INTERNATIONAL JOURNAL OF

ELECTRONICS ENGINEERING

ISSN : 0973-7383

Volume 11 • Number 1 • 2019

Challenges of Information Retrieval in Indian Language Big Data: ODIA Movie Review Context

Sanjib Kumar Sahu[a], D.P. Mahapatra[b] and R.C. Balabantaray[c]

[a]*Dept. of Computer Science & Application, Utkal University, Bhubaneswar, Odisha, India. Email: sahusanjib1977@gmail.com*
[b]*Dept. of Computer Science & Engineering, NIT, Rourkela, Odisha, India*
[c]*Dept. of Computer Science & Engineering, IIIT, Bhubaneswar, Odisha, India*

Abstract: This is the time of information technology. Today the most important thing is how to get right information at right time. Many data repositories are now available online. Information retrieval systems or search engines are used to access electronic information available on the internet. Different search engines are using different techniques for efficient retrieval of information, as per user needs. The accessibility of these databases is very low because the efficient search engines/retrieval systems supporting these languages are very low. Currently the web is dominated by English speakers, now many of the existing WebPages are in English. Retrieving information in Odia language is clearly important for many users, as many people are speaking this language in Odisha state of India. In this retrieval, the challenges occur because the queries are in the same language as the collection being accessed. To remove the gap between Natural Language Processing and Information Retrieval many tools have been developed. Today scenario many people in Odisha are watching movie both in Hindi and in Local Language (Odia). Rarely any websites are available in Odia language for display the opinion of user about movie review, in spite of so many people are speaking this language and do not understand the English language. For classification of review of different movie depend on syntax based attributes or lexical attribute or on both. The main purpose of this analysis of the opinion is to obtain the viewer's feelings about movie whether positive or negative. In this paper the need for analysis of sentiment is explained for Odia Language.

Keywords: Odia Information Retrieval, Indian Language Information Search, Cross Lingual Information Retrieval, Product Review, Movie Review.

1. INTRODUCTION

Oriya language is the official language of Odisha (Officially name of the state is changed from Orissa to Odisha in November 2011). It is a state in that 42 million people are speaking this language as per 2011 census. The region is also known as "Utkal" and is described in India's national anthem. It has a rich heritage, culture and knowledge is stored in many document form written in Odia (Oriya) text. According to Salton's classic textbook definition, Information retrieval is a divided or classified with the structure, analysis, storage, organization,

searching and retrieval of information. Now a day more and more data is available on internet. The internet is basically dominated by English language since its inception. In the time of Natural Language Processing (NLP) and now web is shifting towards multi-lingual. Indian languages are not well equiped in conceptualizing the technology. The good thing is that 22 Indian languages are adopted by NLP, but still others remain untouched. Though in these Indian languages, the understanding or acceptability of NLP is not uniform. Since India is a country having a wide range of regional languages, in the Indian context, the IR approach should be such that it can handle multilingual document collections. However, from a Natural Language Processing point of view, the language is resource poor. NLP in Odia language is far behind in information retrieval and other related applications. As the state government is interested for digitize all of its old documents to Odia (Oriya) language. As per information of CDAC website, many Odia Natural Language Processing tools are developed by CDAC and released by the Ministry of Information and Communication Technology, Govt. of India like true type font, keyboard driver, spell checker, Microsoft Word tool. The present work focuses on past, present and future of an efficient information retrieval system for Odia language.

1.1. About Odia (ଓଡ଼ିଆ) Language

Odia is an Indian language spoken by more than 33 million people mainly in the Indian state of Odisha, and also in West Bengal, Jharkhand, and Gujarat. Odia is closely related to Bengali and Assamese, and used to be known as Oriya, and Odisha used to be known as Orissa in English.

As per details from Census 2011, Orissa has population of 4.2 Crores, an increase from figure of 3.68 Crore in 2001 census. Total population of Orissa as per 2011 census is 41,974,218 of which male and female are 21,212,136 and 20,762,082 respectively. In 2001, total population was 36,804,660 in which males were 18,660,570 while females were 18,144,090[1].The literacy rate of male and female is 82.40 and 64.36 respectively. Figure 1 describes the district wise details as per 2011 census data.

S.No.	*State/District*	*Total Population*	*Males*	*Females*	*Percentage decidal growth 2001-2011*	*Sex Ratio*	*Male Literacy Rate*	*Female Literacy rate*
1	ODISHA	41.947.358	21.201.678	20,745,680	13.97	978	82.40	64.36
2	BARC-ARH	1478833	748332	730501	9.84	976	84.28	65.84
3	JHARSUGUDA	579499	297014	282435	12.56	951	86.27	70.05
4	SAMBALPUR	1044410	529424	514986	12.24	973	85.17	68.47
5	DEBAGARH	312164	158017	154147	13.88	976	82.62	63.36
6	SUNDERGARH	2080664	1055723	1024941	13.66	971	82.13	65.93
7	KENDUJHAR	1802777	907135	895642	15.42	987	79.22	58.70
3	MAYURBHANJ	2513895	1253633	1260262	13.06	1005	74.92	53.18
9	BALESWAR	2317419	1184371	1133048	14.47	957	88.06	72.95
10	BHADRAK	1506522	760591	745931	12.95	981	89.92	76.49
11	KENDRAPADA	1439891	717695	722196	10.59	1006	92.45	79.51
12	JAGATSINGHPUR	1136604	577699	558905	7.44	967	93.20	80.88
13	CUTTACK	2618708	1339153	1279555	11.87	955	90.51	77.64
14	JAIPUR	1826275	926053	900217	12.43	972	87.36	73.37
15	DHENKANAL	1192948	612597	580351	11.82	947	87.08	71.40
16	ANUGUL	1271703	654898	616805	11.55	942	87.06	70.44
17	NAYAGARH	962215	502194	460021	11.30	916	86.63	71.08

