

Dr Sentiment Creates SentiWordNet(s) for Indian Languages Involving Internet Population

Abstract

The discipline where sentiment / opinion / emotion has been identified and classified in human written text is well known as sentiment analysis. A typical computational approach to sentiment analysis starts with prior polarity lexicons where entries are tagged with their prior out of context polarity as human beings perceive using their cognitive knowledge. Till date, all research efforts found in sentiment analysis literature deal mostly with English texts. In this article, we propose an interactive gaming (Dr Sentiment) technology to create and validate SentiWordNet for three Indian languages, Bengali, Hindi and Telugu by involving Internet population. Dr Sentiment is an online game introduces a fictitious character, interact with players using series of questions and finally reveal the behavioral or sentimental status of any player and store the lexicons as the players polarized during playing. A number of automatic, semiautomatic and manual validations and evaluation methodologies have been adopted to measure the coverage and credibility of the developed SentiWordNet(s).

1 Introduction

In order to identify sentiment from a text, lexical analysis plays a crucial role. As example *love*, *hate*, *good*, *favorite* etc words directly indicate sentiment or opinion. Various previous works (Pang et al., 2002; Wiebe and Mihalcea, 2006; Esuli et. al., 2006) have already proposed techniques for making dictionaries for those sentiment words. But identification of polarity orientation of those words is another vital research issue, called polarity identification.

Polarity Identification and classification of such sentiment lexicons is a hard contextual semantic disambiguation problem. The regulating aspects of semantic orientation of a lexicon are natural language context information (Pang et al., 2002) language properties (Wiebe and Mihalcea, 2006), domain pragmatic knowledge (Aue and Gamon, 2005) and lastly most challenging is the time dimension (Read, 2005).

The following two examples show that the polarity tag associated with a sentiment word depends on context / domain knowledge and time dimension.

Example 1: I prefer Limuzin as it is **longer** than Mercedes.
Avoid **longer** baggage during excursion in Amazon.

In the previous two examples the word **long** has been used as a sentiment/opinion word. But in the first sentence the word long depicts positive sentiment and in the second example it express as a negative sentiment.

Example 2: During 90's mobile phone users generally reported in various online reviews about their *color-phones* but in recent times *color-phone* is not just enough. People are fascinated and influenced by touch screen and various software(s) installation facilities on these new generation gadgets.

Therefore lexicon level polarity assignment is bit difficult. Previous researches (Wiebe and Mihalcea, 2006; Aue and Gamon, 2005) proposed corpus heuristic based polarity assignment at lexicon level. That means total occurrence of a particular word in a domain corpus counted and the distribution of the word as positive or negative. Suppose total occurrence of a word "**long**" in a domain corpus is n . The positive and negative occurrence of that word is S_p and S_n respectively.

Therefore in a developed sentiment lexicon the assigned positivity and negativity score of that word will be as follows:

$$\begin{array}{ll} \text{Positivity} & : \quad \frac{S_p}{n} \\ \text{Negativity} & : \quad \frac{S_n}{n} \end{array}$$

These associative scores are called **prior polarity**. Prior polarity is an approximation value and not exact. Prior polarity sentiment lexicons are required for any new language as a foundation to start the exploration of computational sentiment analysis for the language. Although contextual polarity disambiguation techniques are still required for further sentiment/opinion analysis task. Sentiment lexicons only provide a good baseline i.e. without using any NLP techniques only dictionary based approach produce a good performance. The performance of polarity clas-

sifier has been reported in the Section 5.2. Feature ablation method, reported in Table 7 shows that only dictionary based approach give good baseline score.

2 Motivations

Several prior polarity sentiment lexicons are available for English such as SentiWordNet (Esuli et al., 2006), Subjectivity Word List (Wilson et al., 2005), WordNet Affect list (Strapparava et al., 2004), Taboada's adjective list (Voll et al., 2006).

Among these publicly available sentiment lexicon resources we find that SentiWordNet is most widely used (number of citation is higher than other resources¹) in several applications such as sentiment analysis, opinion mining and emotion analysis. SentiWordNet is an automatically constructed lexical resource for English that assigns a positivity score and a negativity score to each WordNet synset. Therefore we decided to develop SentiWordNet for new languages.

