**NOVATEUR PUBLICATIONS**
**International Journal of Research Publications in Engineering and Technology [IJRPET]**
**ISSN: 2454-7875**
**VOLUME 3, ISSUE 6, Jun.-2017**

# SENTIMENT ANALYSIS OF MARATHI LANGUAGE

SUJATA DESHMUKH
Department of Information Technology, FRCRCE, Affiliated to Mumbai University, Mumbai, India.
sujata.p.deshmukh@gmail.com

NILEEMA PATIL
Dept. of Physics, LTCE, Affiliated to Mumbai University, Mumbai, India. patil. nileema@gmail.com

SURABHI ROTIWAR
Department of Information Technology, FRCRCE, Affiliated to Mumbai University, Mumbai, India.
surabhi.rotiwar@gmail.com

JASON NUNES,
Department of Information Technology, FRCRCE, Affiliated to Mumbai University, Mumbai, India. jason.nunes@gmail.com

**ABSTRACT:**

**Sentiment analysis offers many benefits and opportunities from business, government and consumer perspective in this digital data explosive age. According to Google, in partnership with KPMG India report (April-2017), titled 'Indian Languages - Defining India's Internet', currently India today has 234 million Indian Language users who are online, compared to 175 million English web users and expected 536 million Indians to use regional languages while online by 2021. However Marathi users are expected to make significant contribution to define India Internet volume. By considering aspects of Marathi language and benefits of sentiment analysis, this paper presents a approach to overcome the barriers and difficulties being faced for analyzing text in Marathi language. The proposed system detects hidden sentiments in text of Marathi language. The system uses sentiment analysis methodology in order to achieve desired functionality. In this system, a corpus based approach is proposed, i.e the creation of a diverse up to date corpus of Marathi keywords, along with their individual polarities, with respect to the Word Net, which is consider as a corpus. The algorithm is used to calculate the cumulative polarity of the text and rank the sentence as positive, negative or neutral on a set scale standard.**

**KEYWORDS: Sentiment Analysis of Marathi Language, sentiment analysis, corpus based approach**

## I. INTRODUCTION:

Sentiment Analysis (SA) or Opinion Mining (OM) is the computational study of people's opinions, attitudes and emotions toward an entity. The entity can represent individuals, events or topics. Sentiment Analysis identifies the sentiment expressed in a text then analyzes it. Therefore, the target of SA is to find opinions, identify the sentiments they express, and then classify their polarity. It 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. It represents a large problem space [2, 3].The analysis of sentiments may be document based, paragraph based or sentence based. In document approach the sentiment in the entire document is summarized as positive, negative or objective. In sentence based where individual sentences, bearing sentiments, in the text are classified. Sentiment Analysis task is considered a sentiment classification problem. Following are the general steps–

- To select text features i.e. obtaining a dataset.
- Identifying Parts of speech (POS)– finding adjectives, as they are important indicators of opinions.
- Identifying Opinion words and phrases– these are words commonly used to express opinions including good or bad, like or hate.
- Negation handling–The appearance of negative words may change he opinion orientation like not good is equivalent to bad.

Calculate sentiment score and summarize as positive, negative or objective [2, 3].

## II. LITERATURE REVIEW:

The following literature review provides, issues, challenges, approaches to understand the system. Walaa Medhat et al.(2014), highlighted that there is still a lack of resources and researches concerning languages other than English sentiment analysis [4]. Authors reviewed Fifty-four of the recently published and cited articles, and were categorized and summarized. This paper provides basic understanding and helped to understand corpus and dictionary based approaches. Soha Ahmed et.al (2013), presented the challenges faced by researchers when conducting Social SA (SSA) on Arabic text [6]. Further

authors suggested some solutions inspired from the literature, and discussed the difficulties faced when performing SSA on Arabic social media text. A pre-processing phase to sentiment analysis is proposed and showed the noticeably improvement to the results of sentiment extraction from Arabic social media data. This paper provided the improvements by pre-processing the text of SA. Hatem Ghorben and David Jacot ( 2012), showed that the combination of lexical, morpho-syntactic and semantic features achieves relatively good performance in classifying French movie reviews according to their sentiment polarity (positive, negative) [7]. In order to extract the semantic orientation of words from SentiWordNet, a standard word translation process is used. This paper provided to use SentiWordNet for corpus based approach for proposed system. Further it argued that dictionary-based approach could contribute better results. Yakshi Sharma et. Al (2015), proposed an approach Sentiment Analysis of Hindi Tweets using SentiWordNet with Subjective lexicon method [8]. Each entry in lexicon is categorized into Verb, Noun, Adjective, and Adverb. The results indicate that proposed algorithm gave better accuracy than unigram presence method. In Hindi language, lexicon coverage can be increased as it has limited coverage now. Further authors suggested to do research on subjective lexicon which can be extended to machine learning, n-grams or both combined. This paper provides motivation to use SentiWordNet and lexicon method. Namrata Godbole et. Al (2012), showed a system that consists of a sentiment identification phase, which associates expressed opinions with each relevant entity, and a sentiment aggregation and scoring phase, which scores each entity relative to others in the same class [5]. Finally, authors evaluated the significance of scoring techniques over large corpus of news and blogs. The above literature summarized in following Table 1.