S.No.	*State/District*	*Total Population*	*Males*	*Females*	*Percentage decidal growth 2001-2011*	*Sex Ratio*	*Male Literacy Rate*	*Female Literacy rate*
18	KHORDHA	2246341	1166949	1079392	19.65	925	92.55	82.06
19	PURI	1697983	865209	832774	13.00	963	91.84	78.67
20	GANJAM	3520151	1777324	1742827	11.37	981	81.85	61.84
21	GAJAPAU	575880	282041	293839	10.99	1042	65.58	43.59
22	KANDHAMAL	731952	359401	372551	12.92	1037	78.41	52.46
23	BAUDH	439917	220993	218924	17.82	991	84.49	60.44
24	SUBARNAPUR	652107	332897	319210	20.35	959	84.78	63.63
25	BOLANGIRI	1648574	831349	817225	23.29	983	77.08	53.77
26	NUAPADA	606490	300307	306183	14.28	1020	71.55	45.21
2?	KALAHANDI	1573054	785179	787875	17.79	1003	73.34	47.27
28	RAYAGADA	961959	469672	492287	15.74	1048	62.61	39.87
29	NABARANGAPUR	1218762	604046	614716	18.81	1018	59.45	37.22
30	KORAPUT	1376934	677864	699070	16.63	1031	61.29	38.92
31	MALKANGIRI	612727	303913	308814	21.53	1016	60.29	38.95

Figure 1: Odisha Population Chart District Wise with Literacy Rate

The Figure 2 describes the percentage of people residing in urban and rural area of odisha as per census 2011 data.

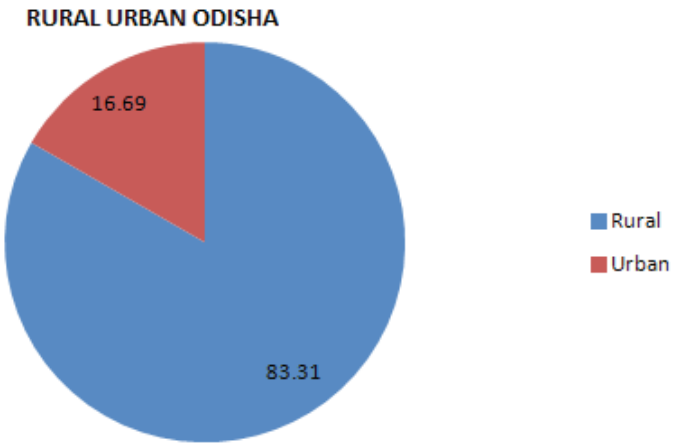

Figure 2: Percentage of People Residing in Urban and Rural Area of Odisha

The percentage value of people speak major Indian Languages are described in Figure 3 and Figure 4. As per 2001 census data, nearly 2.6 crore people are speaking odia language [2]. It consists of plain Odia, Bhatri, Proja, Relli, Sambalpuri and others. The graphical data is described in Figure 5.

The script is written from the Kalinga script, this is one of the many descendents of the Brahmi script of ancient India.

The roadmap to the remaining part of the paper is as follows. In Section 2 we discuss Indian languages in NLP and it's related work. Section 3 proposes information retrieval in odia language and related work. Section 4 we are discussing about sample collection, result analysis and procedure to a methodology for Movie review. Finally, we present the conclusion and future development for our investigation in Section 5.

HINDI	25.071
BENGALI	8.016
TELUGU	7.176
MARATHI	6.97
TAMIL	5.896
URDU	5.009
GUJURATI	4.444
KANNADA	3.669
BHOJPURI	3.217
MALAYAM	3.209
ODIA	3.121
PUNJABI	2.737
RAJASTHANI	1.784
MAGADHI	1.359
CHATISHGARHI	1.289
ASSAMESE	1.242
MAITHILLI	1.184
HARYANVI	0.777
MARWARI	0.771
SANTALI	0.578
MALVI	0.541
KASHMIRI	0.521

Figure 3: Percentage Value of People Speak Major Indian Language

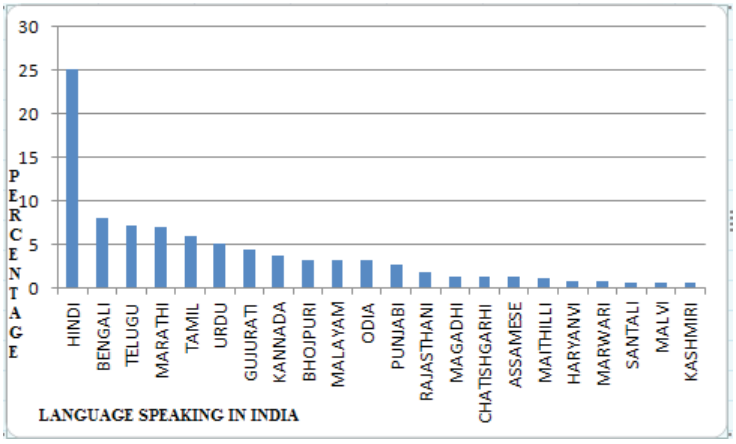

Figure 4: Graphical View of Percentage wise People Speak Indian language

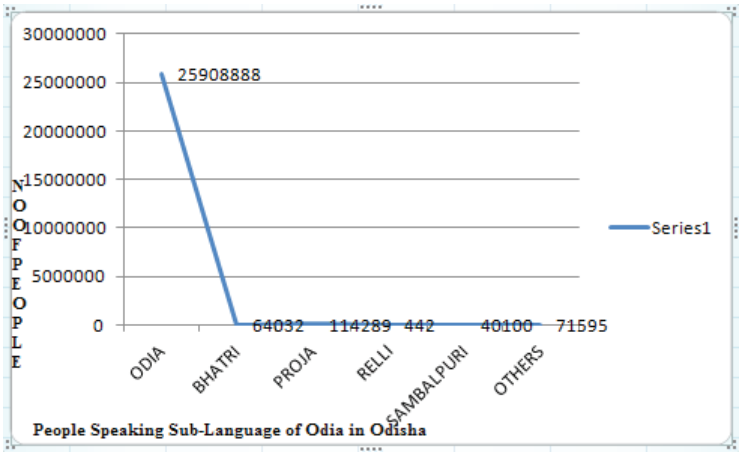

Figure 5: Total no people speaking Odia and it's sub-languages

2. INDIAN LANGUAGE IN NLP

Indian Language is not well equipped in absorbing the technology. The good thing is that 22 Indian languages are adopted by NLP, but still others remain untouched. Though in these Indian languages, the understanding or acceptability of NLP is not uniform. The NLP work first gets into Hindi and then shifted to other languages.