There are numbers of research endeavor could be found in literature for creation of Sentiment Lexicon in several languages and domains. These techniques could be broadly categorized in two genres, one follows classical manual annotation (Andreevskaia and Bergler, 2006);(Wiebe and Riloff, 2006); (Mohammad et al., 2008) techniques and either proposes various automatic techniques (Tong, 2001). Both types of techniques have few limitations. Manual annotation techniques are undoubtedly trustable but it took long time. Especially high numbers of annotators are needed to overcome one's senti-mentality. Automatic processes are good but still it demands manual validations. Automatic processes may fail to cover the multiple domains as automatic processes trust on specific corpus.

Literature survey strongly proves that the polarity of sentiment lexicons depend on multiple factors such as: language specific, domain specific, time specific and may be other hidden multiple aspects. Moreover sentiment is a social understanding which we, the human being learn from the society by cognitive interaction day by day. Therefore sentiment is one's or more than one's out of context (as prior polarity has no contextuality) psychology regarding any topic or concept.

Therefore involving people is the best way to capture the sentiment of the human society. But

as stated earlier human annotators are quite unavailable. Hence we created an online game to attract internet population for the creation of SentiWordNet(s) automatically. Involvement of Internet population is good idea as the population is very high in number and ever growing (approx. 360,985,492)², there are peoples with various languages, cultures, age etc. Therefore Internet population is not biased towards any domain, language or particular society.

The developed online game "*Dr Sentiment*", revolutionize the idea of making prior polarity sentiment lexicon for any new language by involving internet population. We evaluated the coverage and credibility of the generated sentiment lexicons by Dr Sentiment.

3 Source Lexicon Acquisition

SentiWordNet and Subjectivity Word List have been identified as the most reliable source lexicons. A merged sentiment lexicon has been developed from both the resources by removing the duplicates. It has been observed that 64% of the single word entries are common in the Subjectivity Word List and SentiWordNet. The new merged sentiment lexicon consists of 14,135 numbers of tokens. Several filtering techniques have been applied to generate the new list.

A subset of 8,427 sentiment words has been extracted from the English SentiWordNet, by selecting those whose orientation strength is above the heuristically identified threshold of 0.4. The words whose orientation strength is below 0.4 are ambiguous and may lose their subjectivity in the target language after translation. A total of weakly subjective 2652 words are discarded (Rada et al., 2007) from the Subjectivity word list.

In the next stage the words whose POS category in the Subjectivity word list is undefined and tagged as "*anypos*" are considered. These words may generate sense ambiguity issues in the next stages of subjectivity detection.

Some words in the Subjectivity word list are inflected e.g., memories. These words would be stemmed during the translation process, but some words present no subjectivity property after stemming (memory has no subjectivity property). A word may occur in the subjectivity list in many inflected forms. Individual clusters for the words sharing the same root form are created and then checked in the SentiWordNet for validation.

¹ <http://citeseerx.ist.psu.edu/>

² <http://www.internetworldstats.com/stats.htm>

Table 1: English SentiWordNet and Subjectivity Word List Statistics

Words	SentiWordNet		Subjectivity Word List	
	Single	Multi	Single	Multi
Entries	115424	79091	5866	990
Unambiguous	20789	30000	4745	963
Ambiguous	Threshold	Orientation Strength	Subjectivity Strength	POS
	86944	30000	2652	928

If the root word exists in the SentiWordNet then it is assumed that the word remains subjective after stemming and hence is added to the new list. Otherwise the cluster is completely discarded to avoid any further ambiguities. Various statistics of the English SentiWordNet and Subjectivity Word List are reported in Table 1.

4 Dr Sentiment

There are several motivations behind developing an intuitive game to automatically create multilingual SentiWordNet(s). Sentiment lexicon generation from any source language to target language has several issues or limitations i.e.

