Sentiment Classification of Hindi Language using Natural Language Processing Techniques

Pritendra Kumar Malakar[1]

Center for Technology
Studies, School of Language
MGIHU, Wardha (Maharashtra)
pritmalakar@gmail.com

Pravesh Kumar Dwivedi[2]

Center for Technology
Studies, School of Language
MGIHU, Wardha (Maharashtra)
dpraveshkumar@gmail.com

Aarti Kashyap[3]

Assistant Professor
Govt. E. R. R P.G. Science
College, Bilaspur (C.G.)
aarti.kas2009@gmail.com

Abstract: *This paper presents a Hybrid Approach to automatically determine the sentimental word or phrase from Hindi text by using a Hindi sentiment lexicon and classify them into different polarity as Positive, Negative and Neutral.*

Keywords: *NLP, Sentiment Classification, Hybrid Approach.*

1. INTRODUCTION

Sentiment Classification is the computational study of online expressions posted by Users, Peoples or Writers about any particular Product, Service or Topic. It uses some Natural Language Processing (NLP) techniques to classify those expressions into Positive, Negative or Neutral.

Today Web 2.0 applications (Social media, Blogs, Forums, Review channels, wikis and chats) are very popular medium among citizens to express their Sentiment or Opinion publically towards any subject, a large amount of this User Generated Content (UGC) stored on the web in digital form. This information may be meaningful for Government, Organizations or Individuals to make their decision correct and much effective. This brings various challenges when it comes to analyze these expressions automatically. Sentiment Classification is widely in field of Market Research, Business Classification, Policy Making, Social status monitoring and other Decision-Making systems.

2. BACKGROUND

A few amount of research work has been conducted for Sentiment Classification in Indian Languages. Akshat Bakliwal and Piyush Arora (IIIT Hyderabad) [1], developed a **HSL** (Hindi Subjective Lexicon) of all possible and closely related Synonyms and Antonyms and uses n-Gram Modeling and Machine learning technique to analyze the sentiments from the text. Aditya Joshi, Balamuraly and Pushpak Bhattacharya (IIT Bombay) [3], using **H-SWN** (Hindi-SentiWordNet) in which all sentimental word is classifying into Positive and Negative class with a fixed numerical score.

3. PROBLEM STATEMENT

The Sentiment Classification Model of Indian Languages extract only Sentimental Words (Adverb, Adjective, negation words such as- तेज, सुस्त, सुंदर, खराब, नही, etc) from the given piece of text and classifies into Positive, Negative or Neutral level. Parameters used to support Sentiment Classification include terms, part of speech, syntactic dependencies and negation. A Computational Model is required that has ability to identify, interpret and understand such Sentimental Words automatically and produce better Sentiment Classification results with greater accuracy.

4. EXPERIMENT SETUP

A. Dataset

After introducing Unicode (UTF-8) standard for Indian Languages, the number of Websites, Blogs, and Discussion Forums etc. in Hindi has grown rapidly, so the amount of Public Sentiment has increases on the web. After collecting data from these resources, a manual annotation scheme is

performed and creating a Dataset of 1000 sentences to test and trend the proposed Sentiment Classification Model for Hindi. These sentences come from some specific domains like- Social and Political discourse.

B. Resource Generation

Hindi is scarce resources language because all of the resources are in under development phase. So firstly we have to generate the essential resources required for the Sentiment Classification. In order to generate the resources, we create a small Tagset for Hindi POS (Part-of-Speech) describing (in the Table No.1) given below.

Table No.1. *POS (Parts-of-Speech) Tag for Hindi Language*

Category	Type	Hindi Words
Noun (N)	Common(NC)	छात्र, कर्मचारी, जनता
	Proper(NP)	राम, श्याम, मोहन, अरबिंद
Pronoun (P)	Pronominal (PN)	यह, वह , तुम, उसको
Nominal Modifier(J)	Adjective (ADJ)	अच्छा, खराब, सुंदर, कुरूप
Verb (V)	Main Verb (VM)	खाना, पीना, सोना
	Auxiliary Verb(VA)	रहा, चुका, है, था, थे, थी
Adverb(A)	Manner (AMN)	तेज, धीमा, सुस्त, जल्दी,

5. WORKING PROCEDURE
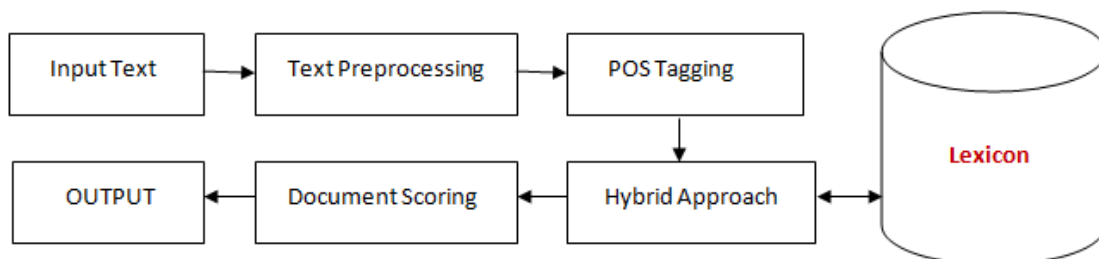
Figure 1. *Working Procedure*

STEP-1: (Input Phase)

Input any text into system. It breaks the entire text into sentences because proposed Sentiment Classification Model works on sentence level.