TABLE I. LITERATURE REVIEW TO IDENTIFY GAP

| Title of paper and authors | Conclusion discussed | Relevance | Research Gap |
|---|---|---|---|
| Sentiment analysis algorithms and applications: A survey Walaa Medhat, Ahmed Hassan, Hoda Korashy 2014[4] | The various sentiment analysis algorithms were noted and why one is better than the other is found. | Corpus based approach was found to be the best approach for our project. | Various other techniques are also available and a combination of techniques can also be favourable. |
| Large-Scale Sentiment Analysis for News and Blogs Namrata Godbole, Manjunath Srinivasiah, Steven Skiena, 2012 [5] | Authors concluded that sentiment can vary by demographic group, news source or geographic location. | Mapping should be done to find the best opinion while using corpus based techniques | In large scale data it is difficult to map the opinions and also due to the various combinations of the words. |
| Key Issues in Conducting Sentiment Analysis on Arabic Social Media Text Soha Ahmed, Michael Pasquier, Ghassan Qadah, 2013 [6] | This paper highlights key issues researchers are facing and innovative approaches that have been developed when performing SSA on Arabic text in general and Arabic social media text in particular. | Paper presented simple pre-processing steps and showed how these steps improve classification accuracy. | Same approach can be applied to Indian languages. |
| Sentiment Analysis of French Movie Reviews Hatem Ghorben, David Jacot, 2012 [7] | In order to extract the semantic orientation of words from SentiWordNe, a standard word-translation process was found to be used. | Although translation does not necessarily preserve the semantic orientation of words due to the variation of language common usage, in spite of all its side it has been argued that dictionary-based approach could achieve better results. | The concept of SentiWordNet can be applied to Indian languages. |
| A Practical Approach to Sentiment Analysis of Hindi Tweets, Yakshi sharma, Veenu mangat, Mandeep Kaur, 2015 [8] | An unsupervised lexicon based approach for SA is discussed | Words like, "NAHI" can invert the polarity of the sentence. So, these words are also considered in finding Polarity of text. | Lexicon coverage can be increased that present one which is limited. |
| Using SentiWordNet for Multilingual Sentiment Analysis Kerstin Denecke, 2008 [9] | Document is translated into English using standard Translation software. Then, the translated document is classified according to its sentiment into one of the classes "positive" and "Negative". | By means of SentiWordNet, scores for positivity and negativity are determined for these words. An interpretation of the scores then leads to the document polarity. | Performance of system is depending on translation software. |
| Sentiment analysis algorithms and applications: A survey Walaa Medhat, Ahmed Hassan, Hoda Korashy 2014[4] | The various sentiment analysis algorithms were noted and why one is better than the other is found. | Corpus based approach was found to be the best approach for this system. | Various other techniques are also available and a combination of techniques can also be favourable. |

## III. PROBLEM STATEMENT:

The aim of this system is to address the overall problem of sentiment analysis being faced in the Marathi language. The idea behind the development of this system is twofold; Firstly, to create a diverse up to date corpus consisting of Marathi keywords like adverbs, adjectives etc. along with their individual polarities. This in keeping with other sentiment analysis literature can be called Marathi Word Net. Secondly, aim of this system is to carry out advanced sentiment analysis using optimal algorithm to obtain the cumulative polarity of text. Thus, this twofold problem statement is converted into following objectives:

• To create an system that that uses a corpus based approach and can carry out advanced Sentiment Analysis for a vernacular language i.e. Marathi that detects hidden sentiments in text and analyzes the content accordingly.

• To create an system that that uses a corpus based approach and can carry out advanced Sentiment Analysis for a vernacular language i.e. Marathi that detects hidden sentiments in text and analyzes the content accordingly.