- Source language word may have no sentiment value in target language (cross language limitation)
- Sentiment score may not be equal to source language
- Relative sentiment score is needed rather than absolute score
- Language / Culture specific lexicons should be included
- Sentiment score should be updated by time

In the history of Information Retrieval research there is a milestone when ESP³ game (Ahn et al., 2004) innovate the concept of a game to automatically label images available in World Wide Web. It has been proven as most reliable strategy to automatically annotate the online images. We are highly motivated by the success of the Image Labeler game and thus proposed an intuitive game to create and validate sentiment lexicons for three Indian languages. Our endeavor is to use successfully high number of Internet population to create and validate SentiWordNet(s) for three Indian languages. A brief statistics about Internet population has been reported in Table 2.

Dr Sentiment is an interactive game⁴. Dr Sentiment is a fictitious character, will ask a player a

set of simple questions and can reveal his/her sentimental status. This strategy revolutionize over every technique we discussed above. The lexicons tagged by this system are credible as it is tagged by human being moreover all the aspect of limitations has been covered using this strategy. As player can play in their native language so there is no issue of cross language limitations. Different tables are maintained for different languages. Relative sentiment score has been calculated by question type 2 (described in 4.1.2 Section). Language or culture specific words are being captured by question type 3 and 4 (described in 5.4.3 and 5.4.4 sections respectively). It has no limitations as a static sentiment lexicon set as it is updated regularly. Almost 100 players per day are currently playing it throughout the world in different languages. A snap of different screens from the game is presented in Figure 1.

For word based translation Google translation⁵ service has been used. It is a nice web service that translates at least at word level without any ambiguity. To avoid biased output for retrieved images from Google we randomize images from first ten results by Google.

4.1 Strategy

There are four types of questions as Q1, Q2, Q3 and Q4. Dr Sentiment asks 30 questions to each player. There are predefined distributions of each question type as 11 for Q1, 11 for Q2, 4 for Q3 and 4 for Q4. The questions are randomly asked to keep the game interesting and out of monotonous or boring.

4.1.1 Q1

An English word from the English SentiWordNet is randomly chosen. A Google image search API fired with the word as a query. An image along with the word itself is shown in the Q1 page of Dr Sentiment game. A word along with an image