STEP-2: (Text Preprocessing Phase)

In order to achieve higher accurate results, some pre-processing operation applied on the given text or sentence because there are some noise (special character, symbols, logos, question mark, URLs, smilies etc) in the text that doesn't return any meaning and by the presence of those noises we cannot get the correct Sentiment Classification results.

STEP-3: (POS Tagging Phase)

Not all the words in given sentences are useful for Sentiment Classification, because some of the words hold sentiments. So after the POS (Part-of-Speech) tagging, the sentiment bearing words or group of word is characterized by their associated tag.

STEP-4: (Hybrid Approach Phase)

There are two models used under the Hybrid Approach, they are:

1. **RULE-BASED MODEL**

    This model includes a set of rules for negation handling because in some conditions polarity of sentence could be changed into Positive to Negative and vice versa. The Negation handling is biggest problem to perform Sentiment Classification.

2. STATISTICAL-BASED MODEL

To achieve accuracy, a database of all Sentimental words with their statistical score is developed under this model. The positive Sentimental words assign a positive value (between 0 to 1) and negative to (-1 to 0.)

STEP-5: **(Output Representation Phase)**

The system get the value of every Sentimental Words exists in the sentence and sum will be calculated. If the calculated sum is Positive means, Sentence is Positive otherwise Negative. If the sum is 0 means Sentence in neutral.

6. RESULT EVALUATION

To test our Sentiment Classification Model, A dataset of 1000 sentences have been taken and from which 500 sentences are Positive and 500 are Negative. These sentences are taken as input for Sentiment Classification. After Classification, our system is providing 70 % correct results.

7. CONCLUSION

The proposed Sentiment Classification method on Hybrid Approach which provides better accuracy and efficient results than previous research. It explores new dimension by including Multi-Word Expressions to improve the quality of Sentiment Classification for Hindi Language.

REFERENCES

[1]. Arora, Piyush (2013). Sentiment Classification for Hindi Language (MS Thesis), IIT Hyderabad.

[2]. Das, D. and Bandyopadhyay, S. (2010). Labeling emotion in bengali blog corpus a fine grained tagging at sentence level. In *Proceedings of the Eighth Workshop on Asian Language Resouces*, pg. 47–55, Beijing, China.

[3]. Joshi, A., Balamuraly, A. R. and Bhattacharya, Pushpak (2010). A Fall-back Strategy for Sentiment Classification in Hindi : A Case Study. Proceedings of ICON: 8th International Conference on Natural Language Processing, Macmillan Publishers, India.

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

[5]. Sharma, Richa ; Nigam, Shweta and Jain, Rekha (2013). Opinion Mining In Hindi Language : A Survey. IJFCST, Vol.4, No.2.

AUTHORS' BIOGRAPHY

Pritendra Kumar Malakar received his B. Sc. degree in Computer Science from Guru Ghasidas University, Bilaspur (C.G.) of India in 2006.He received Master of Computer Application (MCA) degree from Chhattisgarh Swami Vivekananda Technical University (CSVTU), Bhilai (C.G.) in 2009.He worked for Indian Institute of Handloom Technology, Champa (C.G.) as a visiting faculty member from Aug 2009 to March 2010. .He was a faculty member of the Department of Information Technology in Dr. C. V. Raman University, Bilaspur (C.G.) from March, 2010 to June 2013. Now he is pursuing Ph.D. in Informatics and Language engineering from Center for Technology Studies, MGIHU, Wardha (Maharashtra). His research interests include Natural Language Processing, Sentiment Analysis and Information Retrieval.

Pravesh Kumar Dwivedi received his Master of Science (computer science) degree from Makhanlal Chaturvedi National university of journalism and communication, Bhopal (M.P.) in 2009. He received Master of Informatics and Language engineering from Center for Technology Studies, Mahatma Gandhi International Hindi University, Wardha (Maharashtra) in 2012. He received M.Phil. in Computational Linguistics from MGIHU, Wardha (Maharashtra) in 2013. Now he is pursuing Ph.D. in Informatics and Language engineering form Center for Technology Studies, MGIHU, Wardha (Maharashtra). He is actively participated in various national, International seminars. His Research interest areas are Natural Language Processing, Speech Processing, Machine Translation and Information Retrieval.

Aarti Kashyap completed her B.C.A. from Makhanlal Chaturvedi Rashtriya Patrakarita Vishwavidyalaya (MCRPV), Bhopal (M.P.). Also completed her Master of Computer Application (M.C.A.) degree from Chhattisgarh Swami Vivekanand Technical University (CSVTU), Bhilai (C.G.) in 2009. She worked for Govt. Girls Bilasa (Autonomous) P.G. College, Bilaspur (C.G.) as a contractual Assistant Professor from August 2010 to May 2013. Currently she is working in Govt. E.R.R. P.G. Science College, Bilaspur (C.G.) as a Contractual Assistant Professor.