A Rule Based Sentiment Analysis System for Hindi Language

Pritendra Kumar Malakar
Research Scholar, Center for Information & Language Engineering,
MGAHV, Wardha (M.S.) India.

ABSTRACT
Hindi is one of the most spoken languages of the world. Today Hindi language users has better input mechanism to express their sentiments easily on Social Media, so a large volume of User Generated Content in Hindi are digitally store on the internet. It is being seen as an important source of information, but no computational system has yet been available to analyzing these contents. A Sentiment Analysis system has been proposed to solve this problem that analyzes these Hindi contents automatically. The basic principles of Software Engineering and Natural Language Processing have been implemented to design this system. It is a rule based system that follows some linguistics rules to classify any input text into Positive, Negative or Neutral. To evaluate this system, a dataset of 4000 sentences has been created by compiling User Generated Content from Twitter and e-Newspapers. The accuracy to Polarity Classification of the system for Known and Unknown dataset has been measured about 69% and 52%, respectively

Keywords: Hindi, User Generated Content**,** Sentiment Analysis, Bhav Vishleshak, Natural Language Processing

I USER GENERATED CONTENT IN HINDI

Hindi is one of the most spoken languages of the world. Approximately 4.70 % of world's population uses Hindi for their daily life communication [1]. According to a study by KMPG in India and Google, the total number of Hindi language users on the Internet is 254 millions [2]. In view of this scope, encoding standard and linguistic tools have been developed to support Hindi on the internet which has empowered Hindi users to express their sentiments on Social Media. Today Hindi users have been expressing their sentiments towards any subject regularly, so a large volume of User Generated Contents are digitally available on the internet in Hindi. It is being seen as an important source of information by Government, Business Organizations or Individuals to facilitate their decision making processes [3-5].

II PROBLEM STATEMENT

As discussed, a huge amount of User Generated Contents in Hindi available on the Internet, but it brings many serious challenges when it comes to analyze these contents manually. The manual analysis of the contents requires more time and effort that is very complex and tedious task. No computational system has yet been developed to analyzing these Hindi contents. Although a little but important works has been conducted to Sentiment Analysis for Hindi contents. Some of the major work is following:

(a) Akshat Bakliwal and Piyush Arora (IIIT Hyderabad) developed a Hindi Subjective Lexicon of all possible and closely related Synonyms and Antonyms words. They have performed n-Gram Modeling and Machine learning technique to analyze the sentiments from the text [3].

(b) Aditya Joshi, Balamuraly and Pushpak Bhattacharya (IIT Bombay) using **H-SWN** (Hindi-SentiWordNet) in which all sentimental word is classifying into Positive and Negative class with a fixed numerical score [4].

To overcome this problem a Sentiment Analysis system has been developed for Hindi. The detailed description about system development and working procedure has been given below.

III BHAV VISHLESHAK: A SENTIMENT ANALYSIS SYSTEM FOR HINDI

Bhav Vishleshak is a computational system developed to Sentiment Analysis of Hindi contents. The system is mainly designed to classify the given piece of text into Positive, Negative or Neutral automatically. Bhav Vishleshak works only on those Hindi contents that has been written in Unicode based Devanagari Script (Such as-Mangal and Kokila font).

(a) Principles and Technologies used- The basic principles of Software Engineering are applied to develop this system. All the functions of Bhav Vishleshak have been defined according to Natural Language Processing. C#.Net is used to design and code the system using Microsoft Visual Studio 2008. To create the database of the system MS-Access 2007 has been used.