³ <http://www.espgame.org/>

⁴ <http://www.amitavadas.com/Sentiment%20Game/>

⁵ <http://translate.google.com/>

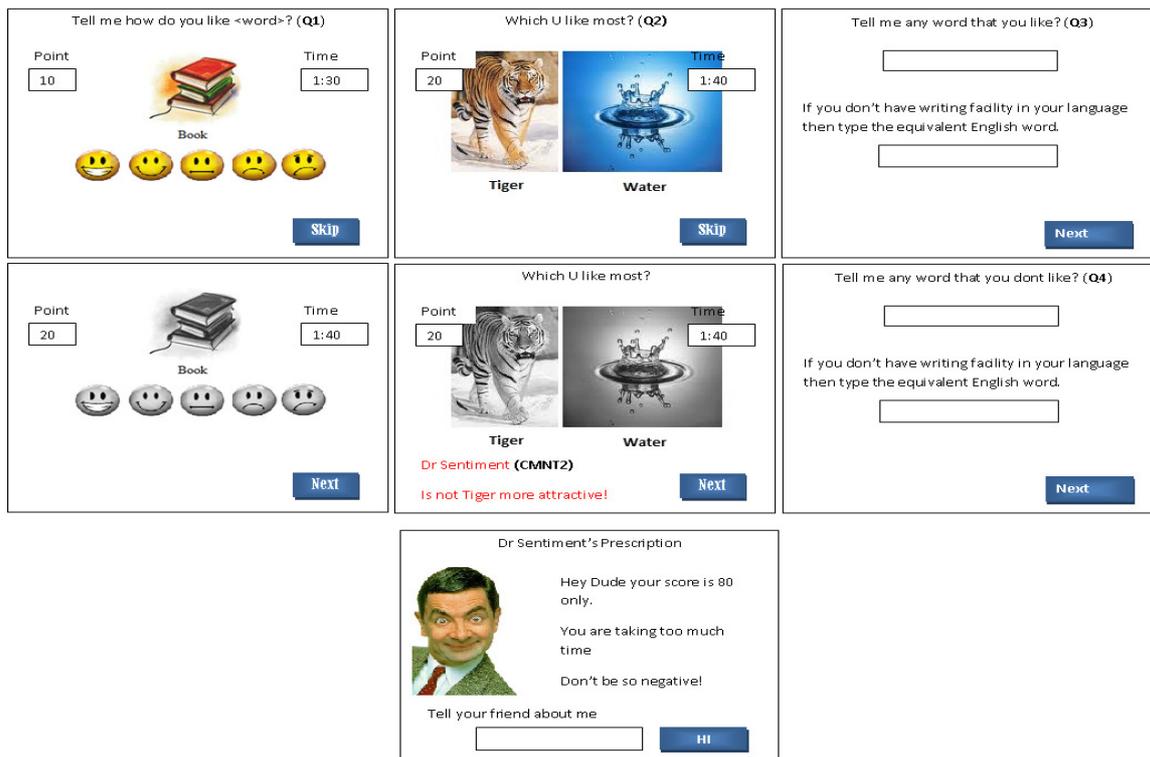


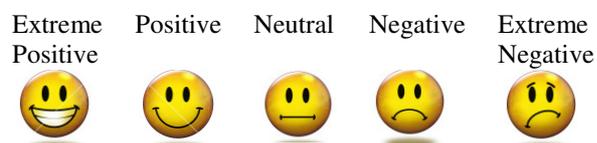
Figure 1: Snaps from Dr Sentiment Game

Table 2: Internet Usage and Population Statistics

WORLD INTERNET USAGE AND POPULATION STATISTICS						
World Regions	Population (2010 Est.)	Internet Users Dec. 31, 2000	Internet Users Latest Data	Penetration (% Population)	Growth 2000-2010	Users % of Table
Africa	1,013,779,050	4,514,400	110,931,700	10.9 %	2,357.3 %	5.6 %
Asia	3,834,792,852	114,304,000	825,094,396	21.5 %	621.8 %	42.0 %
Europe	813,319,511	105,096,093	475,069,448	58.4 %	352.0 %	24.2 %
Middle East	212,336,924	3,284,800	63,240,946	29.8 %	1,825.3 %	3.2 %
North America	344,124,450	108,096,800	266,224,500	77.4 %	146.3 %	13.5 %
Latin America/Caribbean	592,556,972	18,068,919	204,689,836	34.5 %	1,032.8 %	10.4 %
Oceania / Australia	34,700,201	7,620,480	21,263,990	61.3 %	179.0 %	1.1 %
WORLD TOTAL	6,845,609,960	360,985,492	1,966,514,816	28.7 %	444.8 %	100.0 %

is more attractive rather than only a word. The words are shown in player's own language as he/she specified in the login page.

The sentiment score calculated by the different emoticons pressed by different players and scale of sentiment score assigned accordingly as extreme positive (pos: 0.5, neg: 0.0), positive (pos: 0.25, neg: 0.0), neutral (pos: 0.0, neg: 0.0), negative (pos: 0.0, 0.25), extreme negative (pos: 0.0, neg: 0.5).



For Languages other than English the word along with its associate property (POS, Offset) are inserted into the language table. The new positivity score and negativity score are being stored according to the previous strategy over original score on the English word's score and copied to the language table.

4.1.2 Q2

Randomly n (presently 2-4) words should be chosen from the source English table. According images are retrieved from the Google. Player will ask to choose either one of them. The relative score is calculated accordingly and stored in corresponding language table.

4.1.3 Q3

It is very simple to ask a player about any positive word. The words are added to the corresponding language table as pos: 0.5 and neg: 0.0 score.

4.1.4 Q4

It is very simple to ask a player about any negative word. The word will be added to the corresponding language table as pos: 0.0 and neg: 0.5 score.

4.2 Comment Architecture

There are three types of Comments here as CMNT1, CMNT2 and the final comment as Dr Sentiment's prescription.

4.2.1 CMNT1

Comment type 1 has 5 variations as. The comment table is as Table 3.

- Positive word may have tagged as negative. (PN)
- Positive word may have tagged as positive. (PP)
- Negative word may have tagged as positive. (NP)
- Negative word may have tagged as negative. (NN)
- Neutral. (NU)

Comments are retrieved from comment type table according to their category as described and randomly.

