Proceedings of the Third International Conference on Computing Methodologies and Communication (ICCMC 2019)
IEEE Xplore Part Number: CFP19K25-ART; ISBN: 978-1-5386-7808-4

Indic SentiReview: Natural Language Processing based Sentiment Analysis on major Indian Languages

Nidhi Hadiya
Computer Engineering Department
Sarvajanik College of Engineering & Technology(SCET)
Surat, India
nidhihadiya123@gmail.com

Dr. Nirali Nanavati
Computer Engineering Department
Sarvajanik College of Engineering & Technology(SCET)
Surat, India
nirali.nanavati@scet.ac.in

***Abstract*—Sentiment Analysis(SA) can be a natural language processing(NLP) task that extracts opinions from the given text and classifies them as a negative or positive. Research works in SA is mostly conducted in English. Nowadays, the web indexes and other websites related to reviews also support non-English languages. It is therefore necessary to perform SA for these languages as well. There are numerous works found in the literature for SA in other languages worldwide. However, SA for Indian languages needs exploration. In this paper, we discuss various available lexicon resources and often used SA techniques in some Indian languages. Moreover, we present the theoretical parametric evaluation of our studied techniques and we also discuss challenges, which were identified during SA in Indian Languages.**

Keywords—Sentiment Analysis, opinion mining, Indian languages, Natural Language Processing, information retrieval.

I. INTRODUCTION

Accessing social media has now become customary for people and they post their opinions on a particular product or person. People share their feelings in a very informal way. It is therefore a very crucial task to recognize the exact sentiment attached with the text. Therefore, SA is required. The majority of the systems developed for SA so far are only in English and other European languages.

India has 22 official languages as being a multi-lingual country. Because of India's linguistic multiplicity, a broad area has always been open for NLP researchers. With the introduction of UTF-8, the web content in Indian languages is growing enormously. Most of Indians like to share their views on social media in their own religion languages. This increase in the accessibility of extensive Indian languages information has therefore motivated scientists to investigate this field.

There are numerous applications of SA. For example, for a review-related website such as product reviews, film reviews, for detecting heated language in mails, for knowing attitudes and trends of consumers, for knowing public opinions towards the political leaders or political parties [1][2].

The rest of the paper is organized as follows: In section II, we discuss overall process of SA, levels of SA, various pre-processing, feature selection and feature extraction techniques. Section III describes evaluation of Indian languages for SA and available lexicon resources. Section IV describes various techniques used for SA. Section V contains survey of existing approaches for SA in Indian languages. Section VI describes parametric evolution of our studied approaches. Section VII specifies challenges identified during SA in Indian languages. Finally, section VIII contains conclusion and references.

II. THEORITICAL BACKGROUND

Recently, numerous studies have been carried out on SA. Fig. 1 shows overall process of SA starts from pre-processing step and continuous to the sentiment classification through the various available methods including machine learning and lexicon based.

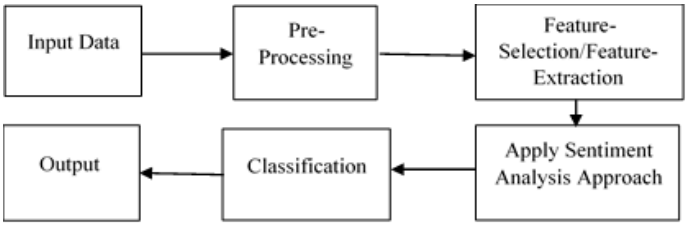

Fig. 1. Overall process of Sentiment Analysis [1]

A. Preprocessing Techniques

Generally used pre-processing techniques are discussed in this section [13][15][18]. In the first step called as a tokenization, the text is often split into smaller tokens. In Stop Word Removal, the words which haven't any impact on the ultimate results of SA are removed. Stemming and Lemmatization are used to trim words while not losing their original significance. Part of speech (POS) tagging is the technique of annotating the text which corresponds to certain parts of the speech like verb, noun, adjective, etc.

