A Sentiment Analysis of Gujarati Text using Gujarati Senti word Net

Lata Gohil, Dharmendra Patel

Abstract*: *Sentiment Analysis plays vital role in decision making. For English language intensive research work is done in this area. Very less work is reported in this domain for Indian languages compared to English language. Gujarati language is almost unexplored for this task. More data in form of movie reviews, product reviews, social media posts etc are available in regional languages as people like to use their native language on Internet which leads to need of mining these data in order to understand their opinion. Various tools and resources are developed for English language and few for Indian languages. Gujarati is resource poor language for this task. Motive of this paper is to develop sentiment lexical resource for Gujarati language which can be used for sentiment analysis of Gujarati text. Hindi SentiWordNet (H-SWN) [1] ***and synonym relations of words from IndoWordnet (IWN)*** [2] [3] ***are used for developing Gujarati SentiWordNet. Our contribution is twofold. (1) Gujarati SentiWordNet (G-SWN) is developed. (2) Gujarati corpus is prepared in order to evaluate lexical resource created. Evaluation result shows the usefulness of generated resource.***

Index Terms*: *Gujarati SentiWordNet, SentiWordNet, Gujarati Corpus, Sentiment Analysis, Opinion Mining

I. INTRODUCTION

In today's era of Internet, multi-modal content increases on Internet at fast pace. Text content is written in various languages. Mining this text is important task to extract knowledge. Sentiment Analysis is one such task which computationally identifies the opinion from unstructured or semi-structured text. Several sentiment lexical resources are developed for English language such as Taboada's adjective list [4], SentiWordNet [5], Subjectivity Word List [6]. Sentiment lexicon stores opinionated terms and their part of speech tag and polarity classes along with scores.

This paper proposes technique for construction of Gujarati SentiWordNet. Two resources are developed: Gujarati SentiWordNet (G-SWN) and Gujarati Corpus.

The rest content of this paper is as follows: Related work is described in Section 2. Motivation of this research work is presented in Section 3. Methodology applied for creation of Gujarati SentiWordNet is discussed in Section 4. Section 5 explains corpus generation followed by evaluation of resource created. A possible future work suggestions for extendable version of the Gujarati SentiWordNet (G-SWN) is mentioned in Section 6.

Revised Manuscript Received on July 05, 2019.

Lata Gohil, Computer Science and Engineering Department, Institute of Technology, Nirma University, India

Smt. Chandaben Mohanbhai Patel Institute of Computer Applications (CMPICA), CHARUSAT, Changa, India

Dharmendra Patel, Smt. Chandaben Mohanbhai Patel Institute of Computer Applications (CMPICA), CHARUSAT, Changa, India

II. RELATED WORK

Sentiment analysis is useful method to understand opinion expressed in text. "It is one of the most active research areas in natural language processing and is also widely studied in data mining, web mining, and information retrieval" [7]. Beginning research work on sentiment analysis was mostly focused on English language. However increasing non-English language content on Internet led demand to work towards other languages. There are two approaches namely Lexicon and Machine Learning are widely explored for sentiment analysis. Large amount of annotated data is the key requirement of Machine learning approach. Resource-poor languages lack such annotated corpora and therefore lexicon approach is suitable for them to start with. Various approaches have been explored to prepare sentiment lexical resources such as dictionary based, wordnet based, corpus based. Though lexical resources contain out of context polarity scores of terms, they have been proven to give good baseline.

Several sentiment lexical resources are developed for English language. SentiWordNet [5] contains over 3 million terms along with positive, negative and objective scores. Subjectivity Lexicon [6], which is a part of OpinionFinder, comprise of word, POS tag, polarity and subjectivity as either strongly or weakly subjective. Opinion Lexicon [8] is prepared by extracting adjectives from opinion sentences of annotated twitter corpora. AFINN-111 [9] contains words which are manually rated for valence.

