Building a SentiWordNet For Odia

Gaurav Mohanty and **Abishek Kannan** and **Radhika Mamidi**
Language Technologies Research Center
Kohli Center on Intelligent Systems
International Institute of Information Technology, Hyderabad
{gaurav.mohanty,abishek.kannan}@research.iiit.ac.in
radhika.mamidi@iiit.ac.in

Abstract

As a discipline of Natural Language Processing, Sentiment Analysis is used to extract and analyze subjective information present in natural language data. The task of Sentiment Analysis has acquired wide commercial uses including social media monitoring tasks, survey responses, review systems, etc. Languages like English have several resources which aid in the task of Sentiment Analysis. SentiWordNet and Subjectivity WordList are examples of such tools and resources. With more data being available in native vernacular, language-specific SentiWordNet(s) have become essential. For resource poor languages, creating such SentiWordNet(s) is a difficult task to achieve. One solution is to use available resources in English and translate the final source lexicon to target lexicon via machine translation. Machine translation systems for the English-Odia language pair have not yet been developed. In this paper, we discuss a method to create a SentiWordNet for Odia, which is resource-poor, by only using resources which are currently available for Indian languages. The lexicon created, would serve as a tool for Sentiment Analysis related task specific to Odia data.

1 Introduction

For resource-poor languages, one popular approach is to use readily available resources in English to generate a source lexicon. The source lexicon is then translated using a Machine Translation system or a bilingual dictionary to create the final target lexicon (Bakliwal et al., 2012). In case of the English-Odia language pair, a good Machine Translation system is absent. The online bilingual dictionaries for the same have very few word pairs. Manual translation is expensive in terms of human resource and time. Another approach is to use available parallel corpora for the language pair and use a word-alignment tool in order to get a one-to-one mapping between words. For this method, a sufficiently large corpus is required in order to get an appropriate number of unique word pairs. Such a large corpus is unavailable for the English-Odia language pair. In fact, larger corpora is available for Odia and other Indian language pairs. WordNets developed under the IndoWordNet structure (Bhattacharyya, 2010) do not map words directly but they match synsets instead. These WordNets for Indian languages serve well in translation from source to target lexicon. The SentiWordNets present for such Indian languages helps in assignment of polarity to the final collection of words.

Odia SentiWordNet is built using WordNets and SentiWordNets available for other Indian languages. WordNets include those of Bengali, Tamil, Telugu and Odia itself. SentiWordNets used include those of Bengali, Tamil and Telugu.

The paper is divided into various sections. Section 2 comprises of previous work and progress towards building SentiWordNets for Indian languages. Section 3 describes resources used for creation of Odia SentiWordNet. Section 4 contains a detailed explanation of procedure followed for the same and defines the evaluation scheme for verification of resource thus created. An insight on future work and extensibility of the SentiWordNet is provided in Section 5.

2 Previous Work

Since its introduction in 1961 by IBM, Sentiment Analysis has been a fast growing area in computer

science. Research on Sentiment Analysis began in English. However with increasing demand, several researchers have developed various tools and resources for many other languages. Odia (ISO 639 language code: ori)[1], being a resource-poor language, lacks necessary tools to perform Sentiment Analysis.

Since opinion mining has proved extremely useful in online review and survey systems and since data is more readily available than ever, Sentiment Analysis serves as an effective method to achieve automated scoring of products, movies, etc.

Turney worked on classifying customer reviews (Turney, 2002). They adopt an unsupervised learning technique to predict the semantic orientation of phrases. Hatzivassiloglou (Hatzivassiloglou and R. McKeown, 1997) and Turney (Turney and Littman, 2003) describe methods of using a set of words gathered a priori as a seed list to classify the semantic orientation of phrases. The former method (Hatzivassiloglou and R. McKeown, 1997) was the first to deal with opinion classification in phrases. The approach mainly uses adjectives for Sentiment Analysis. However, sufficient pre-processing was carried out using available tools for English before the phrases were successfully classified.

Even though sentiment depends on context, lexical resources have proven to give a good baseline for further studies. The English language has several lexical resources such as the SentiWordNet as described by Esuli (Esuli and Sebastiani, 2006). It contains over 3 million tokens assigned with polarity and objectivity score. The resource has been improved over the years as demonstrated in literature (Baccianella et al., 2010). Another such important resource is the Subjectivity Lexicon (Wilson et al., 2005) which is a part of OpinionFinder[2].