## IV. PROPOSED SYSTEM:

The proposed sentiment analysis approach is a Corpus based approach. It is twofold and involves two modules. First module provides foundation to the proposed system by creating of feasible corpus for Marathi language from English SentiwordNet and second module presents mapping of keywords to analysis sentiments. The main tasks of creation of feasible corpus from English SentiwordNet are Categorization of keywords with respect to parts of speech (POS) and allocating polarity to keywords. The main steps for Mapping of keywords to analysis of sentiments are accept Marathi content as input for analysis and use the devised algorithm to find polarity and degree of given input.

### A. CORPUS GENERATION FROM ENGLISH SENTI WORD NET:

This system proposed the creation of a Marathi corpus. In order to generate the corpus, steps are as follows:

1. Collection of data online: Find text files which contained lists of various Marathi keywords and their meanings etc.
2. Segregation and lemmatization: It pre-processed the data to categorize words as PoS i.e parts of speech, ie Adjective etc and another corresponding lemma file.
3. Conversion to csv format: After pre-processing, it converted the text files into 11 csv format in order to be able to index the keywords properly. Every keyword was attributed a unique identification number. Then Data File is created, which has the same corresponding UIDs and the meaning of the words, followed by its synonyms to facilitate mapping.
4. Attributing polarity to each keyword: polarity of each word is calculated for word using English sentiwordnet. Then, for the purpose of easy indexing, the text files are converted to csv format.

### B. ALGORITHM

The algorithm first find individual polarity of each word in the sentence and then find the cumulative polarity to determine if the sentiment is positive, negative or neutral and to what degree. Sentence in Marathi is a input for the algorithm. The steps are follows:

• Elimination of stop words. Any words that do not attribute any specific polarity to the sentence are found to be redundant while calculating the overall polarity of the sentence. Hence they are known as stop words.

• Obtaining the polarity of relevant keywords from the corpus. Here, it is +0.75 and +0.5.

• Calculating cumulative polarity.

The cumulative polarity P is calculated as P=(a+b+c+….)/n where; a,b,c are the individual polarities mapped from the corpus, n= No of words.

## V. RESULTS AND DISCUSSION:

This system is implemented in Java. Following screen shot shows the output of the system.


Figure 1: Highly positive output

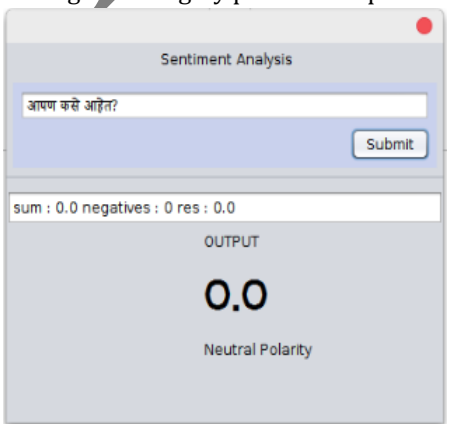


Figure 2 Neutral polarity output

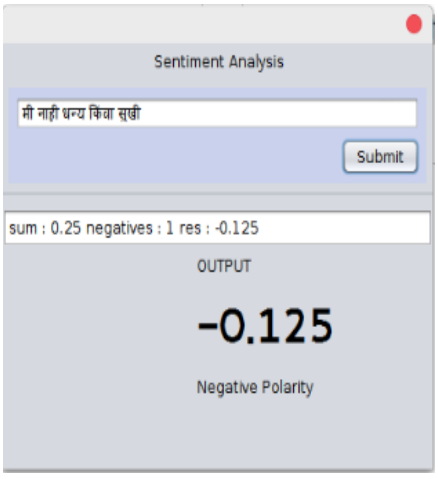


Figure3: Negative polarity output

Following Table 2 provides different test cases its polarity.

Table 2: Test cases with Result

| Input | Output |
|---|---|
| मला वाटते ती, अतिशय सुंदर आहे | 0.84, Highly Positive |
| किती सुंदर आहे ती! | 0.75 Highly Positive |
| ती आहे म्हणून दु: खी? | 0.0 Neutral Polarity |
| मी खूप आनंदी आहे | 0.14 Positive Polarity |
| मी आनंदी नाही | -0.12 Negative Polarity |
| मी नाही धन्य किंवा सुखी | -0.125, Negative polarity |
| तो नाही आहे की, मी आनंदी नाही | 0.125, Positive Polarity |
| "तुम्ही कसे आहात?" | 0.0 Neutral Polarity |
| ""मी दु: खी आहे" | 0.0 Neutral Polarity |

After careful analysis of the test cases, the few discrepancies in the obtained results were mainly due to the following 3 factors:

1) Corpus word limit and Translation accuracy : If words are not present in corpus built, then online Yandex translator is used and sentiment is analyzed with English SentiWordNet. The freely available Yandex translator that system is using provides an accuracy in the range of 60-70%.Thus, especially for Asian, vernacular languages, the translation is only partially accurate. Thus, these inaccurate translations affect the overall polarity of the sentences which can result in minor discrepencies in the overall output. However, for smaller and simple sentences, the translation accuracy is comparatively higher than that of complex compound sentences. It was an observation that the framing of the sentences makes a difference in the sentiment analysis.

