International Journal of Advanced Research and Development
ISSN: 2455-4030
Impact Factor: RJIF 5.24
www.advancedjournal.com
Volume 2; Issue 5; September 2017; Page No. 292-294

Hindi idioms processor: A computational tool to support sentiment analysis

Pritendra Kumar Malakar
Centre for Informatics and Language, Engineering School of Language, MGAHV, Wardha, Maharashtra, India

Abstract
Idioms handling is one of the major and challenging problems of Sentiment Analysis. It is comparatively difficult to recognize and classify idioms automatically in a text than other components of language. So in this research paper, a Computational Tool is proposed which identifies and categorizes Hindi Idioms electronically. It will be highly valuable for both research and practical applications. This system includes various NLP techniques with Hybrid Approach to produce better results with a significant higher level of accuracy.

Keywords: sentiment analysis, idioms, nlp, hybrid approach

1. Introduction

Sentiment Analysis

Sentiment analysis, also called opinion mining, is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes. Sentiment Analysis is the process of classifying the polarity of given piece of text into Positive or Negative (and in some cases Neutral). It uses some NLP techniques (Text Normalization, POS Tagging, Chunking, NER, WSD, etc.) to extract the sentimental features (Noun, Adjective, Adverb, Negation, Multiword Expressions, Syntactic dependency, etc) from the text and classify them according to their polarity (Positive Negative or Neutral).

2. Idioms Processing: Major problem of Sentiment Analysis

Idioms processing is a harder problem to develop Computational Systems for Sentiment Analysis. Idioms processing includes Store, Identification, Analysis and Classification of idioms automatically. It is comparatively difficult to process idioms automatically than other components of language (Noun, Pronoun, Verb, Adjective, etc). Therefore, Hindi idioms have been selected as a problem in proposed research work. The main objectives of selection of the above problem are as follows:

1. To develop novel methodology for processing Hindi idioms.
2. To build Hindi Idioms based Computational Application.
3. To Co-operate in the field of Hindi-based Language Engineering.

3. Key Challenges

- Hindi language is considered to be poor in terms of research resources because all of the resources are in under development phase. This brings challenges to perform Sentiment Analysis for Hindi. So firstly we have to generate the essential resources required for the Sentiment analysis.
- Hindi is morphologically rich and Free Word Order language, so identification of Subject, Object and Verb is very difficult.
- Negation handling is the biggest problem in Sentiment Analysis because some negation word can reverse the polarity of sentences.

4. Literature Survey

The work of collection or compilation of idioms has been prominently highlighted in most of the books related to Hindi idioms. They provide information about literal meaning and usage of idioms in sentences. There are two major research works have been done related to the Linguistic analysis of phrase and idioms of Hindi is following:

1. Muhavara-Mimansa (Dr. Omprakash Gupta, 1960) [2].
2. Hindi Muhavare: Arthee Sanrachana evam Manobhashikata (Dr. Punita Pachauri, 2006) [1].

The research works presented above referenced to help in the study, analysis and classification of idioms and to develop the basic concept and framework of research.

5. Data collection

A total of 4000 Hindi Idiomatic Expression (idioms) has been collected for the linguistic study from 3 major sources mentioned in following table:

Table 1: Sources for Data Collection

S. No	Title	Author	Publisher	ISBN
1	Manak Samanya Hindi	Dr. Prithvinath Pandey	Arihant Publications Pvt. Ltd.	978-81-8348-695-8
2	Manak Hindi Muhavara Lokokti Kosh	Acharya Dr. Harivansh Tarun	Prakashan Sansthan	81-7714-161-9
3	Saraswati Muhavara Kosh	Rajveer Singh 'Darshanik'	Saraswati Vihar	978-81-216-1705-5

6. Computational System

Resource Generation

From the point of view of language engineering Hindi is a scarce resource language because the research work being done in Hindi is in the developing stage. Therefore it is necessary to collect or build resources for the research. The following resources have been created in this research:

- **Database**

The literary meaning of each of the collected idioms in Hindi is classified as positively negative and unbiased by the annotation method giving the main word and form.

Table 2: Structures of Database

S. No	Idioms	Literal Meaning	Keyword	Forms	Polarity
1	कमर कसना	तैयार होना	कमर	कमर कस	**Positive**
2	आंख का अंधा	मूर्ख होना	आंख	आंख का अंधा, आंख की अंधी	**Negative**

- **Set of Rules**

Rules have been demonstrated to determine the polarity of expressions. Under certain circumstances, the polarization of expressions does not change due to the words (positive or negative from opposite), so some pre-defined rules have been developed to control such problem, on the basis of which our proposed system will be implemented.