4.2.2 CMNT2

The strategy here is as same as the CMNT 1. Comment type 2 has only 2 variations as.

- Positive word may have tagged as negative. (PN)
- Negative word may have tagged as positive. (NP)

4.3 Dr Sentiment's Prescription

The final comment depends on various factors as total positive, negative or neutral tagging and total time taken. Some more rules are incorporated as positive words tagged as negative, negative words tagged as positive etc.

5 Senti-Mentality

Several analyses have been done on the developed sentiment lexicons to understand the sentimental behavior of people depending upon location, age, sex, profession and etc. The login form of the "Dr Sentiment" ask to provide several information such as country, city, age, sex, profession etc. A tracking system keeps track of every player's tagged words. Player specific separate log has been maintained for tagging. A word previously tagged by a player is avoided by the tracking system for the next time playing as our intension is to tag more and more words involving Internet population. We hope this strategy help to keep the game interesting and ever new to the players as a proof we found that a large number of returning players increased after this change. Statistical analyses reveal some interesting data as described below.

Table 3: Comments

PN	PP	NP	NN	NU
You don't like <word>!	Good you have a good choice!	Is <word> good!	Yes <word> is too bad!	You should speak out frankly!
You should like <word>!	I love <word> too!	I hope it is a bad choice!	You are quite right!	You are too diplomatic!
But <word> is a good itself!	I support your view!	I don't agree with you!	I also don't like <word>!	Why you hiding from me? I am Dr Sentiment.

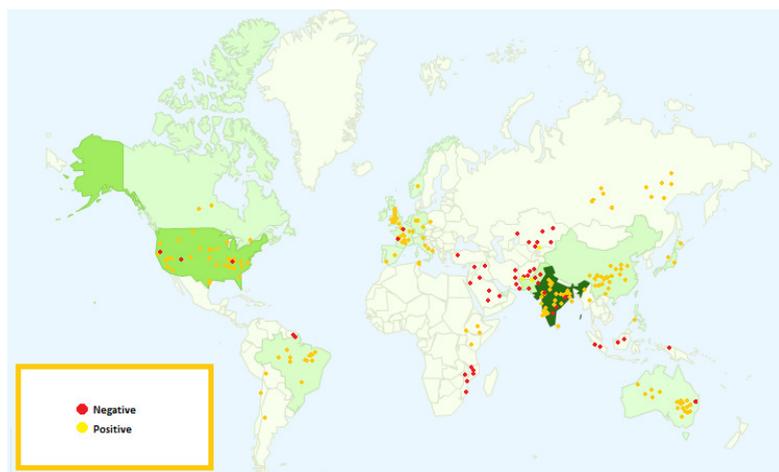


Figure 2: Geospatial Senti-Mentality

5.1 Concept-Culture-Wise Analysis

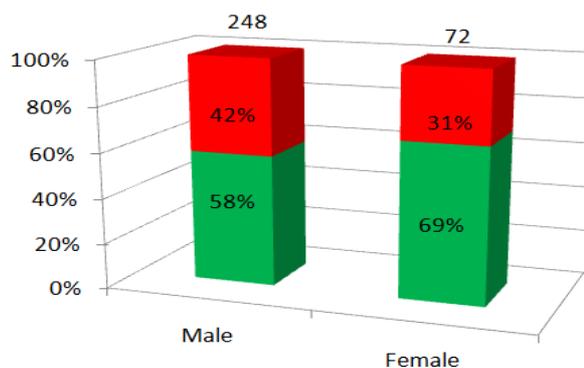
During analysis we found an interesting outcome. The word “blue” get tagged by different players around the world. But surprisingly it has been tagged as positive from a portion of the world and negative by a different portion of the world. A graphical illustration may illustrate the problem well. The observation is most of the negative tagging are coming from middle-east and especially from Islamic countries. I start finding the root cause of this peculiar behavior and found a line in Wiki⁶ (see in Religion Section) may give a good explanation: “Blue in Islam: In verse 20:102 of the Qur’an, the word زرق zurq (plural of azraq 'blue') is used metaphorically for evil doers whose eyes are glazed with fear”. May be some other explanations could be there but it is undoubtedly an interesting observation for sentiment lexicon creation.

