PsychoSentiWordNet

Amitava Das
Department of Computer Science and Engineering
Jadavpur University
amitava.santu@gmail.com

Abstract

Sentiment analysis is one of the hot demanding research areas since last few decades. Although a formidable amount of research has been done but still the existing reported solutions or available systems are far from perfect or to meet the satisfaction level of end user's. The main issue may be there are many 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; govern 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. PsychoSentiWordNet is an extension over SentiWordNet that holds human psychological knowledge and sentiment knowledge simultaneously.

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. Various previous works (Pang et al., 2002; Wiebe and Mihalcea, 2006; Baccianella et. al., 2010) have already proposed 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.

What previous studies proposed is to attach prior polarity to each sentiment lexicon level. Prior polarity is an approximation value based on corpus heuristics based statistics and not exact. The probabilistic fixed point prior polarity scores do not solve the problem completely rather it shoves the problem into next level, called contextual polarity classification.

The hypothesis we started with is that the summation of all the regulating aspects of sentiment orientation is human psychology and thus it is called multi-faceted problem (Liu, 2010). More precisely what we meant by human psychology is the all known and unknown aspects, 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 over the existing SentiWordNet to hold the possible psychological ingredients, governs the sentiment understandability of us. The PsychoSentiWordNet holds variable prior polarity scores, could be fetched depending upon those psychological regulating aspects. An example may illustrate the definition better for the concept "***Rock_Climbing***":

Aspects (Age)	Polarity
Null	Positive
50-54	Negative
26-29	Positive

Proceedings of the ACL-HLT 2011 Student Session, pages 52–57,
Portland, OR, USA 19-24 June 2011. ©2011 Association for Computational Linguistics

In the previous example the described concept "***Rock_Climbing***" is generally positive as it is adventurous and people have it to make fun or excursion. But it demands highly physical ability thus may be not as good for aged people like the younger people.

PsychoSentiWordNet provides good coverage as it an extension over SentiWordNet 3.0 (Baccianella et. al., 2010). In this paper, we propose an interactive gaming (Dr Sentiment) technology to collect psycho-sentimental polarity for lexicons.

In this section we have philosophically argued about the necessity of developing PsychoSentiWordNet. In the next section we will describe about the technical proposed architecture for building the lexical resource. Section 3 explains about some exciting outcomes that support the usefulness of the PsychoSentiWordNet. What we believe is the developed PsychoSentiWordNet will help automatic sentiment analysis research in many aspect and other disciplines as well, described in the section 4.The data structure and organization is described in section 5 and finally the present paper concluded with 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.

There are several motivations behind developing an 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) innovate 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 literature for creation of Sentiment Lexicon in several languages and domains. These techniques can 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 the others proposed various automatic techniques (Tong, 2001). Both types of techniques have few limitations. Manual annotation techniques are undoubtedly trustable but it generally takes time. Automatic techniques demands manual validations and are dependent on the corpus availability in the respective domain. Manual annotation technique required a large number of annotators to balance one's sentimentality in order to reach agreement. But human annotators are quite unavailable and costly.

But 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. 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.

The lexicons tagged by this system are credible as it is tagged by human beings. In either way it is not like 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.

The game has four types of question templates. For further detailed description the question templates are named as Q1, Q2, Q3 and Q4. To make the gaming interface more interesting images has been added with the help of Google image search API [4] and to avoid biasness we have randomized among the first ten images retrieved by Google. Snapshots of different screens from the game are presented in Figure 1.

[1] http://www.amitavadas.com/Sentiment%20Game/
[2] http://www.espgame.org/
[3] http://www.internetworldstats.com/stats.htm
[4] http://code.google.com/apis/imagesearch/

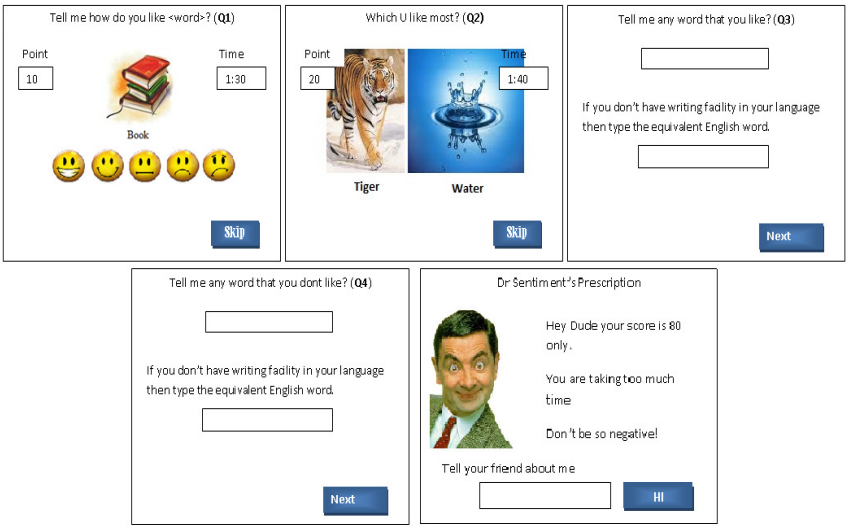

Figure 1: Snapshots from Dr Sentiment Game

2.1 Gaming Strategy

There are four types of questions: 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. There is no thumb rule behind the cardinals rather they are arbitrarily chosen and randomly changed for experimentation. The questions are randomly asked to keep the game more interesting.