Several tools have been developed to overcome the gap between NLP and IR. As per information extracted from CDAC website, CDAC developed many NLP and released by the Department of Information technology, Govt. of India.

2.1. Related Work on CLIR

Manoj kumar et. al., [7]. described a query based translation approach using dictionaries in their paper on Hindi to English and Marathi to English CLIR System. They use query words that are not available in the dictionary. The resultant translation is then compared with the unique words of the corpus.

D. Thenmozhi et. al., [8] Presents a Tamil English cross lingual information Retrieval System for Agriculture Society. In their research, they developed a CLIR system in agriculture domain for the farmers of Tamil Nadu which helps them to specify their information need in Tamil and to retrieve the documents in English. Local word reordering is performed according to subject-verb-object pattern in order to preserve the relative dependency across the words. Word sense disambiguation is performed that identifies the correct sense of an ambiguous word that is being used in a query.

B. Herbert et. al., [9] presents the combination of query translation approaches for cross-language information retrieval (CLIR). They translate queries with Google Translate and extend them with new translations obtained by mapping noun phrases in the query to concepts in the target language using Wikipedia.

Saurabh Varshney et. al., [10] In their research proposed an algorithm for improving the performance of the English-Hindi CLIR system. He try to use all possible combination of Hindi translated query using transliteration of English query terms and choosing the best query among them for retrieval of documents.

Debasis mandal et. al., [11] explained CLIR system in Bengali and Hindi to English. The cross lingual task includes the retrieval of English documents in response to queries in two most widely spoken Indian Languages, Hindi and Bengali. They use automatic Query Generation and Machine Translation approach.

Nandkishor Vasnik et. al., [12], in his research use the NLP techniques approaches for improving the quality of the search on Internet. In which query extensions and improving the quality of information retrieved using NLP based systems. The Main goal of system, TALASH: A Hindi search Engine is to improve the result provided by Google Search Engine through the extension of user Hindi input. Result of search engine is depending on database present and how structured is it. This research aims to provide linguistic mechanisms that transform and extend the user query by integrating Hindi Word Net semantic database, and user context.

Mallamma V Reddy et. al., [13] described a CLIR system on Kannada English and Telugu English. A query translation based approach using bi-lingual dictionaries is used in this paper. When a query word is not available in the dictionary then the words are translated to utilize the corpus to return by using a rule based approach.

3. INFORMATION RETRIEVAL IN ODIA LANGUAGE

Day By Day the volume of internet users are increasing, so the amount of odia language community. In India more than 90% websites are designed and written in English language and in Odisha state this percentage value

will be more. In spite of being more than 83% living in rural area [1], very less or negligible amount websites are giving product review or Movie review in Odia language.

In past few years, the data repository in odia language has also increased. Govt. of Odisha is planning all its old heritage documents are to be in Odia digitized form. There are no tools and techniques available for the efficient retrieval of information in odia language. However, now the web is getting multi-lingual. In case of odia language, the research is still in preliminary steps. There is vast scope of research in this field. With the coming of the internet revolution, electronic documents are becoming a principle media of business and academic information. Thousands and thousands of electronic documents are prepared and made available on the internet. The study of information retrieval is limited to Odia(Oriya) language and for the years from 2000 to 2015, some studies are in developing stage for other Indian Language, but limited. Department of electronics and information technology (DEITY) came up with the mission of consortia for Machine Translation (MT) systems. English to Indian Language Machine Translation (EILMET) consortium has been formed as a part of this mission. Anuvadaksh is a solution that allows translating the text from English to six other Indian languages, Odia (Oriya) is one of them.

The DIT, Govt. of India is also giving the efforts for marketing language technology in India. Other bodies like HRD (Human Resource Development), DRDO (Defence Research and Development Organisation), AICTE (All India Council of technical Education), and UGC (University Grant Commission) are also putting effort in research & development of Language technology. Odia language is part of that.

Odia speaking people are staying in main cities of India i.e. Kolkota, Mumbai, Delhi, Chennai, Bangalore, Hyderabad, pune, Pondicherry, Gurgaon etc. Odia language also spoken in some of renowned countries of the world i.e. USA, UK, Canada, UAE, Sri Lanka, Singapore, Malaysia, Burma, and Indonesia. There are 45 million odia speaking people living in globally [2].

3.1. Information Retrieval in Odia Language: Related Work

In paper [14], K. R. Shabadi presents the morphological processing of verbal forms in odia in a finite state automation. The paper also proposes a prototype for designing a morphological analyzer for odia verbal forms, which can provide syntactic and lexical information for each lexical unit in the analyzed verbal form.

In paper [15] Jena et. al. describes the work done on creating a morphological analyzer for odia language by using the paradigm approach. The paradigm presents all the words from a given stem and also provides a structure associated with every word and the paradigm have been created.

In paper [16], s.biswas et. al., presents a HYBRID system that applies maximum entropy model with HMM and some rules of linguistic to recognize odia language. In this paper, first maxEnt defines named entities (such as person, degination, abbreviation, organization, location) in odia corpus, then tagging them temporarily as reference.

In paper [17], s.mohanty et. al., describes, how to create an automatic, efficient and effective tool that is able to classify large documents quickly.

In paper [18], s.mohanty et. al., describes the analysis & design of odia Morphological Analyzer (OMA). OMA system is designed related to object oriented analysis to increase its reusability, robustness and extensibility.

In paper [19], s.Mohanty et. al., describes an object oriented model for orient system, this concept uses object oriented programming like java offers to represent and retrieve odia language related information.

In paper [20], sampa et. al., describes a suffix stripping algorithm for Odia stemmer. This paper uses the suffix stripping algorithm to remove the inflectional (bibhakti) suffixes. That algorithm predicts more than 88% result. Lastly she draws a diagram of stemmer using finite automata.