5.2 Age-Wise Analysis

Another interesting observation is sentiment understanding may vary age-wise. For better understanding we should provide the total statistics and the age wise distribution of total players. Total 533 players have been taken part till date. The total number of distribution of players age wise are shown at top of every bar. In the Figure 1 the horizontal bars are divided into two colors (Green depicts Positivity and Red depicts negativity) according to the total positivity and negativity scores, gathered during playing. It could be treated as a good sociological study. It gives an idea that how the overall senti-mentality has been changed of a human being during various stage of his/her life.

5.3 Gender Specific

It is incredibly noted that women are more positive than a man. May be this sentimentality varies subject wise but it is an overall statistics.



5.4 Other-Wise

We have witnessed two important observations as stated in two previous sections. Although there are still multiple dimension are left to be explored.

Some of the important dimension may be country, city, age, sex, profession etc. Combination of dimension may reveal some interesting study. Combinational dimension pairs such as location-age, location-profession, sex-wise, language-location etc could be possible. Interesting we found that woman are more positive than man.

6 Expected Impact of the Resources

There may be a hidden question, that if Google translation services produce any wrong translation, then what will be the impact into the targeted language-specific SentiWordNet(s)?

We have manually checked Google word-level translation for Indian languages and there were

⁶ <http://en.wikipedia.org/wiki/Blue>

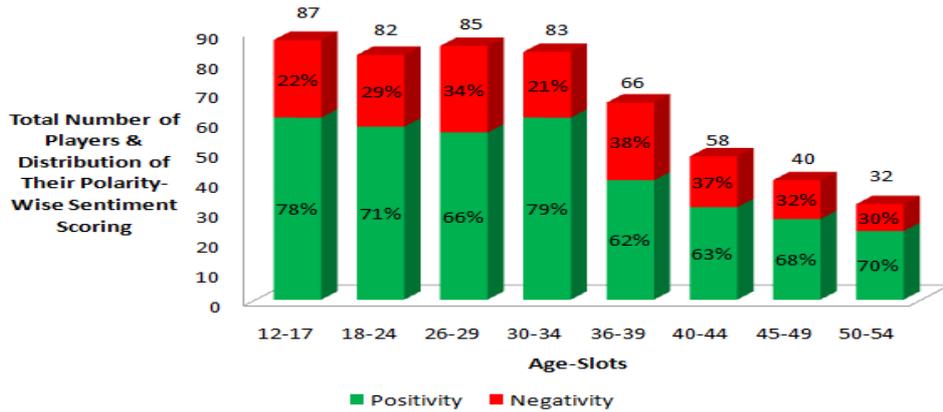


Figure 3: Age-Wise Senti-Mentality

very little error. Let assume Google produces some wrong word-level translation then the question is: what should our system do to handle this? Google system has consistency, i.e., for any particular word Google produces same erroneous output every time. So the same erroneous output of any source word gets tagged by native speakers (players). The background database of the system stores data into language specific tables, so there is no inter-language ambiguity. May be for the erroneous outputs by Google rise difficulties for cross-lingual use but still developed SentiWordNet(s) are useful for monolingual use. Presently Bengali SentiWordNet contains 20,546, Hindi SentiWordNet contains 13,889 and Telugu SentiWordNet contains 10,204 unique entries.

Undoubtedly the generated lexicons are important resources for three languages for sentiment/opinion or emotion analysis task. Moreover the other non linguistic dimensions are very much important for further analysis and in several newly discovered sub-disciplines such as: Geospatial Information retrieval (Egenhofer, 2002), Personalized search (Gaucha et al., 2003), and Recommender System (Adomavicius and Tuzhilin, 2005) etc.

7 Evaluation

Andera Esuli and Fabrizio Sebastiani (Esuli and Fabrizio, 2006) (The inventors of the SentiWordNet) have calculated the reliability of the sentiment scores attached to each synsets in the SentiWordNet. They have tagged sentiment words in the English WordNet with positive and negative sentiment scores. We extend our vision and proposed two extrinsic evaluation strategies. The evaluation strategies have been adopted for the developed Bengali SentiWordNet based on the two usages of the sentiment lexicon, subjec-

tivity classifier and polarity identifier. The Hindi and Telugu SentiWordNet(s) have been partly evaluated. SentiWordNet(s) for other languages have not been evaluated yet, may it is our future direction of research.