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 (Fig 2) to express their sentimentality. The interface keeps log records of each interaction.

Figure 2: 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 1 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.

Now about the technical solution how we did it. 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 log table.

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

Table 1: Relative Sentiment Scores from SentiWordNet

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 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 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.

2.7 CMNT1

Comment type 1 has 5 variations as shown in the Comment table in Table 3. 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)

2.8 CMNT2

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

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

2.9 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 values of accumulating all the above factors.

This is only the appealing factor to a player. The provoking message for players is Dr Sentiment can reveal their sentimental status: whether they are extreme negative or positive or very much neutral or diplomatic etc. 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 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.

We are not demanding that the revealed status of a player by Dr Sentiment is exact or ideal. It is only to make fun but the outcomes of the game effectively help to store human sentimental psychology in terms of computational lexicon.

3 Senti-Mentality

PsychoSentiWordNet gives a good sketch to understand the psycho-sentimental behavior of 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

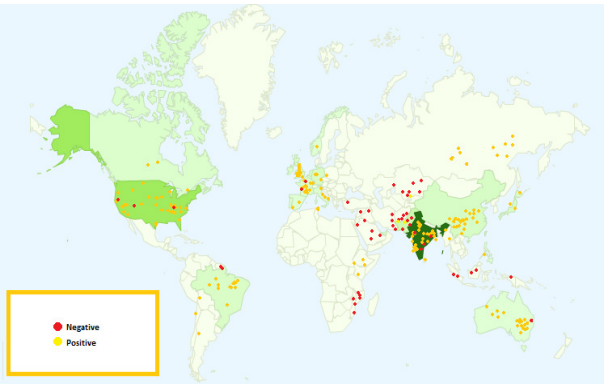

Figure 3: Geospatial Senti-Mentality

The word "***blue***" get 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 3 explains the situation. The observation is that most of the negative tags are coming from the middle-east and especially from the Islamic countries. We found a line in Wiki[5] (see in Religion Section) that 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". But other explanations may be there for this. This is an interesting observation that supports the effectiveness of PsychoSentiWordNet. This information could be further retrieved from the developed source by giving information like (blue, Italy), (blue, Iraq) or (blue, USA) etc.

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

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

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 top of every bar. In the Figure 4 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 that variation of sentimentality with age. This information could be further retrieved from the developed source by giving information like (X, 36-39) or (X, 45-49) etc.

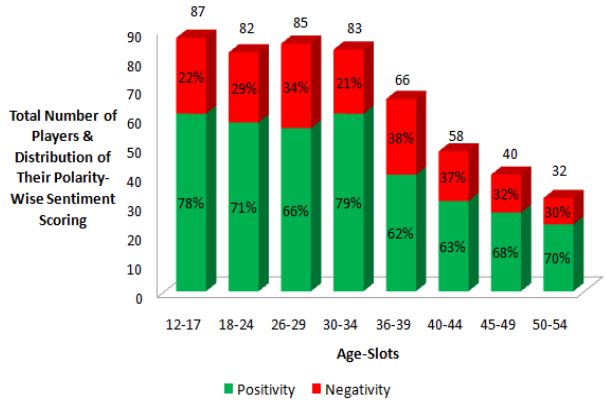

Figure 4: Age-Wise Senti-Mentality

3.3 Gender Specific

It is observed from the statistics collected that women are more positive than a man. The variations in sentimentality among men and women are shown in the following Figure 5.

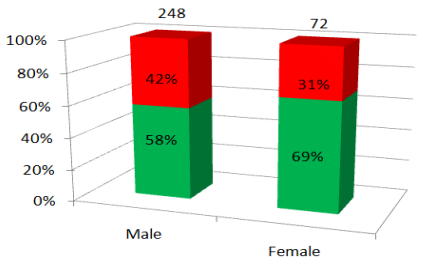

Figure 5: Gender Specific Senti-Mentality

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 are important resource for sentiment/opinion or emotion analysis task. Moreover the other non linguistic psychological 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.

5 The Data Structure and Organization

Deciding about the data structure of this kind of special requirement was not trivial. Presently RDBMS (Relational Database Management System) has been 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 thus we decided to generate web service API by that people could access latest prior polarity scores. We do believe this method will over perform than a static sentiment lexicon set.

6 Conclusion & Future Direction

In the present paper the development of the PsychoSentiWordNet 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 meets psychology.

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

Acknowledgments

The work reported in this paper was supported by a grant from the India-Japan Cooperative Program (DST-JST) 2009 Research project entitled "***Sentiment Analysis where AI meets Psychology***" funded by Department of Science and Technology (DST), Government of India.

References

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.

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

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.

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

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

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

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

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.

Wiebe Janyce and Riloff Ellen. Creating Subjective and Objective Sentence Classifiers from Unannotated Texts. In the Proc. CICLING, Pages 475–486, 2006.

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.