Sentiment lexical resources are also developed for majority of Indian languages. Hindi-SentiWordNet (H-SWN) [1] is built by exploiting English SentiWordNet [5] and English-Hindi WordNet Linking [10]. Hindi Subjective Lexicon [11] have been generated using initial seed words whose synonyms and antonym relations are taken from Hindi WordNet using graph based expansion method. [12] have prepared Bangla SentiWordNet through English-Bengali bilingual dictionary and SentiWordNet [5]. SentiWordNet(s) for Indian languages [12] have been developed using wordnet, antonym generation, corpus and game based approaches. [13] have developed PsychoSentiWordNet by leveraging human psychological knowledge through involvement of Internet population. Odia lexicon resource [14] is prepared using SentiWordNet for Indian languages [12] and IndoWordNet [2] by taking word from SentiWordNet for Indian languages and mapping its corresponding Synset ID from Odia wordnet whose POS tag is either adjective or adverb. [14]

have used translation approach for building SentiWordNet for Tamil language. Four English language resources named English SentiWordNet 3.0 [5], Subjectivity Lexicon [6], AFINN-111 [9] and Opinion Lexicon [8] are used for the purpose of increasing reliability of lexicon. Urdu lexicon resource [15] is built using word level translation approach. Lexicons acquisition done from English opinion words, English SentiWordNet (SWN), Urdu modifiers and English to Urdu bilingual dictionary. Three different methods namely SentiWordNet-based, manual-driven and corpus-based are used for calculating polarity scores.

For development of lexical tools, resource-poor languages usually make use of existing resources of other languages. In this study, we have used Hindi SentiWordNet (H-SWN) and IndoWordNet (IWN) for building Gujarati SentiWordNet (G-SWN).

III. MOTIVATION

Extensive research endeavor is reported for Sentiment Analysis in English and other international languages and ample amount of resources and tools are developed and available for use. Regional languages are less explored for this task and thus they are resource-poor languages.

People incline to use their regional language on Social Media and other platforms on Internet. "Gujarati is an Indo-Aryan family language evolved from Sanskrit language and it is one of the twenty-two official languages and fourteen regional languages of India" [16]. There are very large Gujarati immigrant communities not only in other parts of India but in world also.

Large amount of user generated content on Internet is available in Gujarati language and thus there is need to mine this content for extracting useful information. Gujarati is resource-poor language. This motivate us to develop sentiment lexical resource for Gujarati language.

IV. METHODOLOGY

Sentiment lexical resource consists of prior polarity of terms. This polarity is out of context. However for resource poor language, lexical resource is good option to start with. Our proposed lexicon resource is generated using Hindi SentiWordNet (H-SWN) [1] and IndoWordnet (IWN) [2][3]. We named this resource Gujarati SentiWordNet (G-SWN).

H-SWN is automatically generated sentiment lexicon resource for Hindi language by exploiting two resources English-Hindi WordNet linking [10] and English SentiWordNet [5]. Each synset of H-SWN is recorded with three polarity scores named positive, negative and objective whose values lie between 0 and 1 and summation of this three values is 1. For example, synset "बंधन" is recorded as: a 6268 0.375 0.125 बंधन,बद्धी,अनुबंध,अनुबन्ध,बन्धन,अनुबन्ध,अलान,आलान. The scores 0.375, 0.125 and 0.50 (i.e. 1 - 0.375 - 0.125) are positive, negative and objective scores respectively.

"IndoWordnet (IWN) is a connected structure of wordnets of major Indian languages from Indo-Aryan, Dravidian and Sino-Tibetan families" [2]. Hindi wordnet is the source wordnet from which other wordnets in IWN have been prepared using expansion approach. Identification number of each synset has been preserved across languages which helps to access synset from different languages with ease. Java-based API [3] and Python-base API [17] are available for IndoWordNet.

Assumption for preparing G-SWN is made that synset preserves sentiment across languages. Therefore, polarity scores of term in H-SWN is projected to corresponding synset of IWN. Synonym relations of synsets of IWN are used for generating G-SWN.

