ISSN: 2321-7782 (Online)

Volume 3, Issue 9, September 2015

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

Domain based Opinion Lexicon Generation for Sentiment Analysis of the Academic Resources

Archana Shukla[1]
Computer Science & Engineering Technology
Sham Higginbottom Institute of Technology Allahabad
Allahabad, India

Hari Mohan Singh[2]
Computer Science & Engineering Technology
Sham Higginbottom Institute of Technology Allahabad
Allahabad, India

Abstract: In this paper, we have design and augment an existing sentiment lexicon dictionary in the context of academic domain to provide academic community members facility to judge the quality of research articles by analyzing whether it has positive sentiments or negative sentiments. We have also developed a system which is able to identify opinionated words and phrases from the documents based on online dictionary such as WordNet. Our system uses SentiWordNet to assign sentiment scores to each word found in the document. Sentiments of words are assigned three sentiment scores: Positivity, Negativity and Objectivity with a word and lies in between the range of [0-1]. Our system provides user friendly interface to enable any member of the community to contribute in opinion lexicon generation by providing sentiment words and their polarity values.

Keywords: Sentiment Analysis, Sentiment words, Literature survey, Research community, Opinion Lexicon.

I. INTRODUCTION

Students enrolled in research based degree programs are required to learn research methodology and to contribute in the body of the knowledge and to apply innovative applications of existing knowledge which make a significant impact at national or international level. They may have to go through course work, do literature survey and present periodically the progress of their works through seminars. Out of these activities, literature survey accounts for the major part of the duration of the degree program. The most difficult part of masters or doctoral degree program course arguably understands the research paper while identifying problem area. Students of these programs are required to perform research activity related to their domain of interest. They collect research papers and other publications either from web sites of professional societies likes IEEE, ACM etc. of from printed copy of journal available in the library. While going through these publications, they write their notes, observations, and remarks, questions either on the same document or on the separate sheet of paper. These comments/ observations may be about entire paper or part of them. These comments/ observations are very valuable knowledge resource not only for the current reader but also for the future generation which helps in decision making to judge the quality of publications, as they are not available in electronic form and are not sharable.

Our work is motivated by desire to develop a Sentiment Lexicon dictionary for academic domain which facilitates research community members to judge the quality of papers whether it contains positive sentiments or negative sentiments. It may also reduce the time taken in literature survey. We have developed a system which analyzed two hundred twenty five research papers to identify opinionated words. These words consist of adjective, adverb, verbs etc. These opinionated words are only those words which have weight higher than the average weight of all words found in the analysis. Our system uses SentiWordNet [21] to assign sentiment scores to each word found in the document. Sentiments of words are assigned three sentiment scores: Positivity, Negativity and objectivity with a word and lies in between the range of [0-1]. After analyzing, system highlights all the words which may express the opinion in the web document. Manual inspection is carried out by three authors to verify whether the word is sentiment word or not. Our system discard those words which were not identifies as sentiment word. It gives average score for the words which have already assigned a sentiment score by other users. Our system provides user friendly interface to enable any member of the community to contribute in opinion lexicon generation by providing sentiment words and their polarity values.

This paper is organized into six sections. Second section presents the related work. Section third describes the architecture of the system and the algorithm for the extraction of sentiment words and computation of its polarity values. Section 4 presents the details description about the tool.