B. Feature Extraction and Feature Selection Techniques

Mostly researchers use N-gram as a feature extractor for SA. The n-gram is an adjoining sequence of n words from a text. A n-gram of size 1 is brought up as a "unigram" [4-6] [9][11][13][16-18], a pair could be a "bigram" [6][11][13], size

3 could be a "trigram" [11][13]. Another widely used technique is Bi-tagged, these functions are extracted selectively using fixed patterns based on POS. The most prominent feature selection techniques are Term Frequency(TF), Feature Present(FP) and Term Frequency-Inverse Term Frequency(TF-IDF) [4][8]. TF is the amount of times the document exhibits a feature, FP checks whether a feature appears in the document or not and TF-IDF is used to measure the significance of a word to a document. The worth of the word will increase proportionally to the amount of times the word seems.

C. *Levels of Sentiment Analysis*

SA can be applied on various levels of text, such as document-level [4-6][14][15][18][22], sentence-level [7-13][16-19] and entity-level [23]. In document-level, the document as a whole is classified as a positive, negative or neutral opinion, Sentence-level categorizes individually expressed feelings in each sentence and Entity-level SA aims to classify feelings with regards to the specific features of entities.

III. EVOLUTION OF SA FOR INDIAN LANGUAGES

The advancement of SA in Indian languages began when Joshi et al. [4] initially tried to implement SA in Hindi. They successfully carried out sentimental analysis in Hindi and later began to work on various other languages like Punjabi, Marathi, Tamil, Kannada, Telugu, Punjabi, Bengali, Gujarati and many more.

From extensive survey, we found that SA techniques for non-English languages are generally classified into 2 main categories such as machine translation(MT) and In-language SA [2][3]. MT has been utilized as a tool for making multi-lingual SA frameworks. MT has been employed in these frameworks for translating corpora of various other languages into English. After the translation, sentiment classifiers are built in English. MT is additionally used to create resources and corpora for non-English languages by translating English corpora into that language [4]. By following in-language MT approach, sentiment classifier is built in religion language itself as like we tend to perform in English.

Lexicon Resources available for Indian Languages are as follows.

- **WordNet [24]:** For any language, WordNet is a semantic dictionary. It bunches words into equivalent word sets referred to as synsets, and records the different semantic relationships between these synonym sets [24]. Word-Net is accessible in Bangla, Bodo, Gujarati, Hindi, Malayalam, Punjabi, Tamil, Telugu etc. These WordNets have been created using the expansion approach from Hindi and English WordNet.
- **SentiWordNet [25**]: It is a lexical opinion mining resource. SentiWordNet assigns 3 sentiment scores to each WordNet synset, which are positivity, negativity, objectivity. SentiWordNet is available for Hindi, Tamil, Telugu.

IV. SENTIMENT ANALYSIS APPROACHES

By the literature survey, it has been seen that SA techniques are commonly classified into 3 main categories such as machine learning, lexicon-based and hybrid as appeared in Fig. 2. These techniques are briefly described as follows.