(b) Structure of the system

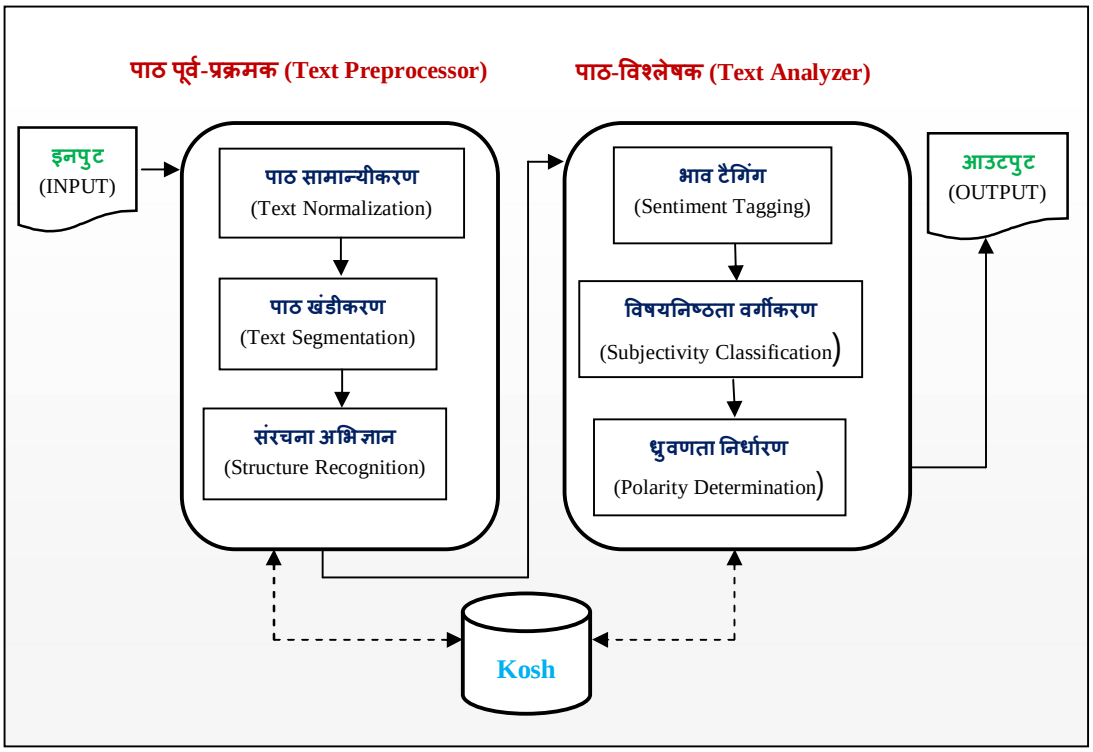

Fig. 1: **System Structure of Bhav Vishleshak**

Above figure shows the internal structure of the Bhav Vishleshak. Two components namely Text Preprocessor and Text Analyzer have been used in Bhav Vishleshak. These are depicted by the rounded rectangle in the figure. Some NLP based processes have been followed by these components to execute these functions. These processes are displayed by the rectangle inside the components. The order of the execution of these processes is sequential. After the completion of one process, the next process begins. It is represented by the arrow in the figure. The Database of the system namely **'Kosh'** is represented by the cylinder shape in the figure. Both the components associated bi-directionally with the database that is depicted by dashed lines.

(c) Text Preprocessor-It is not necessary that input text always be acceptable by the system. Sometimes system is unable to process the text different from standard format. Therefore input is firstly converted into system understandable format by the Text Preprocessor. Following three processes have been used under the Text Preprocessor:

(i) Text Normalization-Text Normalization is a process to remove the undesired elements from the input text in terms of Sentiment Analysis. Various types of symbols, logos, URLs, emoticons, special characters, etc. are used by Social Media users in their expressions which may be affects the accuracy and quality of Sentiment Analysis. Therefore all the unnecessary elements are removed from the text through Text Normalization.

(ii) Text Segmentation-Since Bhav Vishleshak is mainly designed for Sentence level Sentiment Analysis so here Text Segmentation has been used to split the text at sentence level. Initially the input text is divided into sentences and each sentence analyzed separately.

(iii) Structure Recognition-Structure Recognition is a process defined under the Text Preprocessor to identify the structure of the sentences. The structure of sentences is either simple or non-simple. A sentence having more than one sub-sentence is called non-simple sentence. All the sentences are tagged as simple and non-simple. The purpose of finding the structure of sentences here is to identify the sub-sentences used in the sentences, so that they can be analyzed separately as a sentence.