In paper [21], R.C. Balabantaray et. al., describes an Odia text summarization using stemmer. They summarize the Odia paragraph using stemming algorithm.

In paper [22], R.C. Balabantaray et. al., describes the affix removal method. This paper explains how to store the root word in a dictionary, a stop word list in another dictionary.

In paper [23], Dhabal mentions the technique which removes the suffix in paper "design of lightweight stemmer for odia derivational suffixes". First the test has been used the derivational suffixes using suffix stripping algorithm and he found the result 66.25%. Because some words are over stemming and some words are under stemming. To solve that over-stemming problem, an algorithm has been designed to solve the over stemming problem. This algorithm predicts 85% result approximately.

3.2. About ODIA Classifier: An Idea

A noun is enumerated in Odia, it takes a group of morphemes called 'classifier', when the number indicates only 'one' then structure of the numeral phrase is

JANE (ଜଣେ) PILA (ପିଲା) ------ ONE CHILD

DUEE JANA (ଦୁଇ ଜଣ) PILA (ପିଲା) ------ TWO CHILDREN

The classifiers divided into two groups that are Qualifiers and Quantifiers. A mass noun, can be counted with the help of Quantifiers

GILASE PANI (ଗିଲାସେ ପାଣି) ------ ONE GLASS OF WATER

PENTHAE KADALI (ପନ୍ଦେଥଯ କଦଳୀ) ------ ONE BUNCH OF BANANAS

The class of qualifiers to noun is dominated by the classifier

GOTIE (ଗେ ।ଟିଏ) KADALI (କଦଳୀ) ------ ONE BANANA

GOTIE (ଗେ ।ଟିଏ) SEO (ସେଓ) ------ ONE APPLE

DUITIE (ଦୁଇଟି) PILA (ପିଲା) ------ TWO CHILDREN

DUITIE (ଦୁଇଟି) KADALI (କଦଳୀ) ------ TWO BANANAS

Noun is further divided into living and non-living

JANE (ଜଣେ) PILA (ପିଲା) ------ ONE CHILD

KHANDEI BAHI (ଖଣ୍ଡିଏ ବହି) ------ ONE BOOK.

4. PROPOSED WORK AND ANALYSIS: MOVIE REVIEW CONTEXT IN ODIA LANGUAGE

Here in this paper, the actual use of the review analysis can be done for movie. The reviews can be bifurgated as positive or negative based on the reveiew or opinion of the viewer. The huge number of opinions or reviews can be easily sorted into positive or negative. This can help the prospective customer to make their decision to watch the movie based on the feedback. Inspite of huge number of people staying in Rural Area of Odisha, speaking odia language and not understanding the English language. This type of analysis or Movie Review website rarely available in Internet. This makes the viewer to intelligently select the movie and also save their

valuable time. It focuses them to summarize information and potentially apply them in watching movie. This type analysis for mobile product reveiw work done in our earlier paper" Challenges for Information Retrieval in Big data: Product Review Context" by sanjib et. al., [25]. In past while classifying the sentence, the frequency of the term is important, but then it was observed that the presence of one single negative word can make the whole sentence negative. So the point to be noted that the presence of the negative word in appropriate place in the sentence can make the complete sentence negative. Though the complete text is positive but presence of negative word at the end of the sentence make the complete text negative. for example the music is excellent, the background score is superb but the acting of actor cannot hold the viewer.

4.1. Proposed System for ODIA Movie Review

We have followed the following steps for getting results:

1. Collected the reviews from different movie websites.
2. Build training data by using some ODIA text corpus
3. Extracted the features from the reviews in ODIA language like.
4. Tagged the opinion of the reviews
   (i.e. +ve &-ve) by our software tool
5. Does training by taking sample data.
6. Does testing by taking sample data.
7. Obtaining the result.

4.2. Description of the System

Following steps to implement the +ve and –ve review classification for ODIA Language

i) Create and read ODIA text corpus from text File (UTF-8).

ii) For each sentence, point out the attribute and mark them.

iii) Mark and remove the stop words which are very common words.

iv) Analyze the sentiment polarity for each word from the tool.

v) Create and generate the polarity matrix file.

vi) Build a classifier model by correlating words.

vii) Use SVM or Weka tool classifiers for classification.

viii) Build Test data set.

ix) Apply model on new test data and compare the results.

4.3. Step by Step Approach

The complete Odia sentences are separated into words by using the space separator. The complete sentence is broken down into many words using the blank space between the two words. The next step is marking and removing the stop words. Stop words are those words, which do not have any use in determining the sentiment (ଭାବ) of the sentence.

The most common words used in movie review (ଚଳଚିତ୍ରର ସମୀକ୍ଷା) like Music (ସଗୀତ), cinematography (ପତଚିତ୍ର), acting (ଅଭିନୟ), story (କାହାଣୀ), direction (ନିଦ୍ଦେର୍ଶିତ), choreography (ନାଚ), box office (ବକ୍ସ ଅଫିସ), expensive (ବ୍ୟୟବହୁଳ), screenplay (ବିଷୟବସ୍ତୁ), dialog (ଡାୟଲୋଗ), production (ନିମର୍ାଣ), expenditure (ଖଚର୍), background score (ପୃଷ୍ଠଭୂମି), surroundings (ଚତୁପର୍ାଶ୍ୱର୍), imaginary (କଳ୍ପନାକୃତ), screenplay (ସ୍କ୍ରିନପ୍ଲେ), comedy (ହାସ୍ୟପଚ), storyline (ବିଷୟବସ୍ତୁ) etc are collected and made a list of it. These are called attribute list. The odia sentences are then compared with the list of odia stop words and those when identified are removed from the sentence.