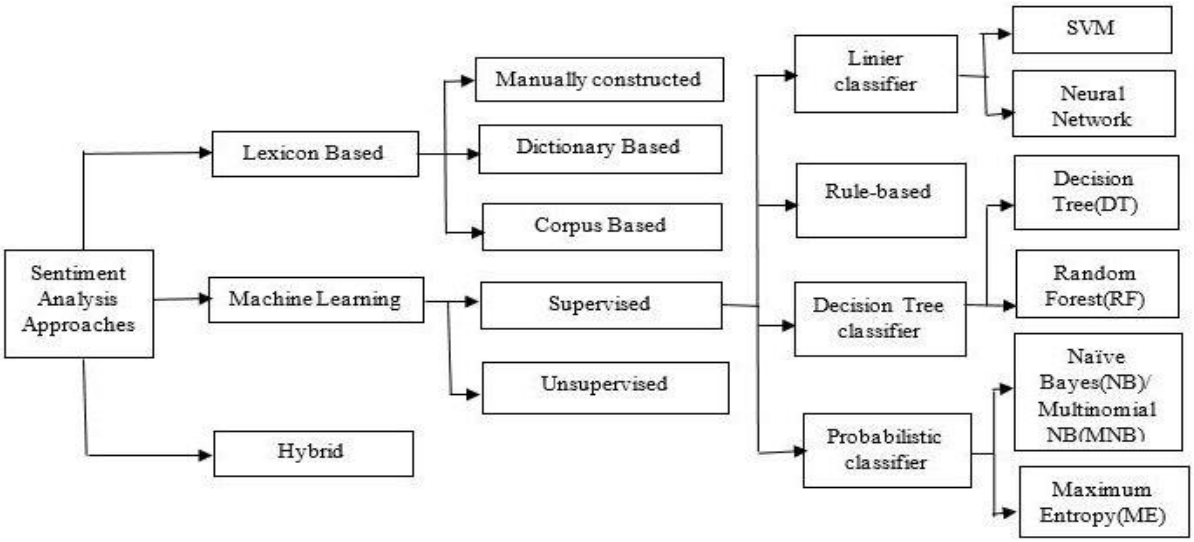

Fig. 2. Approches of Sentiment Analysis

A. *Machine Learning (ML)*

It empowers computers to learn while not being explicitly programmed. ML focuses on developing computer programs that may be instructed to learn when new data are detected [1]. These approaches are additional classifies as supervised or unsupervised. In supervised systems, each input and desire output pairs are provided. Input and output data are tagged for classification to produce a learning basis for future data processing [4][6-8][10-14][17][19][20]. The sub-division of supervised ML techniques is given within the figure 2. In Unsupervised systems, data given to algorithm isn't tagged, that means, only the input variables are given while not a corresponding output [6][9].

B. *Lexicon Based*

These kinds of methodologies use lexicons to determine the polarity of each word and aggregate their scores to calculate the whole polarity of the given text [1]. Lexicon based approaches are additional classified into three sub approaches: manual, corpus based and dictionary based approach. In manual approach, polarities are physically allotted to each sentiment word [15]. Corpus based method uses a seed-list of opinion words whose polarity are already known and then with the assistance of co-occurrence patterns, the new opinion words are determined from a large corpus. In dictionary based method, a seed list of opinion words with their manual polarity is constructed and then with the assistance of synonyms and antonyms of respective words, the seed list is expanded [18].

C. *Hybrid*

In the hybrid approach, the various combination of both the ML and the lexicon based approaches are used to improve the performance of SA [1].

Based on the aim of SA, there are several pros and cons of using these SA strategies. The best thing about the ML approach is that, it having ability to create trained models for distinct purposes, however the limitation is that, it is hard to incorporate into a classifier. Lexicon-based approaches have a main benefit of a more extensive inclusion of general knowledge sentiment lexicons, however these approaches have two main limitations. Firstly, the quantity of words in the lexicons is finite. Secondly, sentiment lexicons tend to assign a fixed sentiment orientation to the words, no matter the context during which these words are employed in a text.

V. SURVEY ON SA WORKS IN MAJOR INDIAN LANGUAGES

This section represents the research work stated in the field of SA for some Indian languages.

Hindi: Initially Joshi et al. [4] tried to work in Hindi for SA. By developing their own lexicon resource HindiSentiWordNet (HSWN), they implemented a fall-back system supported by three approaches, including in-language SA, machine translation and lexicon resource based SA. Mittal et al. [5] planned SA using HSWN on Hindi film reviews and conjointly handled issues like negation and discourse relationships. Jha et al. [6] implemented a system named as HOMS (Hindi opinion mining system) using unsupervised approach and NB classifier. Prasad et al. [7] enforced SA using decision tree under unconstrained and constrained condition. Venugopalan et al. [8] performed SA using SVM and Decision Tree on SAIL-2015. The authors conjointly managed Emoticons, Hashtags and Punctuation. Kumar et al. [10] and Sarkar et al. [11] carried out SA of Hindi tweets on SAIL-2015 using Multinomial NB and SVM. Mumtaz et al. [9] proposed the Senti-lexicon algorithmic program to seek out the polarity of a Hindi movie reviews. The authors handled negation and emoticons as well. Se et al. [12] Implemented SA on Hindi tweets using NB and categorized those tweets such as positive, negative and neutral.