(d) Text Analyzer-This is the second and most important component of the system. To analyze the text it uses following three processes sequentially:

(i) Sentiment Tagging-Sentiment Analysis mainly works on the basis of sentimental words and idioms presented in the text. These are the essential elements for doing Sentiment Analysis, so here Sentiment Tagging has been used to identify all the sentimental words and idioms in the text.

(ii) Subjectivity Classification-It is a process used to classify the sentences into Subjective or Objective. A sentence that contains any sentimental words and/or idioms is called Subjective otherwise it is called Objective.

(iii) Polarity Determination-Polarity Determination is a process used to decide the polarity of each subjective sentence presented in the text. Subjective sentence is classified as Positive, Negative or Neutral by this process. It follows some predefined classification rules to determine the polarity of the text

(e) Kosh-It is a database created under the system. It stores all the information necessary for the system.

(f) Interface of the system

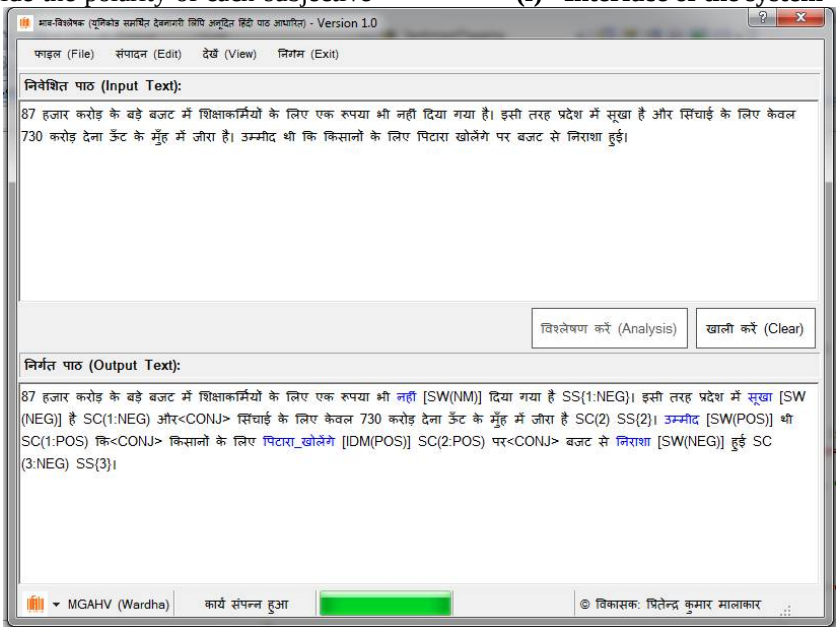

Fig. 2: Interface of Bhav Vishleshak

The above figure shows the main interface of the system in which the names of all the controls for the convenience of users are kept in Hindi and English. Different options have been provided to the user through menu displayed below the titlebar in the interface. The interface contains two textboxes; one to receive input from the user and second one to display output to the user. There are two buttons are used in the interface; one for analyze the input text and second to clear the input box.

(g) Evaluation of the system-The Bhav Vishleshak has been evaluated on Known and Unknown Dataset.

(i) Known Dataset: A Dataset that is referred to develop any system is called Known Dataset. All the sentences are analyzed to Rule Derivation and Database Creation for the system. Since the system is developed on the basis of such dataset, it is treated as Known Dataset.

(ii) Unknown Dataset: A Dataset that is not referred to develop any system is called Unknown Dataset. Normally systems produce higher accuracy on the Known Dataset, so to make system more versatile it should also be evaluated on the Unknown Dataset.