7.1 Coverage

We experimented with NEWS and BLOG corpora for subjectivity detection. Sentiment lexicons are generally domain independent but it provides a good baseline while working with sentiment analysis systems. The coverage of the developed Bengali SentiWordNet is evaluated by using it in a subjectivity classifier (Das and Bandyopadhyay, 2009). The statistics of the NEWS and BLOG corpora is reported in Table 4.

Table 4: Bengali Corpus Statistics

	NEWS	BLOG
Total number of documents	100	-
Total number of sentences	2234	300
Average number of sentences in a document	22	-
Total number of wordforms	28807	4675
Average number of wordforms in a document	288	-
Total number of distinct wordforms	17176	1235

Table 5: Subjectivity Classifier using SentiWordNet

Languages	Domain	Precision	Recall
English	MPQA	76.08%	83.33%
	IMDB	79.90%	86.55%
Bengali	NEWS	72.16%	76.00%
	BLOG	74.6%	80.4%

For comparison with the coverage of English SentiWordNet the same subjectivity classifier (Das and Bandyopadhyay, 2009) has been applied on Multi Perspective Question Answering (MPQA) (NEWS) and IMDB Movie review corpus along with English SentiWordNet.

Table 6: Polarity Performance Using Bengali SentiWordNet

Features	Overall Performance
SentiWordNet	47.60%
SentiWordNet + Negative Word	50.40%
SentiWordNet + Negative Word + Stemming Cluster	56.02%
SentiWordNet + Negative Word + Stemming Cluster + Functional Word	58.23%
SentiWordNet + Negative Word + Stemming Cluster + Functional Word Parts Of Speech	61.9%
SentiWordNet + Negative Word + Stemming Cluster + Functional Word + Parts Of Speech +Chunk	66.8%
SentiWordNet + Negative Word + Stemming Cluster + Functional Word + Parts Of Speech + Chunk +Dependency tree feature	70.04%

The result of the subjectivity classifier on both the corpus proves that the coverage of the Bengali SentiWordNet is reasonably good.

The subjectivity word list used in the subjectivity classifier is developed from the IMDB corpus and hence the experiments on the IMDB corpus have yielded high precision and recall scores. The developed Bengali SentiWordNet is domain independent and still its coverage is very good as shown in Table 5. We can't evaluate the Hindi and Telugu SentiWordNet(s) as there is no annotated data is available publicly for these languages.

7.2 Credibility of Polarity Scores

This evaluation metric measures the reliability of the associated polarity scores in the sentiment lexicons. A typical approach to sentiment analysis is to start with a lexicon of positive and negative words and phrases. In these lexicons, entries are tagged with their prior out of context polarity. To measure the reliability of polarity scores in the developed Bengali SentiWordNet, a polarity classifier (Das and Bandyopadhyay, 2010) has been developed using the Bengali SentiWordNet along with some other linguistic features. Feature ablation method proves that the generated SentiWordNet gives a good baseline. Although contextual polarity disambiguation techniques are required using multiple feature.

Feature ablation method proves that the associated polarity scores in the developed Bengali SentiWordNet are reliable. Table 6 shows the performance of a polarity classifier using the Bengali SentiWordNet. The polarity wise overall performance of the polarity classifier is reported in Table 7.

Table 7: Polarity-wise Performance Using Bengali SentiWordNet

Polarity	Precision	Recall
Positive	56.59%	52.89%
Negative	75.57%	65.87%

Comparative study with an English polarity classifier that works with only prior polarity lexicon is necessary but no such works have been identified from literature.

An arbitrary 300 words have been chosen from the Hindi SentiWordNet for human evaluation. Two persons are asked to manually check it and the result is reported in Table 8. The coverage of the Hindi SentiWordNet has not been evaluated, as no manually annotated sentiment corpus is available.

