0
          contains(ଓଡିଆ) = True              pos : neg    =      8.8 : 1.0
      contains(ଝଗଡାହେବା) = True              neg : pos    =      7.9 : 1.0
      contains(ଝଗଡାଝାଟି) = True              neg : pos    =      5.8 : 1.0
          contains(ଗାଳି) = True              neg : pos    =      5.6 : 1.0
    contains(ଆକ୍ସିଡେଣ୍ଟ) = True              neg : pos    =      5.3 : 1.0
    contains(ଅନୁପସ୍ଥିତି) = True              neg : pos    =      5.3 : 1.0
```

**Fig 4. Snapshot of 10 most informative features.**

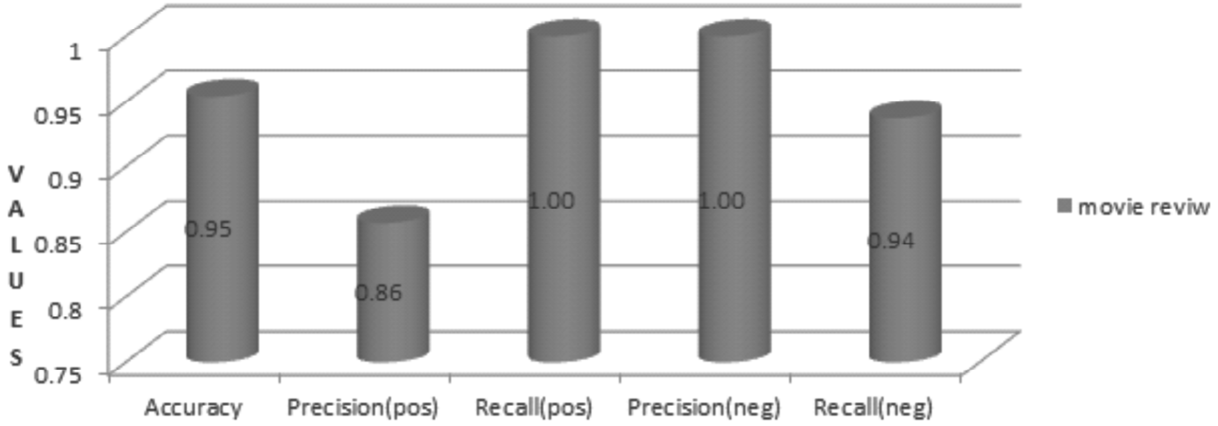


**Fig. 5. Performance of Odia sentiment analysis approach.**

The above figure 5 shows that the system 'Feature extraction for Sentiment Analysis in Odia Language using Naive Bayes classifier' performs well with respect to the movie review domain. Sentiment Analysis for Odia language shows the accuracy of $\geq 90\%$ can be achieve which proves that the system is more efficient.

## 6. CONCLUSION AND FUTURE WORK

In this paper we proposed an approach that gives polarity of the Odia language. Sentiment Analysis is required to be applied for Odia language to understand the sentiment of Odisha's people .The separate positive and negative summarized results are generated, which is helpful for the user in decision making. The experiment results indicate that the approach we used, is performing well in this domain and achieved the accuracy $\geq 90\%$ .

Substantial amount of work is left to be carried out, here we provide beam of light in direction of possible future avenues of research.

- Adding different feature selection mechanisms.
- Doing deeper analysis of sentence as a whole.
- Trying different classifier other than the Naive Bayes Classifier.
- Doing Sentiment Analysis on more type of document.
- Using different approaches of Sentiment Analysis.

## 7. REFERENCES

1. "Odia become sixth classical language "-*The Telegraph,* "Odia get classical language status"- *The Hindu.*
2. *Classical Odia (PDF). Bhubaneswar: Government of Odisha*, history of odia language .p.5 2001.