Languages which have a scarcity of readily available data depend on resource rich languages to build such lexicons. Whalley (Whalley and Medagoda, 2015) describes how the Sinhalese sentiment lexicon was created using the English SentiWordNet 3.0. The SentiWordNet in English was mapped to a Sinhalese dictionary and the scores were copied from one language to another. Another way to achieve this is by linking the WordNets of the source and target language. Joshi proposed a method to create a SentiWordNet for Hindi by linking the English and Hindi WordNets and assigning scores to the synsets in Hindi WordNet (Joshi et al., 2010). Dipankar Das suggested a method to develop WordNet affect lists in Bengali using affect wordlists already available in English. (Das and B, 2010). The method uses a bilingual dictionary to translate words from English to Bengali. Amitava Das (Das and Bandyopadhyay, 2010) (Das and Gambäck, 2012) (Das and Bandyopadhyay, 2011) proposes several ways to generate such lexical resources for other Indian languages. One approach suggests the usage of both English SentiWordNet 3.0 and Subjectivity Lexicon and adopting a translation based approach in order to build the lexicon in three Indian languages (Das and Bandyopadhyay, 2010). A SentiWordNet for Tamil has also been developed using a similar translation based approach for currently available resources in English (Kannan et al., 2016). Due to lack of a sufficiently large parallel corpus or a bilingual dictionary, direct translation techniques from English to Odia could not be applied in-order to build the SentiWordNet in Odia.

3 Prerequisites

For creating Odia SentiWordNet, SentiWordNets of three Indian languages, namely Bengali, Tamil and Telugu are used. Polarity of words for these resources has proved to be reliable (Das and Bandyopadhyay, 2010). Multiple SentiWordNets are used for a better estimate of sentiment for each word and reduction of ambiguities while building the resource. For creation of lexicon for Odia, WordNets for Odia and the other three Indian languages are used. These WordNets have synsets linked via a common synset identification number (ID), without direct word-to-word mapping. The resources used are described below.

3.1 SentiWordNets for Indian Languages

SentiWordNet is a lexical resource explicitly devised for supporting sentiment classification and opinion mining applications. According to Baccianella, SentiWordNet is the result of the automatic annotation of all the synsets of WordNet towards the notions of positivity, negativity, and neutrality (Baccianella et al., 2010). Each synset is associated with three numerical scores : pos(s), neg(s), and obj(s) which indicate positive, negative, and objective i.e., neutral respectively. Senti-

[1] http://www-01.sil.org/iso639-3/codes.asp
[2] http://mpqa.cs.pitt.edu/opinionfinder/

WordNets for Bengali, Telugu and Tamil were created using Das's approach (Das and Bandyopadhyay, 2010). Each of these comprises of four lists under the categories of "Positive", "Negative", "Neutral" and "Ambiguous" which contain words of positive, negative, neutral polarity and ambiguous words, respectively. The Parts-of-Speech tag information for each word is also provided. Table 1 gives detailed statistics for each of the SentiWordNets used.

Language	POS	NEG	NEU	AMB
Bengali	1779	3714	359	648
Telugu	2136	4076	359	1093
Tamil	2225	4447	361	1168

Table 1: Statistics for SentiWordNets

3.2 WordNets for Indian Languages

"Wordnets are lexical structures composed of synsets and semantic relations" (Fellbaum, 1998). A synset comprises a set of synonyms. They are linked by semantic relations like hypernymy (is-a), meronymy (part-of), troponymy (manner-of), etc. WordNets for four different languages are used for building the lexicon for Odia SentiWordNet. These WordNets are linked across languages through common synset IDs. They are part of the linked IndoWordNet structure (Bhattacharyya, 2010). WordNets for Bengali, Tamil and Telugu were used for creating the source lexicon. Odia WordNet was used for generating the target lexicon. Table 2 describes the statistics of the number of tokens present in every Part-Of-Speech category for each language.

LANG	Odia	Bengali	Telugu	Tamil
NOUN	27216	27281	12078	16312
VERB	2418	2804	2795	2803
ADJ	5273	5815	5776	5827
RB	377	445	442	477
Total	35284	36346	21091	25419

Table 2: IndoWordNet Statistics.