To evaluate the Bhav Vishleshak, a dataset of 4000 sentences is created. The Dataset is divided into Known and Unknown dataset. Each dataset is organized into 10 different sets. In order to maintain the balance between sets, the equal number of sentences has been kept in each set. There are 200 sentences in each set, in which there are 100 positive and 100 negative sentences. The following formulas have been used to calculate the accuracy percentage of the Bhav Vishleshak:

Formula 1: (To SUs Identification)

$$= \left(\sum \textbf{Identified (SUs)} / \sum \textbf{Total (SUs)} \right) \times 100$$

Where, SUs means Sentimental Units **Formula 2: (**To Polarity Classification)

= (Number of correctly classified Sentences / Total Sentences) × 100

Table 1
Accuracy Measurement Table for Known Dataset

S.No.	SET	Accuracy (%)	
		SUs Identification	Polarity Classification
1.	A	60	60
2.	B	90	90
3.	C	70	80
4.	D	80	90
5.	E	40	30
6.	F	90	60
7.	G	50	70
8.	H	70	80
9.	I	40	90
10.	J	80	100
Avg. Acc.		**61**	**69**

Table 2
Accuracy Measurement Table for Unknown Dataset

S.No.	SET	Accuracy (%)	
		SUs Identification	Polarity Classification
1.	K	70	80
2.	L	60	60
3.	M	60	50
4.	N	50	60
5.	O	80	70
6.	P	90	60
7.	Q	70	80
8.	R	30	60
9.	S	40	40
10.	T	10	40
Avg. Acc.		**49**	**52**

Comparison of the accuracy to Polarity classification on the known and unknown dataset of Bhav Vishleshak is shown in the following graph:

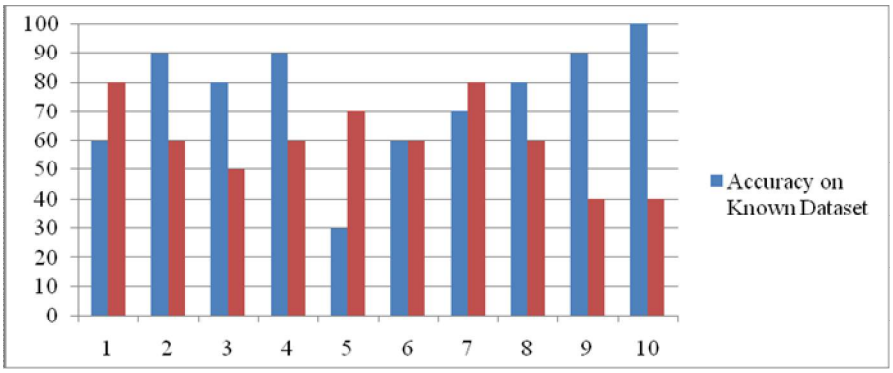

Fig. 3: Comparison graph for known and unknown dataset.

IV CONCLUSION

The above system is an attempt towards Sentiment Analysis of Hindi content. Since it mainly works on predefined rules and knowledge, so its accuracy and quality are limited. If the number of rules and the size of the database are increased, the system will produce better results. To make the system more effective, it has to be customized according to Unknown Datasets. The system is to be also useful for Text Analysis, POS Tagging, Language Teaching, etc.

REFERENCES

[1] Sharma, R., & Bhattacharya, P. (2014). A Sentiment Analyzer for Hindi Using Hindi Senti Lexicon. *ICON 2014*: http://ltrc.iiit.ac.in/icon/2014/proceedings.php.

[2] https://assets.kpmg/content/dam/kpmg/in/pdf/2017/04/Indian-languages-Defining-Indias-Internet.pdf.

[3] Arora, Piyush. (2013) Sentiment Analysis for Hindi Language (MS Thesis), IIIT Hyderabad.

[4] Joshi, A., R, Balamuraly A R.., & Bhattacharyya, P. (2010). A Fall-back Strategy for Sentiment Analysis in Hindi:a Case Study. *Proceedings of ICON 2010: 8th International Conference on Natural Language Processing.* Hyderabad: Macmillan Publishers.

[5] Sharma, R., Nigam,S. & Jain,R. (2013). Opinion Mining In Hindi Language: A Survey. *IJFCST*, Vol.4, No.2.

[6] Pang B., and Lillian Lee ((2008. Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval 2(1-2): 1–135.