2) Limited scope of English SentiWordNet: The English SentiwordNet 3.0 which system is using to obtain the polarities of the respective words, also needs a lot of optimization. Many words which are actually of neutral polarity, are misclassified as adjectives having higher or lower polarity and hence either provide wrong polarities or tamper with the algorithm.

3) Non acceptance of special characters: This system perfectly analyzes and parses basic punctuation like commas, question marks, exclamation marks etc., but a special character like "" etc. is parsed along with the word, the system fails to recognize and process it and doesn't give any output. Hence there is a need to include a separate exception class to eliminate these special characters.

## VI. CONCLUSION:

Sentiment Analysis has been quite popular and has lead to building of better products, understanding user's opinion, executing and managing of business decisions. With rapidly increasing technology, the early approach of word-of-mouth has been shifted towards the mass opinion what the people like and appreciate in majority. The rise in user-generated content for Marathi language across various genres- news, culture, arts, sports etc has opened the data to be explored and mined effectively, to provide better services and facilities in terms of sentiment analysis. The scarcity of resources is one of the biggest challenges while dealing with sentiment analysis for Marathi language. This system focused on resource creation which includes building of an up to date corpus for Marathi language. The algorithm being proposed is used to give cumulative polarity using the Wordnet. This sentiment analysis model proposes a novel and effective approach to achieve desired functionality for Marathi language. This system mainly focused on creation of a vast and diverse up-to-date corpus, efficient mapping of data and generation of accurate sentiments for the data to erase the language barriers faced in the field of Sentiment analysis for Marathi.

## VII. FUTURE SCOPE:

The scope of this system is limited for sentence level which can further be increased so as to analyze the sentiment for paragraphs and even larger text documents. This can be done, by firstly finding the individual polarity of the sentences using our suggested algorithm and then finding the net cumulative polarity of all sentences to obtain the overall polarity of the paragraph or document. There is a need for an optimized and accurate algorithm to do the same. Further limited size of corpus can be increased by not only considering adjectives but also verbs and nouns which are already present in the English Senti Word Net. This while evaluating the sentiment, better accuracy and results can be obtained. Also, increase the dictionary size by regularly updating it with new words. Further real time update of dictionary can be future research direction in the field of sentiment analysis of Marathi language.

## REFERENCES:

1) Report *'Indian Languages - Defining India's Internet'*, a study by KPMG in India and Google April 2017.
2) Sentiment analysis, https:// en.*wikipedia.org/ wiki/ Sentiment_ analysis*, accessed in Aug. 2016.
3) Alessia D'Andrea, Fernando Ferri, Patrizia Grifoni, "*Approaches, Tools and Applications for Sentiment Analysis Implementation*", International Journal of Computer Applications (0975 – 8887) Volume 125 – No.3, September 2015
4) H. K. Walaa Medhat, Ahmed Hassan, "*Sentiment analysis algorithms and applications: A survey*," Ain Shams Engineering Journal (2014) 5, 1093–111, 2014.
5) S. S. Namrata Godbole, Manjunath Srinivasaiah, "*Large-scale sentiment analysis for news and blogs*," ICWSM'2007, Boulder, Colorado,USA, 2007.
6) G. Q. Soha Ahmed, Michael Pasquier, "*Key issues in conducting sentiment analysis on Arabic social media text*," IIT`13, 2013.

7) D. J. Hatem Ghorben, "*Sentiment analysis of french movie reviews*," Proceedings of the 7th Atlantic Web Intelligence Conference, AWIC 2011, pg. no. 19-28, 2011.
8) M. K. Yakshi sharma, Veenu mangat, "*A practical approach to sentiment analysis of Hindi tweets,*" 1st International Conference on Next Generation Computing Technologies (NGCT), pg. no. 677-680, 2015.
9) Denecke, Kerstin. "*Using sentiwordnet for multilingual sentiment analysis*." Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on. IEEE, 2008.