Bengali: Ghosal et al. [13] tested SA by utilizing ML techniques, like SVM, NB, k-NN, RF and DT by using feature extraction techniques such as unigrams, bigrams and trigrams on Bengali horoscope. Kumar et al. [10] and Sarkar et al. [11] carried out SA of Bengali tweets on SAIL-2015 using Multinomial NB and SVM. Se et al. [12] implemented SA for Bengali tweets using NB on SAIL-2015.

Tamil: Se et al. [12] tested SA using NB for Tamil tweets. Se et al. [14] used SentiWordNet to implement SA on film reviews dataset in Tamil by utilizing ML methods like SVM, J48, NB and Maximum entropy.

Konkani: Miranda et al. [15] used Konkani SWN to carry out SA and jointly handled negations and conjunctions to implement an opinion mining system.

Malayalam: Nair et al. [16] implemented a rule-based system to accomplish sentence-level SA on reviews of Malayalam movies. They additionally handled negations and smileys. In their next approach [17], they implemented a hybrid approach by combining ML techniques with rule-based techniques. M.P. et al. [18] proposed a lexicon based document-level SA system for movie reviews.

Punjabi: Kaur et al. [19] performed SA on the basis of a hybrid approach using NB and n-gram on the Punjabi newspaper dataset.

Gujarati: Joshi et al. [20] proposed ML technique for SA on Gujarati tweets. They used POS tagging as feature extractor and SVM as a classifier.

Kanadda: For Kannada documents, Deepamala et al. [22] implemented a lexicon-based method for SA and performed accuracy based comparison of this method with ME and NB. Rohini et al. [23] proposed decision tree based SA on movie reviews in Kannada.

VI. PARAMETRIC EVALUATION

As a shown in table 1, we have analyzed and performed extensive theoretical parametric evaluation of various SA approaches based on our identified parameters such as, which types of classification methods had been used, which types of features were used to carried out SA, on which level SA was performed, which types of issues were handled by that method, which tool was used for SA and lastly how much accuracy they got from their experiment.

Table 1. Theoretical parametric evaluation of existing SA approaches in Indian Languages

Language	Available Lexicon Resource	Ref.	Classification Method	Feature Used	SA Level	Corpus Type	Issues Handled	Tool used	Accuracy
Hindi	Hindi WordNet, Hindi SentiWord-Net	[4]	Machine Translation	TF-IDF	D	Movie Reviews	-	-	65.96%
			Using HSWN	Unigrams				-	60.31%
			ML using SVM	TF-IDF				Rapid Miner	78.14%
		[5]	Using HSWN	Unigrams	D	Movie Reviews	Negation, Discourse Relation	-	80.21%
		[6]	ML using NB	Unigrams, Bigrams	D	Movie Reviews	Negation	NLTK	87.1%
			Unsupervised	POS					
		[7]	C4.5 Decision Tree(J48)	NS	S	SAIL-2015	-	Weka	Constrained: 40.47% Unconstrained: 31.26%
		[8]	SVM & Decision Tree based J48	TF-IDF	S	SAIL-2015	Emoticons, Hashtags, Punctuation	Weka	42.83%(SVM)
		[9]	Lexicon based	Unigrams	S	Movie Reviews	Emoticons, Negation	-	70%
Hindi, Bengali	-	[10]	ML using SVM	N-grams with POS	S	SAIl-2015	-	-	Constrained: 49.68%(Hindi), 43.20%(Bengali) Unconstrained: 46.25%(Hindi), 42%(Bengali)
		[11]	ML using Multinomial NB	Unigrams, Bigrams, Trigrams	S	SAIL-2015	-	Weka	48.82%(Hindi), 40.40%(Bengali)
Hindi, Tamil, Bangali	-	[12]	ML using NB	POS using SentiWord-Net	S	SAIL-2015	-	-	56.67%(Hindi), 39.28%(Tamil), 33.6%(Bengali)
Bengali	Bengali WordNet, Bengali SentiWord-Net	[13]	NB, SVM, K-NN, DT, RF	Unigrams, Bigrams, Trigrams	S	Bengali Horoscope	-	-	98.7% (SVM)
Tamil	Tamil WordNet, Tamil SentiWord-Net	[14]	SVM,DT,NB, ME	POS using SentiWord-Net	D	Movie Reviews	-	-	75.9%(SVM)
Konkani	Konkani WordNet	[15]	Lexicon based	POS using SentiWord-Net	D	NS	Negation, Conjunction	-	NS

