Dr Sentiment Knows Everything!

Amitava Das and **Sivaji Bandyopadhyay**
Department of Computer Science and Engineering
Jadavpur University
India
amitava.santu@gmail.com sivaji_cse_ju@yahoo.com

Abstract

Sentiment analysis is one of the hot demanding research areas since last few decades. Although a formidable amount of research have been done, the existing reported solutions or available systems are still far from perfect or do not meet the satisfaction level of end users'. The main issue is the various conceptual rules that govern sentiment and there are even more clues (possibly unlimited) that can convey these concepts from realization to verbalization of a human being. Human psychology directly relates to the unrevealed clues and governs the sentiment realization of us. Human psychology relates many things like social psychology, culture, pragmatics and many more endless intelligent aspects of civilization. Proper incorporation of human psychology into computational sentiment knowledge representation may solve the problem. In the present paper we propose a template based online interactive gaming technology, called ***Dr Sentiment*** to automatically create the ***PsychoSentiWordNet*** involving internet population. The PsychoSentiWordNet is an extension of SentiWordNet that presently holds human psychological knowledge on a few aspects along with sentiment knowledge.

1 Introduction

In order to identify sentiment from a text, lexical analysis plays a crucial role. For example, words like ***love, hate, good*** and ***favorite*** directly indicate sentiment or opinion. Previous works (Pang et al., 2002; Wiebe and Mihalcea, 2006; Baccianella et. al., 2010) have already proposed various techniques for making dictionaries for those sentiment words. But polarity assignment of such sentiment lexicons is a hard semantic disambiguation problem. The regulating aspects which govern the lexical level semantic orientation are natural language context (Pang et al., 2002), language properties (Wiebe and Mihalcea, 2006), domain pragmatic knowledge (Aue and Gamon, 2005), time dimension (Read, 2005), colors and culture (Strapparava and Ozbal, 2010) and many more unrevealed hidden aspects. Therefore it is a challenging and enigmatic research problem.

The current trend is to attach **prior polarity** to each entry at the sentiment lexicon level. Prior polarity is an approximation value based on heuristics based statistics collected from corpus and not exact. The probabilistic fixed point prior polarity scores do not solve the problem completely rather it places the problem into next level, called contextual polarity classification.

We start with the hypothesis that the summation of all the regulating aspects of sentiment orientation is human psychology and thus it is a multifaceted problem (Liu, 2010). More precisely what we mean by human psychology is the union of all known and unknown aspects that directly or indirectly govern the sentiment orientation knowledge of us. The regulating aspects wrapped in the present PsychoSentiWordNet are **Gender**, **Age**, **City**, **Country**, **Language and Profession**.

The PsychoSentiWordNet is an extension of the existing SentiWordNet 3.0 (Baccianella et. al., 2010) to hold the possible psychological ingredients and govern the sentiment understandability of us. The PsychoSentiWordNet holds variable prior polarity scores that could be fetched depending upon those psychological regulating aspects.

Proceedings of the ACL-HLT 2011 System Demonstrations, pages 50–55,
Portland, Oregon, USA, 21 June 2011.

An example with the input word '***High***' may illustrate the definition better:

Aspects (Profession)	Polarity
Null	Positive
Businessman	Negative
Share Broker	Positive

In this paper, we propose an interactive gaming (Dr Sentiment) technology to collect psycho-sentimental polarity for lexicons. This technology has proven itself as an excellent technique to collect psychological sentiment of human society even at multilingual level. Dr Sentiment presently supports 56 languages and therefore we may call it ***Global PsychoSentiWordNet***. The supported languages by Dr Sentiment are reported in Table 1.

In this section we have philosophically argued about the necessity of developing PsychoSentiWordNet. In the next section 2 we will describe the technical details of the proposed architecture for building the lexical resource. Section 3 explains about some exciting outcomes of PsychoSentiWordNet. The developed PsychoSentiWordNet(s) are expected to help automatic sentiment analysis research in many aspects and other disciplines as well and have been described in section 4.The data structure and the organization are described in section 5. The conclusion is drawn in section 6.

2 Dr Sentiment

Dr Sentiment[1] is a template based interactive online game, which collects player's sentiment by asking a set of simple template based questions and finally reveals a player's sentimental status. Dr Sentiment fetches random words from SentiWordNet synsets and asks every player to tell about his/her sentiment polarity understanding regarding the concept behind the word fetched by it.

There are several motivations behind developing the intuitive game to automatically collect human psycho-sentimental orientation information.

In the history of Information Retrieval research there is a milestone when ESP game[2] (Ahn et al., 2004) innovated the concept of a game to automatically label images available in the World Wide Web. It has been identified as the most reliable strategy to automatically annotate the online images. We are highly motivated by the success of the Image Labeler game.