4 Procedure

A step-by-step procedure to be followed is illustrated in Figure 1. This procedure can be adopted for a different target language, as long as the target language has a WordNet which is linked with other Indian language WordNets. The procedure is divided into three parts:

1. **Creating Source Lexicon:** SentiWordNets from Indian languages are used to assign a polarity to corresponding WordNet synsets. A final list of synsets IDs with the corresponding polarity serves as a source list.

2. **Generating Target Lexicon:** For every synset ID from source, the corresponding words from the target language WordNet are assigned the same polarity as that of the synset ID.

3. **Evaluation of Final Resource:** The created target lexicon needs to be evaluated for errors. This paper adopts manual evaluation by language specific annotators and reports annotator agreement score.

4.1 Creating Source Lexicon

Source Lexicon acquisition begins with SentiWordNets available for the three aforementioned Indian languages. In order to create a reliable baseline for Odia, only words with positive and negative polarity are considered. Currently, ambiguous words or those having neutral polarity are not considered for the creation of source lexicon. For each language, words with positive and negative polarity are extracted from their corresponding SentiWordNets.

The corresponding synset ID of each word is then found from that language's WordNet. This is attained by using a hash-map created over all the words in WordNet for that language. The synset ID for the identified word serves as a key to a dictionary δ. The corresponding value is a list with the polarity of the word as an item. In case δ already has a synset ID as a key, the polarity of the word is appended to the existing list for that key in δ. Such a case would occur when word and its synonym (both part of the same synset) are both present in the SentiWordNet for that language. Such a case can also occur when a word from a different language's WordNet has a synset ID which is already a key in δ. The final dictionary comprises of several synset IDs as key. A total of 6203 synset IDs were identified. For each key the value in δ is a list of polarities (positive or negative) which are observed for words in the sysnet across languages.

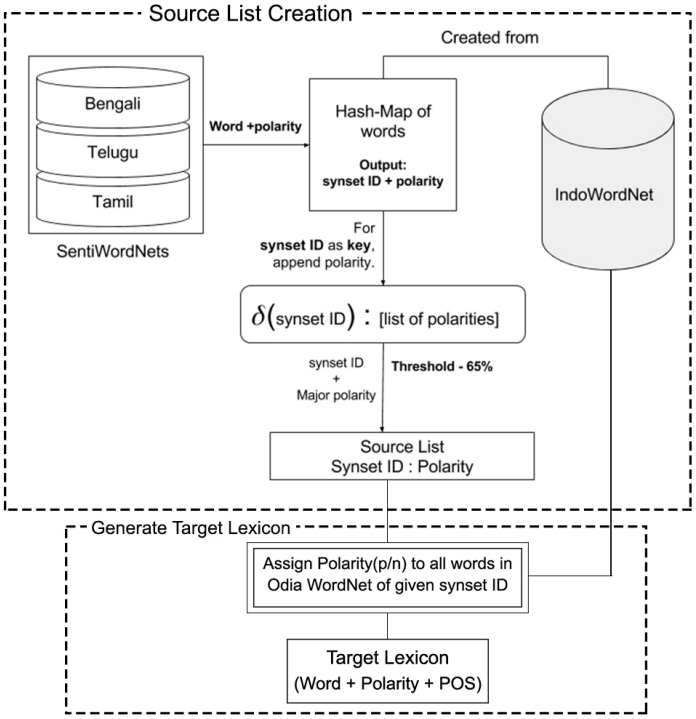

Figure 1: Flow of Design for Odia SentiWordNet

A word and its synonyms should commonly have the same polarity. This should also be true across languages, in an ideal scenario. However, it was observed that in many cases the list of polarities for a given key is not homogeneous. This is because a word with a particular sentiment in one language may not necessarily have the same sentiment in another language. Infact, it was also observed that, in very few cases, a word with a given sentiment in one language sometimes did not have the same sentiment for some of its synonyms in the same language. Every synset ID which exists as a key in δ is to be assigned a single polarity. Any of the synset IDs which have contradicting polarities to a certain degree should be ignored as these will affect the reliability of the Odia SentiWordNet. For a given key (synset ID), the polarity in its list in δ which holds a majority greater than 65% is considered the final polarity for that synset ID. This results in a list of synset IDs with an assigned major polarity. The list serves as the "Source" to map to synsets in Odia WordNet. The source list comprises of 5661 synset IDs along with their major polarity.