3. Patnaik, Durga (1989). Palm Leaf Etchings of Orissa. New Delhi: Abhinav Publications. p. 11 2002.
4. Panda, Shishir (1991). Medieval Orissa: A Socio-economic Study. New Delhi: Mittal Publications. p. 106 1999.
5. Patnaik, Nihar (1997). Economic History of Orissa. New Delhi: Indus Publishing. p. 149 1999.
6. B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, pages 7986. Association for Computational Linguistics, 2002.
7. "Hindi Subjective Lexicon: A Lexical Resource for Hindi Polarity Classification " Akshat Bakliwal, Piyush Arora, Vasudeva Varma, In International conference on Language resources and Evaluation (LREC) May 2012.
8. Richa Sharma, Shweta Nigam and Rekha Jain, " Polarity Detection Of Movie Reviews In Hindi Language" International Journal on Computational Sciences & Applications (IJCSA) , August 2014.
9. Namita Mittal, Basant Agarwal, Garvit Chouhan, Nitin Bania, Prateek Pareek "Sentiment Analysis of Hindi Review based on Negation and Discourse relation", International Joint Conference on Natural Language Processing,Nagoya, Japan, October 2013.

10. A. Das , S. Bandyopadhyay “SentiWordNet for Bangla.” In knowledge Sharing Event-4 Task-2: Building Electronic Dictionary, February 23th to 24th. 2010 Mysore.

11. A. Das and S. Bandyopadhyay (2010) “Phrase-level Polarity Identification for Bengali “ In International Journal of Computational Linguistics and Application (IJCLA) Vol. 1 No. 1-2 . 2010. ISSN 0976-0962.

12. B. Pang and L. Lee. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2):1135, 2008. ISSN 1554-0669.

13. A. Kaur, Neelam Duhan, V. 2013. “A survey on sentiment analysis and opinion mining “.In International Journal of Innovative & Advancement in Computer Science (IJIACS) ISSN 2347-8616 Vol.4 May 2015

14. A. Joshi, B. A. R, and P. Bhattacharyya. A fall-back strategy for sentiment analysis in hindi: a case study, 2010.

15. Parul Arora, Brahmaleen Kaur,” Sentiment Analysis of Political Reviews in Punjabi Language “ . International Journal of Computer Applications (0975 – 8887) Volume 126 – No.14, September 2015

16. P Jayan, Deepu S Nair, Sherly Elizabeth Jisha. “A subjective feature extraction for sentiment analysis in Malayalam language.” In International Journal of Engineering Sciences Vol 14 2015

17. S. Mohanty and P.K. Santi, “Object Oriented design Approach to OriNet System: Online Lexical Database for Oriya Language”, IEEE Proceedings for LEC-2002, University of Hyderabad, Hyderabad India, 2002.

18. Dhabal prasad sethi, “Design of Lightweight Stemmer for Odia Derivational Suffixes”, International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 12, December 2013

19. Jena, S. Choudhery, H. Choudhry, D. M. Sharma, “Developing Oriya Morphological Analyzer Using Lt-toolbox”, ICISIL, Springer,Communications in Computer and Information Science 139, 2011, Page 124-129

20. R. Mohapatra, Lipi Hembram. “Morph-Synthesizer for Oriya Language A computational Approach”, LANGUAGE IN INDIA, Strenght for Today and Bright Hope for Tomorrow, 2010, Volume 10:9 .

21. A. Bharati, M. Bhatia, V. Chaitanya, R. Sangal, “Paninian Grammar Framework Applied to English”, Department of Computer Science and Engineering, Indian Institute of Technology, Kanpur, February 1996.

22. A. Bharati, R. Sangal. “Parsing Free Word Order Languages in the Paninian Framework”, ACL ’93 Proceedings of the 31st annual meeting on Association for Computational Linguistics, 1993, Pages 105-111

23. Masoumeh Zareapoor, Seeja K. R, “Feature Extraction or Feature Selection for Text Classification: A Case Study on Phishing Email Detection “, I.J. Information Engineering and Electronic Business, 2015, 2, 60-65.