A number of research endeavors could be found in the literature for creation of Sentiment Lexicon in several languages and domains. These techniques can be broadly categorized into two classes, one follows classical manual annotation techniques (Andreevskaia and Bergler, 2006);(Wiebe and Riloff, 2006) while the other follows various automatic techniques (Mohammad et al., 2008). Both types of techniques have few limitations. Manual annotation techniques are undoubtedly trustable but it generally takes time. Automatic techniques demand manual validations and are dependent on the corpus availability in the respective domain. Manual annotation techniques require a large number of annotators to balance one's sentimentality in order to reach agreement. But human annotators are quite unavailable and costly.

Sentiment is a property of human intelligence and is not entirely based on the features of a language. Thus people's involvement is required to capture the sentiment of the human society. We have developed an online game to attract internet population for the creation of PsychoSentiWordNet automatically. Involvement of Internet population is an effective approach as the population is very high in number and ever growing (approx. 360,985,492) [3] . Internet population consists of people with various languages, cultures, age etc and thus not biased towards any domain, language or particular society. A detailed statistics on the Internet usage and population has been reported in the Table 2.

The lexicons tagged by this system are credible as it is tagged by human beings. It is not a static sentiment lexicon set [polarity changes with time (Read, 2005)] as it is updated regularly. Around 10-20 players each day are playing it throughout the world in different languages. The average number of tagging per word is about 7.47 till date.

The Sign Up form of the "Dr Sentiment" game asks the player to provide personal information such as Sex, Age, City, Country, Language and Profession. These collected personal details of a player are kept as a log record in the database.

The gaming interface has four types of question templates. The question templates are named as Q1, Q2, Q3 and Q4.

[1] http://www.amitavadas.com/Sentiment%20Game/index.php
[2] http://www.espgame.org/
[3] http://www.internetworldstats.com/stats.htm

Languages							
Afrikaans	Bulgarian	Dutch	German	Irish	Malay	Russian	Thai
Albanian	Catalan	Estonian	Greek	Italian	Maltese	Serbian	Turkish
Arabic	Chinese	Filipino	Haitian	Japanese	Norwegian	Slovak	Ukrainian
Armenian	Croatian	Finnish	Hebrew	Korean	Persian	Slovenian	Urdu
Azerbaijani	Creole	French	Hungarian	Latvian	Polish	Spanish	Vietnamese
Basque	Czech	Galician	Icelandic	Lithuanian	Portuguese	Swahili	Welsh
Belarusian	Danish	Georgian	Indonesian	Macedonian	Romanian	Swedish	Yiddish

Table 1: Languages

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 %

Table 2: Internet Usage and Population Statistics

To make the gaming interface more interesting images have been added. These images have been retrieved by Google image search API[4] and to avoid biasness we have randomized among the first ten images retrieved by Google.

2.1 Gaming Strategy

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. These numbers are arbitrarily chosen and randomly changed for experimentation. The questions are randomly asked to keep the game more interesting. For word based translation Google translation[5] service has been used. At each Question (Q) level translation service has been used to display the sentiment word into player's own language. Google API provides multiple senses for word level translation and currently only the first sense has been picked automatically.

2.2 Q1

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

Players press the different emoticons (Figure 1) to express their sentimentality. The interface keeps log records of each interaction.

Figure 1: Emoticons to Express Player's Sentiment

2.3 Q2

This question type is specially designed for relative scoring technique. For example: ***good*** and ***better*** both are positive but we need to know which one is more positive than other. Table 3 shows how in SentiWordNet relative scoring has been made. With the present gaming technology relative polarity scoring has been assigned to each ***n-n*** word pair combination.

Randomly ***n*** (presently 2-4) words have been chosen from the source SentiWordNet synsets along with their images as retrieved by Google API. Each player is then asked to select one of them that he/she likes most. The relative score is calculated and stored in the corresponding log table.

Word	Positivity	Negativity
Good	0.625	0.0
Better	0.875	0.0
Best	0.980	0.0

Table 3: Relative Sentiment Scores in SentiWordNet

[4] http://code.google.com/apis/imagesearch/
[5] http://translate.google.com/

2.4 Q3

The player is asked for any positive word in his/her mind. This technique helps to increase the coverage of existing SentiWordNet. The word is then added to the existing PsychoSentiWordNet and further used in Q1 to other users to note their sentimentality about the particular word.

2.5 Q4

A player is asked by Dr Sentiment about any negative word. The word is then added to the existing PsychoSentiWordNet and further used in Q1 to other users to note their sentimentality about the particular word.

2.6 Comment Architecture

There are three types of Comments, Comment type 1 (CMNT1), Comment type 2 (CMNT2) and the final comment as Dr Sentiment's prescription. CMNT1 type and CMNT2 type comments are associated with question types Q1 and Q2 respectively.