4.2 Generating Target Lexicon

In order to create the Target Lexicon, Odia WordNet is used. The Odia WordNet is linked to the other three aforementioned WordNets through a common synset ID. A total 5407 synset IDs from the Source List were found to exist in Odia WordNet. For each synset ID in Source list, the corresponding words are extracted from Odia WordNet. Each of these words is assigned the major polarity (positive or negative) corresponding to that synset ID in the Source list. A total of 13917 tokens were assigned polarity. Table 3 provides details on the total tokens extracted from Odia WordNet. Only adjectives and adverbs are added to the final Target Lexicon. Nouns and verbs were not added to the Target Lexicon because the polarity associated with these words is usually context dependent. These are added to a separate list for future inspection.

No. of observed OWN synsets	5661
Adjectives and Adverbs	4747
Nouns and Verbs	9170
Total number of tokens	13917

Table 3: Target Lexicon Statistics

The final Target Lexicon comprises of words along with their sentiment polarity, Part-of-Speech tag and synset ID corresponding to the language's WordNet. The final lexicon contained 1839 pos-

Word	Meaning	Polarity
ଭାଗ୍ୟଶାଳୀ	fortunate	Positive
ସତ୍ଚୋଟପଣିଆ	truthful, honest	Positive
ଆଲୋକିତ	enlightened	Positive
ସୀମିତ	limited	Negative
ଠିକ୍	correct	Positive
ଏକା	alone, deserted	Negative
ଲୋଭୀ	selfish	Negative

Figure 2: Odia words with polarity

itive entries and 2908 negative entries. Figure 2 shows a few examples of Odia words with their corresponding assigned polarity.

4.3 Resource Evaluation

In order to assess the reliability of the Odia SentiWordNet, a random sampling of 2500 words was created from the Target Lexicon. In order to maintain a balanced sample set, 1250 words were randomly picked from each polarity list. This sample set was provided to three manual annotators to be independently annotated as positive or negative. The manual annotators were native Odia speakers and spoke the language on a daily basis. Each of the three annotators were asked to annotate every token of the sample set with the polarity they deemed appropriate. No annotator had prior information about the assigned polarity to a token. This ensured unbiased annotation of tokens.

In order to capture inter-annotator agreement, Fleiss Kappa[3] score for the annotated sample set was also calculated. Fleiss Kappa is calculated using the following formula:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \quad (1)$$

$\bar{P}$ represents the sum of observed agreement. The sum of agreement by chance is denoted by $\bar{P}_e$. Fleiss Kappa score is calculated using three raters for two categories (positive/negative). A substantial agreement score of κ = **0.76** is reported for Odia SentiWordNet.

[3]https://en.wikipedia.org/wiki/Fleiss'˙kappa

In order to further improve upon the Target Lexicon, words with sentiment which none of the annotators agreed to, were removed. This was done only for the sample of 2500 words. Table 4 gives metrics for the Odia SentiWordNet thus created. A total of 98 words with incorrect polarity were removed.

Initial Positive Tokens	1839
Initial Negative Tokens	2908
Final Positive Tokens	1803
Final Negative Tokens	2846
Inter-Annotator Agreement (**Fleiss Kappa**)	
0.76	

Table 4: Evaluation Details

5 Conclusion and Future Work

Odia SentiWordNet will serve as a useful resource for Sentiment Analysis on Odia data. The method adopted is generic and can be used to create similar sentiment lexicons for other Indian languages which are part of the IndoWordNet structure. In order to find the accuracy of the created resource, it needs to be tested on actual user generated data. Odia data is readily available online. Currently, a set of 1000 Odia sentences is being manually annotated. The annotated set would serve as gold data. These sentences are taken from online newspaper articles[4]. Odia SentiWordNet will be tested on these 1000 sentences in order to predict the sentiment associated with each sentence. Comparison with results of manual annotation should give a more accurate insight on how reliable the resource is. The resource serves as a baseline and can be improved in the future. Several resource expansion strategies can be used to enrich Odia SentiWordNet. One particular method involves usage of antonym relations. Antonyms of a word, which are not already present in the resource can be assigned opposite polarity. Antonym creation rules, specific to the language, can be applied to generate antonyms of many words in the resource as suggested previously in literature (Das and Bandyopadhyay, 2010). If a sufficiently large corpus becomes available, SentiWordNet can be used to capture language-specific nuances. The raw corpus can be trained on a word embedding tool (e.g Word2Vec) to create word clusters of similar