Malayalam	Malayalam WordNet	[16]	Rule based, Lexicon based	Unigrams	S	Movie Reviews	Negation	-	85%
		[17]	ML using SVM	Unigrams	S	Movie Reviews	-	-	91%
		[18]	Dictionary Based	Unigrams	S, D	Movie Reviews	-	-	87.5%(S), 90%(D)
Punjabi	Punjabi WordNet	[19]	ML using NB	N-grams	D	Blogs and News Papers	Negation	-	NS
Gujarati	Gujarati WordNet	[20]	ML using SVM	N-grams & POS	S	Normal Tweets	-	-	92%
Kannada	Kannada WordNet	[22]	Dictionary based, ML using ME, NB	NS	D	General Doc.	Negation	-	93% (ME)
		[23]	ML using DT	POS	E	Movie Reviews	-	-	NS

'D' indicates Document, 'S' indicates 'Sentence', 'E' indicated Entity, 'NS' indicates Not-Specified

VII. Challenges identified in Sentiment Analysis for Indian Languages

The area of SA is in developing stage for Indian languages. However, several challenges are identified throughout literature survey. From those, some challenges are outlined in this section.

- Indian languages like Gujarati has been unexplored for the SA domain. Researchers should work on those languages and try to develop SA resources for those languages [15].
- The absence of linguistic resources, tools and corpus adds challenges while dealing with the problem of SA [3].
- One word conveys totally different implications in one languages in several contexts. Human beings can acknowledge it easily. However, the machine finds it exhausting to acknowledge the context within which the word is used [21].
- Most of the time, people type wrong spellings while posting their opinions or comments on social networks. A person can easily understand a spelling error, but a computer can not recognize that word [21].

VIII. Conclusion

The evolution of research within the field of SA for Indian language inspired us to hold this survey. In this paper, we have summarized different SA approaches, various available lexicon resources, challenges and parametric Evaluation of SA in some Indian language. This survey will facilitate researchers to develop an effective SA for their own Indian languages by using various methodology proposed by other researchers that successively can contribute to serve the Indian society.

References

[1] R. Feldman, "Techniques and applications for sentiment analysis," Communications of the ACM, vol. 56, no. 4, p. 82, 2013.

[2] N. Medagoda, S. Shanmuganathan, and J. Whalley, "A comparative analysis of opinion mining and sentiment classification in non-english languages," 2013 International Conference on Advances in ICT for Emerging Regions (ICTer), 2013.

[3] J. Kaur and J. R. Saini, "A Study and Analysis of Opinion Mining Research in Indo-Aryan, Dravidian and Tibeto-Burman Language Families," International Journal of Data Mining And Emerging Technologies, vol. 4, no. 2, p. 53, 2014.

[4] A. Joshi, A. R. Balamurali, P. Bhattacharyya, "A fall-back strategy for sentiment analysis in hindi: a case study," Proceedings of the 8th ICON, 2010.

[5] N. Mittal, B. Agarwal, G. Chouhan, N. Bania, P Pareek, "Sentiment Analysis of Hindi Reviews based on Negation and Discourse Relation," In Proceedings of the 11th Workshop on Asian Language Resources, pp. 45-50, 2013.