This is done efficiently and is with as less error as possible. It tries to acquire the best, better quality and accurate word. After analyzing the review score of each word calculate the score for the whole sentence. Finally we aggregate the review score for each word to form the overall rating of the document. There is very less or no website available for describing the movie review in Odia Language. So the review made in English Language is converted into odia language and store in a Text File (seen in Figure 6-7). The Text Corpus file we have used in this study has different ODIA words, their meaning, and opinion values like positiveness or negativeness based on common use of words. To find the words from ODIA text document we use utf-8 encoding based data files and java reader to read ODIA text files (seen in Figure 11).

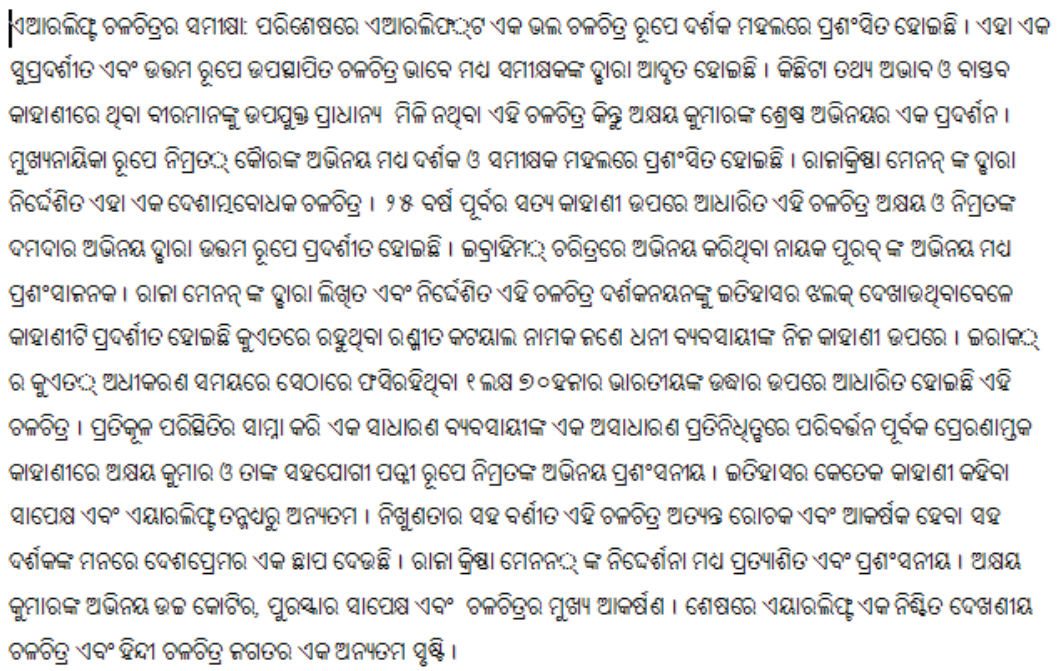
ଏଆରଲିଫ୍ଟ ଚଳଚିତ୍ରର ସମୀକ୍ଷା: ପରିଶେଷରେ ଏଆରଲିଫ୍ଟ ଏକ ଭଲ ଚଳଚିତ୍ର ରୂପେ ଦର୍ଶକ ମହଲରେ ପ୍ରଶଂସିତ ହୋଇଛି । ଏହା ଏକ
ସୁପ୍ରଦର୍ଶୀତ ଏବଂ ଉତ୍ତମ ରୂପେ ଉପସ୍ଥାପିତ ଚଳଚିତ୍ର ଭାବେ ମଧ ସମୀକ୍ଷକଙ୍କ ଦ୍ୱାରା ଆବୃତ ହୋଇଛି । କିଛିଟା ତଥ୍ୟ ଅଭାବ ଓ ବାସ୍ତବ
କାହାଣୀରେ ଥିବା ବୀରମାନଙ୍କୁ ଉପଯୁକ୍ତ ପ୍ରାଧାନ୍ୟ ମିଳି ନଥିବା ଏହି ଚଳଚିତ୍ର କିନ୍ତୁ ଅକ୍ଷୟ କୁମାରଙ୍କ ଶ୍ରେଷ୍ଠ ଅଭିନୟର ଏକ ପ୍ରଦର୍ଶନ ।
ମୁଖ୍ୟନାୟିକା ରୂପେ ନିମ୍ରତ୍ କୌରଙ୍କ ଅଭିନୟ ମଧ ଦର୍ଶକ ଓ ସମୀକ୍ଷକ ମହଲରେ ପ୍ରଶଂସିତ ହୋଇଛି । ରାଜାକ୍ରିଷ୍ଣା ମେନନ୍ ଙ୍କ ଦ୍ୱାରା
ନିର୍ଦ୍ଦେଶିତ ଏହା ଏକ ଦେଶାତ୍ମବୋଧକ ଚଳଚିତ୍ର । ୨୫ ବର୍ଷ ପୂର୍ବର ସତ୍ୟ କାହାଣୀ ଉପରେ ଆଧାରିତ ଏହି ଚଳଚିତ୍ର ଅକ୍ଷୟ ଓ ନିମ୍ରତଙ୍କ
ଦମଦାର ଅଭିନୟ ଦ୍ୱାରା ଉତ୍ତମ ରୂପେ ପ୍ରଦର୍ଶୀତ ହୋଇଛି । ଇବ୍ରାହିମ୍ ଚରିତ୍ରରେ ଅଭିନୟ କରିଥିବା ନାୟକ ପୂରବ୍ ଙ୍କ ଅଭିନୟ ମଧ
ପ୍ରଶଂସାଜନକ । ରାଜା ମେନନ୍ ଙ୍କ ଦ୍ୱାରା ଲିଖିତ ଏବଂ ନିର୍ଦ୍ଦେଶିତ ଏହି ଚଳଚିତ୍ର ଦର୍ଶକନୟନଙ୍କୁ ଇତିହାସର ଝଲକ୍ ଦେଖାଉଥିବାବେଳେ
କାହାଣୀଟି ପ୍ରଦର୍ଶୀତ ହୋଇଛି କୁଏତରେ ରହୁଥିବା ରଣ୍ଜୀତ କଟୟାଲ ନାମକ ଜଣେ ଧନୀ ବ୍ୟବସାୟୀଙ୍କ ନିଜ କାହାଣୀ ଉପରେ । ଇରାକ୍
ର କୁଏତ୍ ଅଧୀକରଣ ସମୟରେ ସେଠାରେ ଫସିରହିଥିବା ୧ ଲକ୍ଷ ୭୦ହଜାର ଭାରତୀୟଙ୍କ ଉଦ୍ଧାର ଉପରେ ଆଧାରିତ ହୋଇଛି ଏହି
ଚଳଚିତ୍ର । ପ୍ରତିକୂଳ ପରିସ୍ଥିତିର ସାମ୍ନା କରି ଏକ ସାଧାରଣ ବ୍ୟବସାୟୀଙ୍କ ଏକ ଅସାଧାରଣ ପ୍ରତିନିଧିତ୍ୱରେ ପରିବର୍ତ୍ତନ ପୂର୍ବକ ପ୍ରେରଣାମ୍ରକ
କାହାଣୀରେ ଅକ୍ଷୟ କୁମାର ଓ ତାଙ୍କ ସହଯୋଗୀ ପତ୍ନୀ ରୂପେ ନିମ୍ରତଙ୍କ ଅଭିନୟ ପ୍ରଶଂସନୀୟ । ଇତିହାସର କେତେକ କାହାଣୀ କହିବା
ସାପେକ୍ଷ ଏବଂ ଏୟାରଲିଫ୍ଟ ତନ୍ମଧ୍ୟରୁ ଅନ୍ୟତମ । ନିଖୁଣତାର ସହ ବର୍ଣୀତ ଏହି ଚଳଚିତ୍ର ଅତ୍ୟନ୍ତ ରୋଚକ ଏବଂ ଆକର୍ଷକ ହେବା ସହ
ଦର୍ଶକଙ୍କ ମନରେ ଦେଶପ୍ରେମର ଏକ ଛାପ ଦେଉଛି । ରାଜା କ୍ରିଷ୍ଣା ମେନନ୍ ଙ୍କ ନିର୍ଦ୍ଦେଶନା ମଧ ପ୍ରତ୍ୟାଶିତ ଏବଂ ପ୍ରଶଂସନୀୟ । ଅକ୍ଷୟ
କୁମାରଙ୍କ ଅଭିନୟ ଉଚ୍ଚ କୋଟିର, ପୁରସ୍କାର ସାପେକ୍ଷ ଏବଂ ଚଳଚିତ୍ରର ମୁଖ୍ୟ ଆକର୍ଷଣ । ଶେଷରେ ଏୟାରଲିଫ୍ଟ ଏକ ନିଶ୍ଚିତ ଦେଖଣୀୟ
ଚଳଚିତ୍ର ଏବଂ ହିନ୍ଦୀ ଚଳଚିତ୍ର ଜଗତର ଏକ ଅନ୍ୟତମ ସୃଷ୍ଟି ।