II. Related Work

Several research efforts have been made by the researcher related to our work in different domains such as business, news, stock finance, Movie etc. Lun- Wei Ku et al [1] and Jiang YANG et al [2] had used generalized dictionary . Former used *General Enquirer* (GI) , *Chinese Network Sentiment Dictionary containing* 2,764 positive and 7, 778 Negative words whereas later used four pre defined dictionary such as *Positive Word Dictionary*, *Negative Word Dictionary*, *A student positive and negative word Dictionary*, *HowNet.* polarity value were assigned by computing average score based on term frequency of word. Positive and Negative value of words were assigned manually by annotators. [3, 4, 5] extracted the sentiment words consisting of adjectives or adverbs or adjective –adverb both. They proved that subjectivity of sentence could be judged according to the adjectives/adverbs in it. Polarity values of each word were assigned by calculating the probability based on term frequency of word. Jeonghee Yi etal and [6, 7] also used generalized dictionary such as *GI, WordNet* consist of 3000 sentiment entries including about 2500 adjectives and less than 500 nouns. They assigned sentiment value to each extracted word based on the number of times it appears in the whole document. They assigned sentiment polarity using sentiment lexicon database. (word, pos, sentiment category). Jaap Kamps et al [8] and J. Kamps and Marx.M [9] used the WordNet Lexical Database to determine the semantic orientation of a word. For a given word, they look at its semantic distance from “good” compared to it semantic distance from ‘bad” based on WordNet. Janice Weibe et al [10], Soo-Min Kim et al [11], and Minquing Hu et al[12] had constructed dictionary initially using small seed words containing verbs (23 positive and 21 negative) and adjectives (15 positive and 19 Negative) and expanded later based on WordNet extracting synonyms and antonyms for adjectives and only synonyms for verbs. Sentiment score was assigned by computing the probability of the word based on the count occurrences of the word synonyms in the dictionary. Hazivasiloglou and McKeown [13, 1997] collected data from Wall Street Journal which contains 21 million words with respect to the news Domain. The technique started with a list of seed opinion adjective words with their orientation and a set of linguistic constraints or conventions on connectives to identify additional adjective opinion words and their orientations. The technique started with a list of seed opinion adjective words with their orientation and a set of linguistic constraints or conventions on connectives to identify additional adjective opinion words and their orientations. One of the constraints was about (AND), which says that conjoined adjectives usually have the same orientation Robert p. Schumaker etal [14] had collected the financial stock related corpus from PRNews Wire, Yahoo Finance, Associated press, Financial Times etc. They had use NLP Processing to find the noun as a sentiment word. Preexisting system were used to compute polarity value of sentiment words. Maite Taboada et al [15] collected the data of 400 reviews from the epinions web sites. They had collected only adjectives and built dictionary consisting of 1, 719 adjectives. They were assigned the sentiment scores +1 and -1 based on the existing dictionary*(GI*). Minquing Hu et al[16] had collected the customer reviews from the Amazon site and c.net site about products. They were built a dictionary with predefined seed set which include only 30 words consisting of only adjectives. They had used NLP processor to generate the part of speech tags. They had used the effective method by utilizing the adjective synonym set and antonym set in WordNet. If antonym/synonym of the given adjective has the known orientation, then the orientation of the given adjective could be set correspondingly. Li Zhuang et al [17] constructed a dictionary for Movie domain. From the 1100 manually labeled reviews, first collect the 100 positive and negative words which was having the highest frequency as a seed words. Then find the synsets of each words. If any of the synsets were present in the seed words, it was put into the final opinion list. Farah Benamara et al [18] collected the corpus from the 200 annotated set of news articles and extracted the adverb-adjective combination (AAC) words as sentiment units. They had assigned score +1 and -1 based on the algorithm. In [19, 20, and 21], the data are collected from the Chinese reviews web sites which contain feedback regarding the products, movies and services. Data were parsed and tag the part of speech using Chinese Words Segmentation developed by ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System). Extracted two-word phrases conforming to certain patterns. Select the RWP (Reference Word Pair) to present extremely positive and extremely negative opinions to determine the semantic orientation of phrase using the following method. Calculate the average SO values of all the extracted phrases in a review to determine it semantic orientation. The opinion will be judged as positive, if its average semantic orientation exceeds a threshold and is negative otherwise.

All the above works have been conducted in context other than research community. Our work on the other hand, focuses on the research communities and their aims at providing a domain specific lexicon for sentiment analysis of research articles whether it contain positive sentiment or negative sentiments.

III. Architecture of the System

We have developed an application using 3-layer architecture as shown in Fig.1. The top most layer is the presentation layer, which manages all the interaction to end user. The middle layer is the application logic layer which includes all the functionalities such as *text extractor module*, *sentiment word extractor module*, *SentiWordNet* and *WordNet* which are used to manage knowledge resources. The bottom layer is the database layer and contains the database for document, document Metadata and sentiment words.

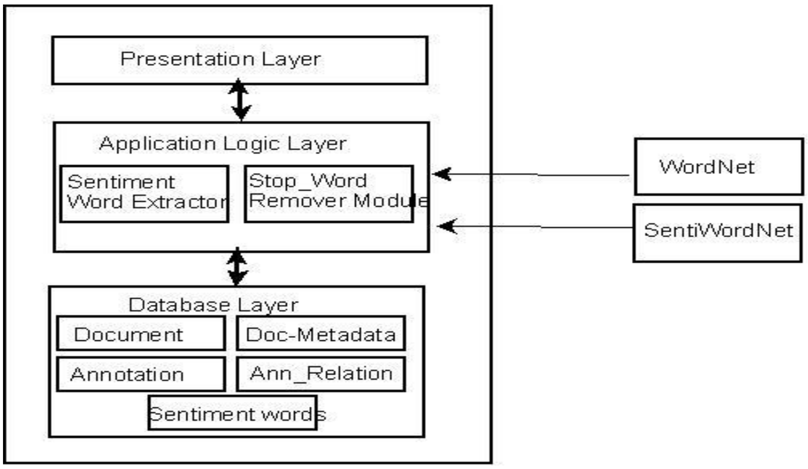


Fig.1. Three Layer Architecture of System

Text Extractor Module: This module extracts the content of document using PDF BOX API such as *getDocumentCatalog()* for extracting the page information. Total number of pages in the PDF document and their count is listed using *getAllPages()* and *size()* function respectively..

Stop Word Removal Module: This module removes stop words and performs stemming using one of the module and consider adjectives, nouns, adverbs and verbs based on POS tagging using *WordNet*. It assigns polarity values in terms of positive, negative and objective using *SentiWordNet*.

SentiWordNet, a lexical resource in which each WordNet synset s is associated to three numerical scores Obj(s), pos(s), neg(s). It is based on the quantitative analysis of the glosses associated to synsets, and on the use of the resulting vectorial term representations for semi-supervised synset classification. These three scores are derived by combining the results produced by a committee of eight ternary classifiers, all characterized by similar accuracy level but different classification behavior.