Table 8: Evaluation of Polarity Score of Developed Hindi SentiWordNet

Polarity	Positive	Negative
Percentage	88.0%	91.0%

For Telugu we rely on the Dr Sentiment with Telugu words on screen. Only 30 users have played the Telugu language specific game till date. Total 920 arbitrary words have been tagged and the accuracy of the polarity scores is reported in Table 9. The coverage of Telugu SentiWordNet has not been evaluated, as no manually annotated sentiment corpus is available.

Table 9: Evaluation of Polarity Score of Developed Telugu SentiWordNet

Polarity	Positive	Negative
Percentage	82.0%	78.0%

8 Conclusion

Indian languages SentiWordNet(s) has been developed by **Dr Sentiment** involving Internet Population and this could be expanded by the using dictionary based approach, WordNet approach, corpus based approaches (Das and Bandyopadhyay, 2010).

Our next target is to assign proper sense id to each SentiWordNet entry from AsianWordNet.

References

- Andreevskaia Alina and Bergler Sabine. CLaC and CLaC-NB: Knowledge-based and corpus-based approaches to sentiment tagging. In Proceedings of the 4th SemEval-2007, pages 117–120, Prague, June 2007.
- Aue A. and Gamon M., Customizing sentiment classifiers to new domains: A case study. In Proceedings RANLP, 2005.
- Das A. and Bandyopadhyay S. (2010). Phrase-level Polarity Identification for Bengali, In IJCLA, Vol. 1, No. 1-2, Jan-Dec 2010, ISSN 0976-0962, Pages 169-182.
- Das A. and S. Bandyopadhyay. SentiWordNet for Indian Languages, In the 8th Workshop on Asian Language Resources (ALR), August 21-22, Beijing, China, 2010.
- Esuli Andrea and Sebastiani Fabrizio. SentiWordNet: A publicly available lexical resource for opinion mining. In Proceedings of LREC, 2006.
- Mihalcea Rada, Banea Carmen and Wiebe Janyce. Learning multilingual subjective language via cross-lingual projections. In Proceedings of ACL, pages 976–983, Prague, Czech Republic, June 2007.
- Mohammad Saif, Dorr Bonnie, and Hirst Graeme. Computing Word-Pair Antonymy. In Proceedings of EMNLP-2008, October 2008, Waikiki, Hawaii.
- Pang Bo, Lee Lillian, and Vaithyanathan Shivakumar. Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of EMNLP, pages 79–86, 2002.
- Read Jonathon. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In Proceedings of the ACL Student Research Workshop, 2005.
- Strapparava, C. and Valitutti, A.(2004). WordNet-Affect: an affective extension of WordNet. In Proceedings of LREC 2004, pages 1083 – 1086, Lisbon, May 2004.
- Voll, K. and Taboada, M. (2007) Not All Words are Created Equal: Extracting Semantic Orientation as a Function of Adjective Relevance. In Proceedings of the 20th Australian Joint Conference on Artificial Intelligence. pages. 337-346.
- Wiebe Janyce and Mihalcea Rada. Word sense and subjectivity. In Proceedings of COLING/ACL-06. Pages 1065--1072.
- Wiebe Janyce and Riloff Ellen. Creating Subjective and Objective Sentence Classifiers from Unannotated Texts. In Proceeding CICLING, Pages 475–486, 2006.
- Wilson Theresa, Wiebe Janyce and Hoffmann Paul (2005). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In Proceedings of HLT/EMNLP 2005, Vancouver, Canada.
- Toward the Semantic Geospatial Web M. Egenhofer ACM-GIS 2002, McLean, VI A. Voisard and S.-C. Chen (eds.), pp. 1-4, November 2002.
- Susan Gaucha, Jason Chaffeeb and Alexander Pretschner. Ontology-based personalized search and browsing. In Proceeding of Web Intelligence and Agent Systems: An international journal. 2003. Pages 219–234. ISSN 1570-1263/03.
- Gediminas Adomavicius and Alexander Tuzhilin. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. In the Proceeding of IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 17, NO. 6, JUNE 2005. ISSN 1041-4347/05. Pages 734-749
- Richard M. Tong. An operational system for detecting and tracking opinions in on-line discussion. In Proceedings of the Workshop on Operational Text Classification (OTC), 2001.