[6] V. Jha, N. Manjunath, P. D. Shenoy, K. R. Venugopal, and L. M. Patnaik, "HOMS: Hindi opinion mining system," 2015 IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS), 2015.

[7] S. S. Prasad, J. Kumar, D. K. Prabhakar, and S. Pal, "Sentiment Classification: An Approach for Indian Language Tweets Using Decision Tree," Mining Intelligence and Knowledge Exploration Lecture Notes in Computer Science, pp. 656–663, 2015.

[8] M. Venugopalan and D. Gupta, "Sentiment Classification for Hindi Tweets in a Constrained Environment Augmented Using Tweet Specific Features," Mining Intelligence and Knowledge Exploration Lecture Notes in Computer Science, pp. 664–670, 2015.

[9] D. Mumtaz and B. Ahuja, "Sentiment analysis of movie review data using Senti-lexicon algorithm," 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), 2016.

[10] A. Kumar, S. Kohail, A. Ekbal, and C. Biemann, "IIT-TUDA: System for Sentiment Analysis in Indian Languages Using Lexical Acquisition," Mining Intelligence and Knowledge Exploration Lecture

Notes in Computer Science, pp. 684–693, 2015.

[11] K. Sarkar and S. Chakraborty, "A Sentiment Analysis System for Indian Language Tweets," Mining Intelligence and Knowledge Exploration Lecture Notes in Computer Science, pp. 694–702, 2015.

[12] S. Se, R. Vinayakumar, M. A. Kumar, and K. P. Soman, "AMRITA-CEN@SAIL2015: Sentiment Analysis in Indian Languages," Mining Intelligence and Knowledge Exploration Lecture Notes in Computer Science, pp. 703–710, 2015.

[13] T. Ghosal, S. K. Das, and S. Bhattacharjee, "Sentiment analysis on (Bengali horoscope) corpus," 12th IEEE Int. Conf. Electron. Energy, Environ. Commun. Comput. Control (E3-C3), INDICON 2015, pp. 1–6, 2016.

[14] S. Se, R. Vinayakumar, M. A. Kumar, and K. P. Soman, "Predicting the Sentimental Reviews in Tamil Movie using Machine Learning Algorithms," Indian Journal of Science and Technology, vol. 9, no. 45, 2016.

[15] D. T. Miranda and M. Mascarenhas, "KOP: An opinion mining system in Konkani," 2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), 2016.

[16] D. S. Nair, J. P. Jayan, R. R. R, and E. Sherly, "SentiMa - Sentiment extraction for Malayalam," 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2014.

[17] D. S. Nair, J. P. Jayan, R. R.r, and E. Sherly, "Sentiment Analysis of Malayalam film review using machine learning techniques," 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2015.

[18] M. P. Ashna and A. K. Sunny, "Lexicon based sentiment analysis system for malayalam language," 2017 International Conference on Computing Methodologies and Communication (ICCMC), 2017.

[19] A. Kaur and V. Gupta, "N-gram Based Approach for Opinion Mining of Punjabi Text," Lecture Notes in Computer Science Multi-disciplinary Trends in Artificial Intelligence, pp. 81–88, 2014.

[20] V. C. Joshi, V. M. Vekariya, "An Approach to Sentiment Analysis on Gujarati Tweets," Advances in Computational Sciences and Technology, pp.1487-1493, 2017.

[21] P. Arora, "Sentiment Analysis For Hindi Language, MS by Research in Computer Science," Master Thesis, IIT Hyderabad, 2013.

[22] D. N. and R. K. P., "Polarity detection of Kannada documents," 2015 IEEE International Advance Computing Conference (IACC), 2015.

[23] V. Rohini, M. Thomas, and C. A. Latha, "Domain based sentiment analysis in regional Language-Kannada using machine learning algorithm," 2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), 2016.

[24] G. A. Miller, "WordNet: a lexical database for English," Communications of the ACM, 1995.

[25] A. Esuli, F. Sebastiani, "SentiWordNet: a high-coverage lexical resource for opinion mining," 2007.