[4]http://thesamaja.in/

words based on the prior and subsequent neighbours of a word in the corpus. Such clusters can be further used to expand the lexicon.

Acknowledgments

The authors would like to thank Pruthwik Mishra, Shastri V. Mohapatra and Ranjita Mohanty for their help in manual annotation and checking the reliability of Odia SentiWordNet. The SentiWordNets for Tamil, Bengali and Telugu were acquired from Amitava Das' website[5]. The IndoWordNet was accessed from CLIFT IIT Bombay website[6].

References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. European Languages Resources Association (ELRA). http://aclweb.org/anthology/L10-1531.

Akshat Bakliwal, Piyush Arora, and Vasudeva Varma. 2012. Hindi subjective lexicon: A lexical resource for hindi adjective polarity classification. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *LREC*. European Language Resources Association (ELRA), pages 1189–1196. http://dblp.uni-trier.de/db/conf/lrec/lrec2012.htmlBakliwalAV12.

Pushpak Bhattacharyya. 2010. Indowordnet. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *LREC*. European Language Resources Association. http://dblp.uni-trier.de/db/conf/lrec/lrec2010.htmlBhattacharyya10.

Amitava Das and Sivaji Bandyopadhyay. 2010. Sentiwordnet for indian languages. In *Proceedings of the Eighth Workshop on Asian Language Resouces*. Coling 2010 Organizing Committee, Beijing, China, pages 56–63. http://www.aclweb.org/anthology/W10-3208.

Amitava Das and Sivaji Bandyopadhyay. 2011. Dr sentiment knows everything! In *Proceedings of the ACL-HLT 2011 System Demonstrations*. Association for Computational Linguistics, Portland, Oregon, pages 50–55. http://www.aclweb.org/anthology/P11-4009.

Amitava Das and Björn Gambäck. 2012. Sentimantics: Conceptual spaces for lexical sentiment polarity representation with contextuality. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*. Association for Computational Linguistics, Stroudsburg, PA, USA, WASSA '12, pages 38–46. http://aclweb.org/anthology/W12-3707.

Dipankar Das and Sivaji B. 2010. Developing bengali wordnet affect for analyzing emotion.

A. Esuli and F. Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. European Language Resources Association (ELRA). http://aclweb.org/anthology/L06-1225.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press, Cambridge, MA.

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *8th Conference of the European Chapter of the Association for Computational Linguistics*. http://aclweb.org/anthology/E97-1023.

Aditya Joshi, AR Balamurali, and Pushpak Bhattacharyya. 2010. A fall-back strategy for sentiment analysis in hindi: a case study. *Proceedings of the 8th ICON* .

Abishek Kannan, Gaurav Mohanty, and Radhika Mamidi. 2016. Towards building a sentiwordnet for tamil. In *Proceedings of the 13th International Conference on Natural Language Processing*. NLP Association of India, Varanasi, India, pages 30–35. http://www.aclweb.org/anthology/W16-6305.

Peter D. Turney. 2002. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '02, pages 417–424. https://doi.org/10.3115/1073083.1073153.

Peter D. Turney and Michael L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Trans. Inf. Syst.* 21(4):315–346. https://doi.org/10.1145/944012.944013.

J Whalley and N Medagoda. 2015. Sentiment lexicon construction using sentiwordnet 3.0. *ICNC'15 - FSKD'15, School of Information Science and Engineering, Hunan University, China* .

Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005. Opinionfinder: A system for subjectivity analysis. In *Proceedings of HLT/EMNLP 2005 Interactive Demonstrations*. http://aclweb.org/anthology/H05-2018.

[5] http://amitavadas.com/sentiwordnet.php

[6] http://www.cfilt.iitb.ac.in/indowordnet/