Figure 6: Text File written in Odia Language for Movie Review

Some of the positive words are in odia language are like Bhala (ଭଲ) means Good, Bada (ବଡ଼) means big, Dhala (ଧଳା) means White, Sahajia (ସହଜିଆ) means Easy, Khushi (ଖୁସି) means Happy. Similarly some of negative words are Damikia (ଦାମିକିଆ) means costly, Choto means Small, Nahan (ନାହଁ) means don't, nahin (ନାହିଁ) means No, Naheen (ନୁହେଁ) means will not. Un success (ବିଫଳ), expensive (ବ୍ୟୟବହୁଳ), very simply (ଅତ୍ୟଧିକ ସରଳତା ପୁଣ୍ଡର୍), irritate (ବିରକ୍ତି), couldn't shown (ଦେଖାଇପାରିନାହିଁ), flop (ଫ୍ଲପ୍), not given thanks (ପ୍ରଶଂସା ପାଇପାରିନାହିଁ), sleepy (ନିଦ୍ରାଜନକଚଳଚିତ୍ର ସାଜିଛି), unidirectional (ଅଦିଗ), ବିଲୁପପ୍ତ, painful (କଷ୍ଟ), scarcity (ଅଭାବ), hopeless (ନିରାଶ), mistake (ତ୍ରୁଟି), not available (ନ ଥିବା), low grade (ନିକୃଷ୍ଟ), comedy less (ହସହୀନ), not eligible (ଅଯୋଗ୍ୟ), Don't Give(ù\C ^ûjó).

Sentiments (ଭାବ) are considered as the manifestation of our feelings and emotions. This field of computer science deals with analyzing and predicting the hidden information stored in the text. This hidden information provides valuable insights about user's intentions, taste and likeliness. Sentiment (ଭାବ) Analysis focuses on

categorizing the text at the level of subjective and objective nature. Subjectivity indicates that the text contains/ bears opinion content whereas Objectivity indicates that the text is without opinion content.