Table 3: Set of Rules for Negation Handling

S. No.	Rules for Negation Handling
1	If Neg_Idiom Then Polarity= Negative;
2	If Pos_Idiom Then Polarity= Positive;
3	If Neg_Idiom+ Neg_Term Then Polarity= Positive;
4	If Pos_Idiom + Neg_Term Then Polarity= Negative;

- **Introduction to the System**

This window-based system has been named "Idiom Processor" developed mainly for of Hindi Idioms. This system will be implemented on Unicode based Devanagari script for Hindi. The following features of this system are:

- Bi-Lingual Interface
- Sentence Level Sentiment Analysis
- Database Management facility

- **System Interface**

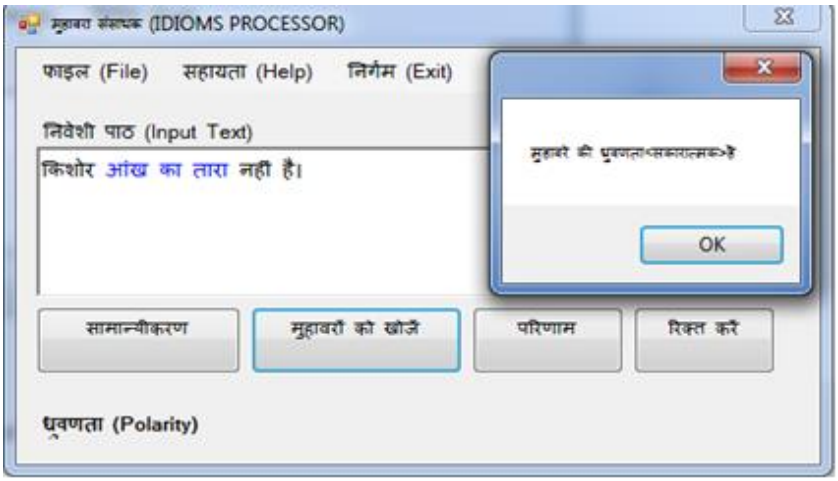

Fig 1

- **Working Procedure**

The proposed computer system will be working in the following steps:

Step 1: **Text Insertion**

In the first phase, the text will be entered into the system.

Step 2: **Data Pre-Processing**

The input text will be splits into sentences because the system will work on sentence-level. Thereafter, undesirable elements (such as special characters, symbols, emoticons, etc.) will be removed from each sentence.

Step 3: Analysis and Classification

This is the most important step in the system. In this stage, the first keyword will be identified from the sentence, which will be present in the Table according to Table 2. After this, all possible forms of the idioms started from that keyword will be extracted from the database and the match in the text. If Idioms is presented then that will be marked and analyzed and classify according to defined rules.

Step 4: the results will be displayed by the system.

- **Evaluation and Testing**

The Idioms Recognition ability of the proposed system is 100% according to the available database. For the evaluation of the system, a corpus of 1,050 simple sentences of Hindi has

been made and the total 8 rules have been defined.

7. Conclusion

The proposed system will prove useful for the processing of Hindi idiomatic expressions. The Morphological Variations, Free Word Order and various syntax phrases of the Hindi language makes the process of analysis and classification process complicated. The field of Hindi language engineering can get new direction from the submitted research paper.

8. References

1. Pachauri Punita. Muhavare: Arthee Sanrachana Evam Manobhashikata. Vaigyanik Evam Takaneeki Shabdavali Ayog, Uttarpradesh, 2006.
2. Gupt Omprakash. Muhavara Meemansa. Bihar-Rashtrabhasha-Parishad Patna, 1881.
3. Ojha, Tribhuvan. Hindi Mein Anekarthakata ka Anusheelan. Vishwavidyalaya Prakashan Varanasi, 1994.
4. Prasad, Dhanji. Theoretical, Applied and Technological Aspects of Linguistics. Priya Sahitya Sadan, Delhi, 2011.
5. Malhotra, Vijay Kumar. Linguistics Applications in Computer. Vani Prakashan, Delhi, 1996.
6. Prasad Dhanji. C# Programming and Linguistics Tools of Hindi. Prakashan Sansthan, New Delhi, 2012.
7. Arora Piyush. Sentiment Analysis for Hindi Language (MS Thesis). IIIT Hyderabad, 2013.