The steps followed for building the G-SWN are given below:

1) Repeat step 2 to 4 for each word (*w*) of H-SWN
2) Map *w* to synset (s-id) of HindiWordNet (HWN)
3) Synset of each lemma of synset (s-id) is taken from GujaratiWordNet (GWN) to prepare new lemma list of synset (s-id)
4) Polarity score of synset (s-id) is projected from H-SWN

The process of building G-SWN is illustrated in Fig. 1.

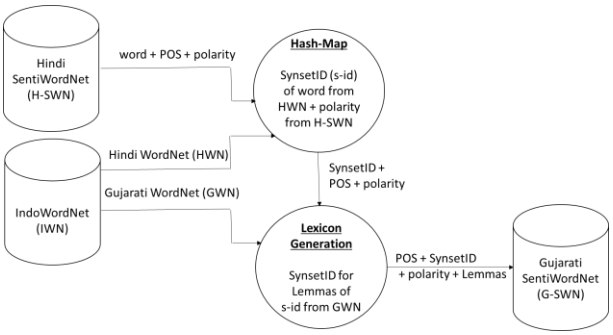

Fig. 1 : Architecture diagram of G-SWN

The resultant G-SWN consists of synsets along with polarity scores and part-of-speech tag. The distribution statistics of synsets as per part-of-speech categories in G-SWN is given in Table I.

Table I : Synset Statistics of G-SWN

POS Category	No. of Synsets
Adjective	1,828
Verb	840
Adverb	89
Noun	3,319
Total No. of Synsets	**6,076**

V. EXPERIMENT

The generated lexicon resource needs to be evaluate for accessing its reliability. We employed unigram presence and simple scoring methods for evaluation of G-SWN. Unavailability of gold standard corpus in Gujarati language leads to prepare it for evaluating G-SWN.

A. Corpus

The construction of Gold Standard Corpora is time and labor intensive task. [18] have suggested alternative mechanisms for preparing annotated corpora such as crowd-sourcing, expert review and active selection schemes which reduce cost in terms of labor and time at the same time retain high quality of

annotation of corpora. However we have followed manual annotation approach.

We collected Twitter data using the public streaming Twitter API by supplying query keywords. After removal of duplication, tweets were annotated for positive and negative polarity tags by two annotators. The Kappa coefficient, statistical measure of inter-rater agreement, is 0.55 for the annotated tweets which shows moderate agreement. Table II presents inter-rater agreement statistics for annotated tweets.

Table II : Inter-rater Agreement Statistics

	Positive	Negative	Total
Positive	442	57	**499**
Negative	200	421	**621**
Total	**642**	**478**	**1120**

B. Result

To evaluate the G-SWN, two classification methods unigram presence and simple scoring are applied on tweet corpus discussed in section V.A ***Corpus***. Unigram presence method counts unigrams of positive and negative polarity and the polarity possess higher count is considered as resultant polarity. While simple scoring method make summation of polar scores of each unigram of tweet present in G-SWN and the polarity with higher value is considered as resultant polarity. With unigram presence method and simple scoring method accuracy achieved is 52.72% and 52.95% respectively. Table III reports the performance of both mentioned classifiers using G-SWN.

Table III : Performance of classifiers using G-SWN

Classifier	Precision	Recall	F-measure	Accuracy
Unigram Presence	52.77	52.72	49.84	52.72
Simple Scoring	53.07	52.95	50.09	52.95

Result of classifiers indicates that the polarity scores of the G-SWN are moderately reliable and G-SWN could be used as baseline.