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 ସାବ୍ୟସ୍ତ ହୋଇଛି । ଏପରି ଏକ ଅଦିଗ ଚଳଚିତ୍ର କିଛି ମାତ୍ରାରେ ବିଲୁପ୍ତ ପ୍ରାୟ ଉତ୍କର୍ଷତାପୂର୍ଣ୍ଣ ଭାବନା ପ୍ରଦର୍ଶିତ କରୁଥିବା ବେଳେ ଦର୍ଶକମାନେ ଏହି ଚଳଚିତ୍ରର ବିଷୟବସ୍ତୁ ଜାଣିବା ପାଇଁ ମଧ୍ୟ ବିଶେଷ କଷ୍ଟ ଅନୁଭବ କରିଛନ୍ତି । ପରୀ ଅପସରା କାହାଣୀ ପରି ପୃଷ୍ଠଭୂମି ଯାହାକି ସାଧାରଣତଃ ଇଂଲାଣ୍ଡର ରୋମାଣ୍ଟିକ ଦୃଶ୍ୟ ଗୁଡିକରେ ଦେଖିବାକୁ ମିଳିଥାଏ, ଚଳଚିତ୍ରରେ ବ୍ୟବହୃତ ମହଲ ଏବଂ ଚତୁର୍ପାର୍ଶ୍ୱର ସବୁକିମା ବଳରେ ତାହା ଉତ୍ତମ ରୂପେ ପ୍ରଦର୍ଶୀତ । କିନ୍ତୁ ମୁଖ୍ୟ ରୂପେ ରୋମାଣ୍ଟିକ ଚଳଚିତ୍ରର ଯାଦୁର ଅଭାବ ପରିଲକ୍ଷିତ ହୋଇଛି । କଳ୍ପନାକୃତ ଏହି ଚଳଚିତ୍ର ଆଶାରୂପକ ଆନନ୍ଦପ୍ରକାଶରେ ମଧ୍ୟ ନିରାଶ କରିଛି । ଚଳଚିତ୍ରଟି ନିଜର ସ୍ୱଭାବ ରୂପକ କାହାଣୀରେ ଲୋକମାନଙ୍କୁ ମୁଗ୍ଧ କରିବା ଅପେକ୍ଷା ଛୁଟିଗୁଡିକରେ ଲୋକଙ୍କୁ ନିରାଶ କରିଛି । ଉକ୍ତ ଚଳଚିତ୍ରରେ କିଛି କଳ୍ପନାଜନିତ ଅଭାବ ରହିଥିବା ବେଳେ ମଜା ନ ଥିବା ଡାୟଲୋଗ ଓ ନିକୃଷ୍ଟ ସ୍କ୍ରିନପ୍ଲେ ଦର୍ଶକଙ୍କୁ ନିରାଶ କରିଛି । କିଛିଟା ହାସ୍ୟପଙ୍ଚ ହସହୀନ ଏବଂ ଅନ୍ୟ କିଛିଟା ବୁଝିବା ଅଯୋଗ୍ୟ ହୋଇଛି ।

Figure 7: Another text file written in Odia Language for Movie Review which is having some more attributes

Some examples

1. **Subjective:** ଏହି ଭାଷାଟି ବହୁତ ସହଜିଆ ଆଟେ

   [This language is very easy]

   This sentence has opinion, it talks about the language and the writer's feeling about the language "easy" and hence it is subjective.

2. **Objective**: ଏଆରଲିଫ୍ଟ ଚଳଚିତ୍ର ରେ ଅକ୍ଷୟ କୁମାର ନିମ୍ରତ୍ କୌର ଅଛନ୍ତି [Airlift movie star Akshay kumar and Nirmit kaur]

This sentence is a fact giving general information rather than opinion or view on something and hence it is objective. The subjective can be further categorized into 3 broad categories based on sentiments expressed in text.

1. **Positive:** ଆମେ ଏଠାରେ ଖୁଶୀରେ ଅଛୁ

   [We are happy here]

2. **Negative:** ଏହା ଠିକ୍ ଶବଦ ମୁହେ

   [This is not a correct word]

3. **Neutral:** ମତେ ଦିପହରକୁ ଭୋକଲାଗି ଯାଏ

   [I get hungry by noon**]**

The feature matrix for odia movie review has been created by the developed software tool. The snapshot of Java code in shown in Figure 11.This feature matrix will be input for Weka tool. For this data, the 10 fold validations have been conducted with the help of Weka tools and sequence of procedures and the respective snapshots of the processing are given below (Figure 8 to 10)

```
@relation 'Sentence classification'
@attribute ସଙ୍ଗୀତ numeric
@attribute ପଟଚିତ୍ର numeric
@attribute ଅଭିନୟ numeric
@attribute କାହାଣୀ numeric
@attribute ନିର୍ଦ୍ଦେଶିତ numeric
@attribute ନାଚ numeric
@attribute ବକ୍ସଅଫିସ numeric
@attribute ଚଳଚିତ୍ର numeric
@attribute ନିର୍ମାଣ numeric
@attribute ଖର୍ଚ୍ଚ numeric
@attribute ନିର୍ଦ୍ଦେଶକ numeric
@attribute ପୃଷ୍ଠଭୂମି numeric
@attribute ଚତୁର୍ପାର୍ଶ୍ୱ numeric
@attribute କଳ୍ପନାକୃତ numeric
@attribute ଡାୟଲୋଗ numeric
@attribute ସ୍କ୍ରିନପ୍ଲେ numeric
@attribute ହାସ୍ୟପଟ୍ଟ numeric
@attribute ବିଷୟବସ୍ତୁ numeric
@attribute କଳ୍ପନାଜନିତ numeric
@attribute class {'positive', 'negative'}
@data
```

Figure 8: Showing Attribute Properties in the Input File

```
@attribute class {'positive', 'negative'}@data
0,0,0,0,0,0,0,18,0,0,0,0,0,0,0,0,0,0,0,positive
0,0,0,0,0,0,0,18,0,0,0,0,0,0,0,0,0,0,0,positive
0,0,6,4,0,0,0,18,0,0,0,0,0,0,0,0,0,0,0,negative
0,0,6,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,positive
0,0,0,0,2,0,0,18,0,0,0,0,0,0,0,0,0,0,0,positive
0,0,6,4,0,0,0,18,0,0,0,0,0,0,0,0,0,0,0,positive
0,0,6,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,positive
0,0,0,4,2,0,0,18,0,0,0,0,0,0,0,0,0,0,0,positive
0,0,0,0,0,0,0,18,0,0,0,0,0,0,0,0,0,0,0,positive
0,0,6,4,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,positive
0,0,0,4,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,positive
0,0,0,0,0,0,0,18,0,0,0,0,0,0,0,0,0,0,0,positive
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,positive
0,0,6,0,0,0,0,18,0,0,0,0,0,0,0,0,0,0,0,positive
0,0,0,0,0,0,0,18,0,0,0,0,0,0,0,0,0,0,0,positive
0,0,0,0,0,0,0,18,0,0,0,0,0,0,0,0,0,0,0,negative
0,0,0,0,0,0,0,18,1,1,0,0,0,0,0,0,0,0,0,negative
```