2.6.1 CMNT1

Comment type 1 has 5 variations as shown in the Comment table in Table 4. Comments are randomly retrieved from comment type table according to their category:

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

2.6.2 CMNT2

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

- Positive word has been tagged as negative (PN)
- Negative word has been tagged as positive (NP)

2.7 Dr Sentiment's Prescription

The final prescription depends on various factors such as total number of positive, negative or neutral comments and the total time taken by any player. The final prescription also depends on the range of the accumulated values of all the above factors.

This is the most important appealing factor to a player. The motivating message for players is that Dr Sentiment can reveal their sentimental status: whether they are extreme negative or positive or very much neutral or diplomatic etc. It is not claimed that the revealed status of a player by Dr Sentiment is exact or ideal. It is only to make the players motivated but the outcomes of the game effectively helps to store human sentimental psychology in terms of computational lexicon.

A word previously tagged by a player is avoided by the tracking system during subsequent turns by the same player. The intension is to tag more and more words involving Internet population. We observe that the strategy helps to keep the game interesting as a large number of players return to play the game after this strategy was implemented.

3 Senti-Mentality

PsychoSentiWordNet gives a good sketch to understand the psycho-sentimental behavior of the human society depending upon proposed psychological dimensions. The PsychoSentiWordNet is basically the log records of every player's tagged words.

3.1 Concept-Culture-Wise Analysis

The word "blue" gets tagged by different players around the world. But surprisingly it has been tagged as positive from one part of the world and negative from another part of the world. The graphical illustration in Figure 2 may explain the situation better. The observation is that most of the negative tags are coming from the middle-east and especially from the Islamic countries.

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.

Table 4: Comments

We found a line in Wiki[6] (see in Religion Section) that may provide 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". But other explanations may be there for this situation. This is an interesting observation that supports the effectiveness of the developed PsychoSentiWordNet. This information could be further retrieved from the developed source by giving information like (blue, Italy), (blue, Iraq) or (blue, USA) etc.

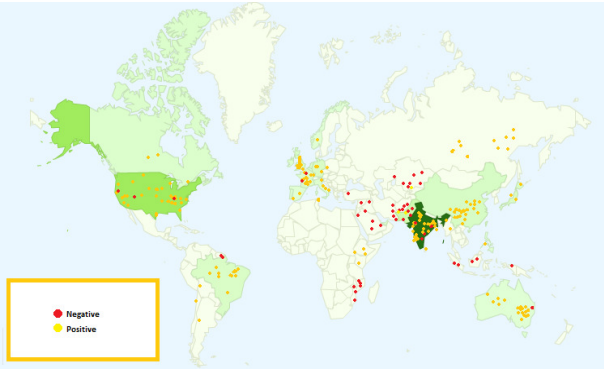

Figure 2: Geospatial Senti-Mentality

3.2 Age-Wise Analysis

Another interesting observation is that sentimentality may vary age-wise. For better understanding we look at the total statistics and the age wise distribution of all the players. Total 533 players have taken part till date. The total number of players for each range of age is shown at the top of every bar.

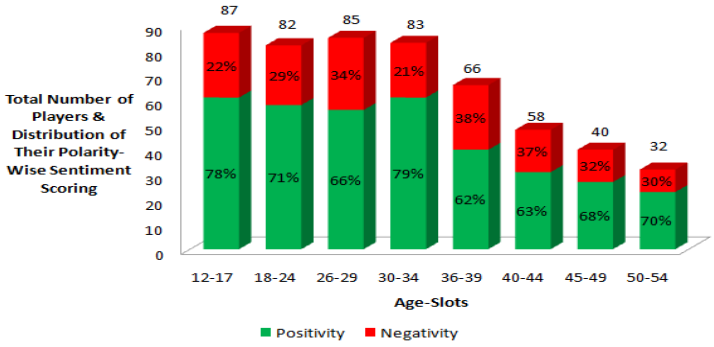

Figure 3: Age-Wise Senti-Mentality

In Figure 3 the horizontal bars are divided into two colors (Green depicts the Positivity and Red depicts the negativity) according to the total positivity and negativity scores, gathered during playing. This sociological study gives an idea on the variation of sentimentality with age. This information may be retrieved from the developed source by giving information like (X, 36-39) or (X, 45-49) etc where X denotes any arbitrary lexicon synset.

3.3 Gender-Wise Analysis

It is observed from the collected statistics that women are more positive than men! The variations in sentimentality among men and women are shown in the following Figure 4.