VI. CONCLUSION AND FUTURE WORK

We proposed method to generate G-SWN using Hindi SentiWordNet (H-SWN) and IndoWordNet (IWN) by exploiting synonym relations. The generated G-SWN resource will be useful for sentiment analysis of Gujarati text. The proposed method can be applied to generate sentiment lexical resources for all languages included in IWN. The Gujarati tweets corpus is developed for evaluation of the generated lexical resource. Corpus was annotated for positive and negative polarity classes by two annotators. Statistical measure inter-annotator agreement Cohen's kappa achieved for this corpus is 0.55. Resultant annotated corpus comprises of 863 tweets out of which 442 are positive tweets and 421 are negative tweets. Evaluation of G-SWN using this gold standard corpora achieved 52.72% and 52.95% accuracy for unigram presence and simple scoring classifiers respectively. Result shows moderate performance of G-SWN. G-SWN provides baseline for further study. In future, this proposed work can be extended by making use of antonym relations. By incorporating Word Sense Disambiguation (WSD) better accuracy can be achieved.

REFERENCES

1. A. Joshi, A. R. Balamurali, and P. Bhattacharyya, "A fall-back strategy for sentiment analysis in hindi: a case study," *Proc. 8th ICON*, 2010.
2. P. Bhattacharyya, "IndoWordNet," in *In Proc. of LREC-10*, 2010.
3. N. R. Prabhugaonkar, A. Nagvenkar, and R. Karmali, "IndoWordNet Application Programming Interfaces," 2012.
4. K. Voll and M. Taboada, "Not all words are created equal: Extracting semantic orientation as a function of adjective relevance," in *Australasian Joint Conference on Artificial Intelligence*, 2007, pp. 337–346.
5. A. Esuli and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining.," in *LREC*, 2006, vol. 6, pp. 417–422.
6. T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 2005.
7. B. Liu, "Sentiment analysis and opinion mining," *Synth. Lect. Hum. Lang. Technol.*, vol. 5, no. 1, pp. 1–167, 2012.
8. M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004, pp. 168–177.
9. F. Å. Nielsen, "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs," *arXiv Prepr. arXiv1103.2903*, 2011.
10. A. K. Karra, "WordNet Linking." Master of Technology Dissertation, CSE Department, IIT Bombay, 2010.
11. A. Bakliwal, P. Arora, and V. Varma, "Hindi subjective lexicon: A lexical resource for hindi polarity classification," in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC)*, 2012, pp. 1189–1196.
12. A. Das and S. Bandyopadhyay, "Sentiwordnet for bangla," *Knowl. Shar. Event-4 Task*, vol. 2, pp. 1–8, 2010.
13. A. Das and S. Bandyopadhyay, "Dr Sentiment knows everything!," in *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies: systems demonstrations*, 2011, pp. 50–55.
14. G. Mohanty, A. Kannan, and R. Mamidi, "Building a sentiwordnet for odia," in *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2017, pp. 143–148.
15. M. Z. Asghar, A. Sattar, A. Khan, A. Ali, F. Masud Kundi, and S. Ahmad, "Creating sentiment lexicon for sentiment analysis in Urdu: The case of a resource-poor language," *Expert Syst.*, p. e12397, 2019.
16. "https://en.wikipedia.org/wiki/Gujarati_language." Last accessed 15 June 2019.
17. R. Panjwani, D. Kanojia, and P. Bhattacharyya, "pyiwn: A Python-based API to access Indian Language WordNets," in *Proceedings of the 9th Global WordNet Conference (GWC 2018)*, 2018, p. 382.
18. L. Wissler, M. Almashraee, D. M. D\'\iaz, and A. Paschke, "The Gold Standard in Corpus Annotation.," in *IEEE GSC*, 2014.

AUTHORS PROFILE

Lata Gohil is Assistant Professor in Computer Science and Engineering Department, Institute of Technology, Nirma University. She has received MCA degree from Gujarat Vidyapith, Ahmedabad. She has qualified GSET and GATE. Her research area is Information Retrieval and Text Mining. She is pursing PhD from CHARUSAT.

Dr Dhrmendra Patel received his Master of Computer Application degree from North Gujarat University. He received his PhD degree in computer science from Kadi Sarva Viswavidyalaya. His area of research is Web Mining, Fog Computing, Image Processing, Interenet of Things etc. He has published 20 papers in national/international journal of repute. Currently he is working as an associate professor at CHARUSAT, Changa.