Figure 9: Showing the Input Matrix for Weka

```
Correctly Classified Instances          16               55.1724 %
Incorrectly Classified Instances        13               44.8276 %
Kappa statistic                          0
Mean absolute error                      0.5
Root mean squared error                  0.5026
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances               29

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         1          0.552     1         0.711      0.349    positive
                 0         0          0         0         0          0.349    negative
Weighted Avg.    0.552     0.552      0.304     0.552     0.392      0.349

=== Confusion Matrix ===

  a  b   <-- classified as
 16  0 |  a = positive
 13  0 |  b = negative
```

Figure 10: Analysis of Precision, Recall and ROC area for Movie Review using Weka Tool

Figure 11: Software Tool Converting Odia Text to Attributes and Matrix Based on Features

5. CONCLUSION

This analysis has helped to determine ones decision of watching movie or choosing movie through extracting opinion associated with the information. This type of information, we may take from various sources like texts, news, article, comments, blogs, tweets and other social media. This type of analysis will be quite popular and help people for choosing better products, understanding viewers, opinion. Prospective customer makes decision based on reviews and opinions. To classify the opinion in odia language can done efficiently in using this proposed system.

In this paper a new approach has been proposed a general process flow for a formal analysis of odia movie reviews. The data must be first extracted and mined before applying any analytics. This paper have surveyed some movie review data. In future, the proposed system will take more reviews i.e. nearly 10-15 movies for the experimentation and classify it for better result in terms of precision and recall. Still lot of improvement is required in this domain for further improving the result stated above, i.e. precision recall, etc. The study is in the process of using semantics for identification of features/attributes to improve the results i.e. to achieve ROC area nearly equal to 1.

REFERENCES

[1] Abstract of Odisha population census of India data 2011.

[2] Abstract of speakers strength of languages and mother tongues-2000 census of India 2001.

[3] Shah, U. Finin, T. Joshi, A. Cost, R. S. and Mayfield, J: 'Information Retrieval on the Semantic Web.' 10th International Conference on Information and Knowledge Management, November 2002.

[4] D. Manas, C. Hasan, S. Debakar, A. Khandakar: "Focused Web Crawling", A Framework for Crawling of Country Based Financial Data. 978-1-4244-6928-4/10, IEEE, 2010

[5] H. Rui, L. Fen, S. Zhongzhi: "Focused Crawling with Heterogeneous Semantic Information", 2008 IEEEWIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 978-0-7695-3496-1/08, IEEE.

[6] A. Eneko, A. Xabier, O. Arantxa. (2010): "Document Expansion Based on WordNet for Robust IR",'10 Proceedings of the 23rd International Conference on Computational Linguistics, Volume, pages 9–17, Beijing. ACM.

[7] Manoj Kumar Chinnakotla, Sagar Ranadive, Om P. Damani and Pushpak Bhattacharyya, "Hindi to English and Marathi to English Cross Language Information Retrieval Evaluation".

[8] D. Thenmozhi, C. Aravindan "Tamil to English Cross Lingual Information Retrieval System for Agriculture Society".

[9] B. Herbert, G. Szarvas, I. Gurevych "Combining Query Translation Techniques to Improve Cross-Language Information Retrieval"

[10] S. Varshney, J. Bajpai, "Improving performance of English-Hindi Cross Language Information Retrieval using Transliteration of query terms"

[11] D. Mandal, S. Dandapat, M. Gupta, P. Banerjee, S.Sarkar, "Bengali and Hindi to English Cross-language Text Retrieval under Limited Resources", At the 8th Workshop of the Cross-Language Evaluation Forum, Budapest, Hungary, 19-21 September 2007.

[12] N. Vasnik, S. Sahu, D. Roy: "TALASH: A Semantic and context based optimized Hindi search engine", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3, June 2012.

[13] Mallamma V Reddy, M. Hanumanthappa: "Kannada and Telugu Native Languages to English Cross Language Information Retrieval", International Journal of Computer Science and Information Technologies, Vol. 2.

[14] K. R. Shabadi, "Finite State Morphological Processing of Oriya Verbal Forms", In proceeding of EACL2003 workshop on computational linguistics.

[15] Jena, S. Choudhery, H. Choudhery, D.M. Sharma "Developing Oriya Morphological Analyser Using Lt-toolbox", ICISIL, Springer, Communications in Computer and Information Science 139,2011, Page124-129.

[16] S. Biswas, S.P. Mishra, S. Acharya, S. Mohanty "A Hybrid Oriya Named Entity Reorganization System: Harnessing the power of rule", International Journal Of Artificial Intelligence and Expert Systems(IJAE), 2010, Volume-1, Issue 1).

[17] S. Mohanty, P. K. Santi, R. Mishra, R. N. Mohapatra, S.Swain "Semantic Based Text Classification Using WordNet: India Language Perspective", PeterSojka, Key_sum choi, christiane, Pick vossen(eds):GWC, Proceedings, pp.321-324Masaryk University, 2006

[18] S.Mohanty, P.K.Santi, K.P.Das Adhikari, "Analysis and Design of Oriya Morphological Analyzer(OMA):Some Tests With OriNet",Proceedings of Symposium on Indian Morphology, Phonology and Language Engineering, IIT Kharagpur, India, 2005

[19] 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

[20] Samapa Ch patnaik, Sohag Sunder Nanda, Sanghamitra Mohanty "A suffix stripping algorithm for odia stemmer" by at international journal of computational linguistic and natural language processing volume 1

[21] R.C. Balbantray, B. Sahoo, M. Swain, D. K, Sahoo presented a paper IIT-Bh FIRE2012 Submission: MET Track odia.

[22] R.C Balabantaray, B Sahoo,D.K Sahoo, M swain: "Odia Text Summarization using Stemmer", International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 1– No.3, February 2012 – www.ijais.org

[23] 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 page 4594-4597

[24] R. Mohapatra, Lipi hembram: "Morph-Synthesizer for Oriya Language A computational Approach", Language In India, Strength for Today and Bright Hope for Tomorrow, 2010, Volume 10:9

[25] Sanjib k sahu, D. p. mahapatra, R. C. balbantaray: "Challenges for Information Retreival in Big data Product Review Context", International Journal of Computer application (IJCA) Vol. 136, February 2016.