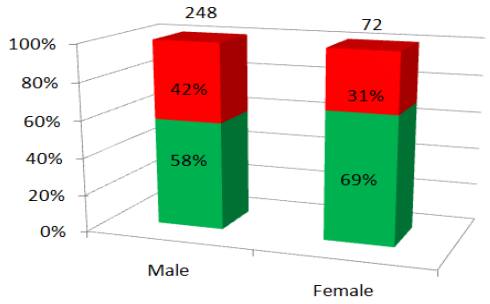

Figure 4: Gender Specific Senti-Mentality

3.4 Other-Wise

We have described several important observations in the previous sections and there are other important observations as well. Studies on the combinations of the proposed psychological dimensions, such as, location-age, location-profession and gender-location may reveal some interesting results.

4 Expected Impact of the Resource

Undoubtedly the generated PsychoSentiWordNet(s) are important resources for sentiment/opinion or emotion analysis task. Moreover the other non linguistic psychological dimensions are very much important for further analysis as well as for several newly discovered sub-disciplines such as: Geospatial Information retrieval (Egenhofer, 2002), Personalized search (Gaucha et al., 2003), Recommender System (Adomavicius and Tuzhilin, 2005), Sentiment Tracking (Tong, 2001) etc.

5 The Data Structure and Organization

Deciding on the data structure for the PsychoSentiWordNet was not trivial. Presently RDBMS (Relational Database Management System) has been

[6] http://en.wikipedia.org/wiki/Blue

used. Several tables are being used to keep user's clicking log and their personal information.

As one of the research motivations was to generate up-to-date prior polarity scores across various dimensions, we decided to generate web service API through which the people can access latest prior polarity scores. The developed PsychoSentiWordNet is expected to perform better than a static sentiment lexicon.

6 Conclusion and Future Directions

In the present paper the development of the ***PsychoSentiWordNet*** for 56 languages has been described. No evaluation has been done yet as there is no data available for this kind of experimentation and to the best of our knowledge this is the first endeavor where sentiment analysis meets AI and psychology.

Our present goal is to collect such corpus and carry out experiments to check whether variable prior polarity scores of PsychoSentiWordNet excel over the fixed point prior polarity score of SentiWordNet.

Automatically picked first sense from Google translation API may cause difficulties for cross lingual projection of sentiment synsets. Erroneous outputs from API may also cause some problems. But these problems lead to another research issue that may be termed as cross lingual sentiment synset linking. Presently we are giving a closer look to the qualitative analysis of developed multilingual psycho-sentiment lexicons.

Acknowledgment

The work reported in this paper was supported by a grant from the India-Japan Cooperative Program (DST-JST) Research project entitled "***Sentiment Analysis where AI meets Psychology***" funded by Department of Science and Technology (DST), Government of India.

References

Adomavicius Gediminas and Alexander Tuzhilin. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. In the Proc. of IEEE Transactions on Knowledge and Data Engineering, VOL. 17, NO. 6, June 2005. ISSN 1041-4347/05. Pages 734-749.

Ahn Luis von and Laura Dabbish. Labeling Images with a Computer Game.In the Proc. of ACM CHI 2004.

Andreevskaia Alina and Bergler Sabine. CLaC and CLaC-NB: Knowledge-based and corpus-based approaches to sentiment tagging. In the Proc. 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 the Proc. Of RANLP, 2005.

Baccianella Stefano, Andrea Esuli, and Fabrizio Sebastiani. SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In the Proc. of LREC-10.

Bo Pang, Lee Lillian, and Vaithyanathan Shivakumar. Thumbs up? Sentiment classification using machine learning techniques. In the Proc. of EMNLP, Pages 79–86, 2002.

Egenhofer M.. Toward the Semantic Geospatial Web. ACM-GIS 2002, McLean, VI A. Voisard and S.-C. Chen (eds.), Pages. 1-4, November 2002.

Gaucha Susan, Jason Chaffeeb and Alexander Pretschnerc. Ontology-based personalized search and browsing. In Proc. of Web Intelligence and Agent Systems: An international journal. 2003. Pages 219–234. ISSN 1570-1263/03.

Liu Bing . Sentiment Analysis: A Multi-Faceted Problem.In the IEEE Intelligent Systems, 2010.

Read Jonathon. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In the Proc. of the ACL Student Research Workshop, 2005.

Richard M. Tong. An operational system for detecting and tracking opinions in online discussion. In the Proc. of the Workshop on Operational Text Classification (OTC), 2001.

Saif Mohammad, Dorr Bonnie and Hirst Graeme. Computing Word-Pair Antonymy. In the Proc. of EMNLP-2008.

Strapparava, C. and Valitutti, A. WordNet-Affect: an affective extension of WordNet. In Proc. of LREC 2004, Pages 1083 – 1086

Wiebe Janyce and Mihalcea Rada. Word sense and subjectivity. In the Proc. of COLING/ACL-06. Pages 1065-1072.