SentiWordNet is useful tool for sentiment analysis because of its wide coverage (all WordNet synsets are tagged according to each of the three labels Objective, Positive, Negative) and because of its fine grain, obtained by qualifying the labels by means of numerical scores.

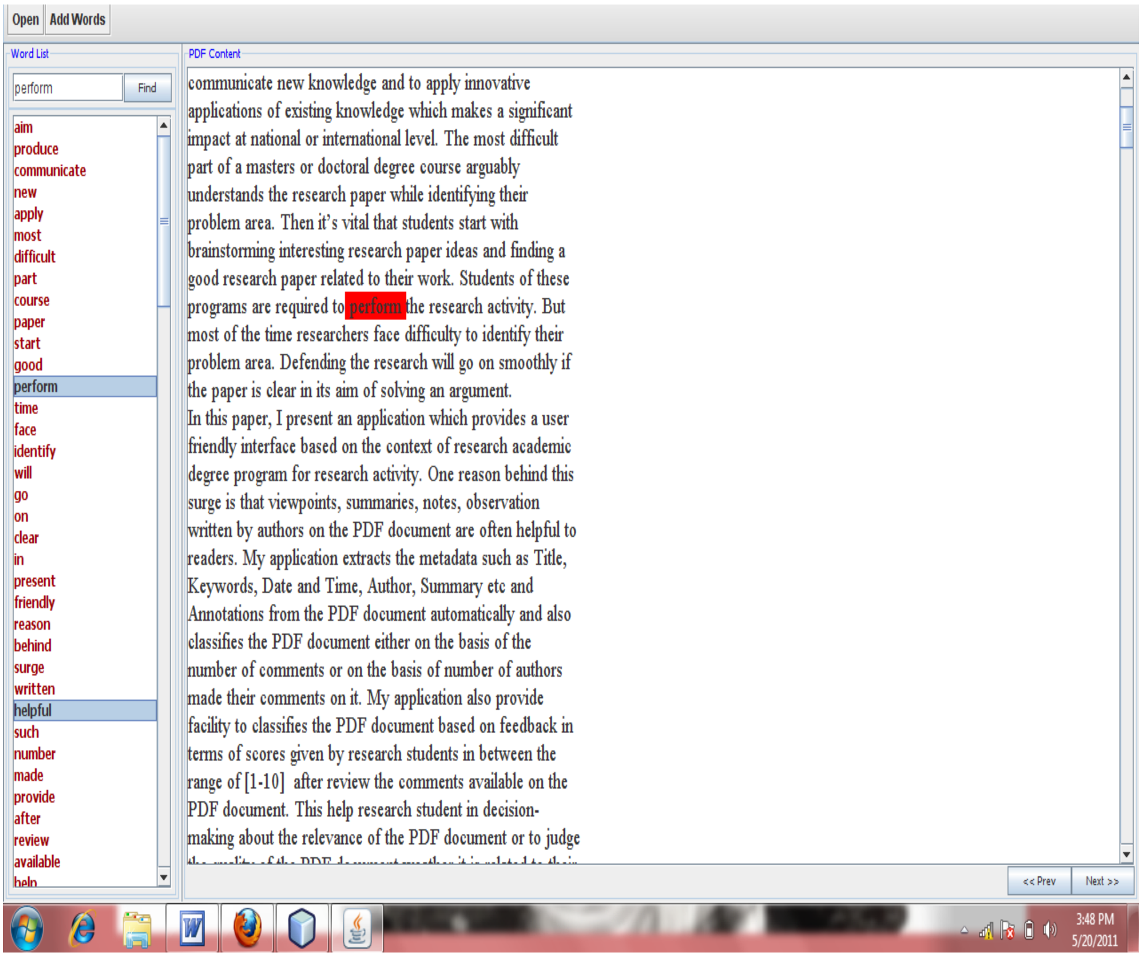
Sentiment scores lie in between the range of [0.0-1.0]. At the time of assignment of scores, our tool also takes care of negation words such as "Not", "Never". If these words are found before any other word (*Adj*), then it interchanges +ve and –ve polarity values of that word which comes after "Not".

IV. Algorithm for Opinion Lexicon Generation

- Extraction of keywords containing adjectives, adverb, verbs and nouns from the corpus available after removing stop words.
- Compute weight for each word based on Term frequency (TF) and Inverse term frequency (ITF).
- Dictionary contains only those words which have weight higher than average weight of all words extracted from the corpus.
- Assign sentiment polarity value and sentiment orientation for each keyword using SentiWordNet.
- Compute sentiment polarity value for phrases which is not available in SentiWordNet by taking average of the sentiment value of the words present in that phrase.

V. User Interface

We have developed a system which is able to identify adjectives, adverbs, verbs and noun from the web document as well as from the research paper automatically based on the online dictionary e.g. WordNet. After analyzing, system highlights all the words which may express the opinion in the web document as shown in Fig.2.



perform

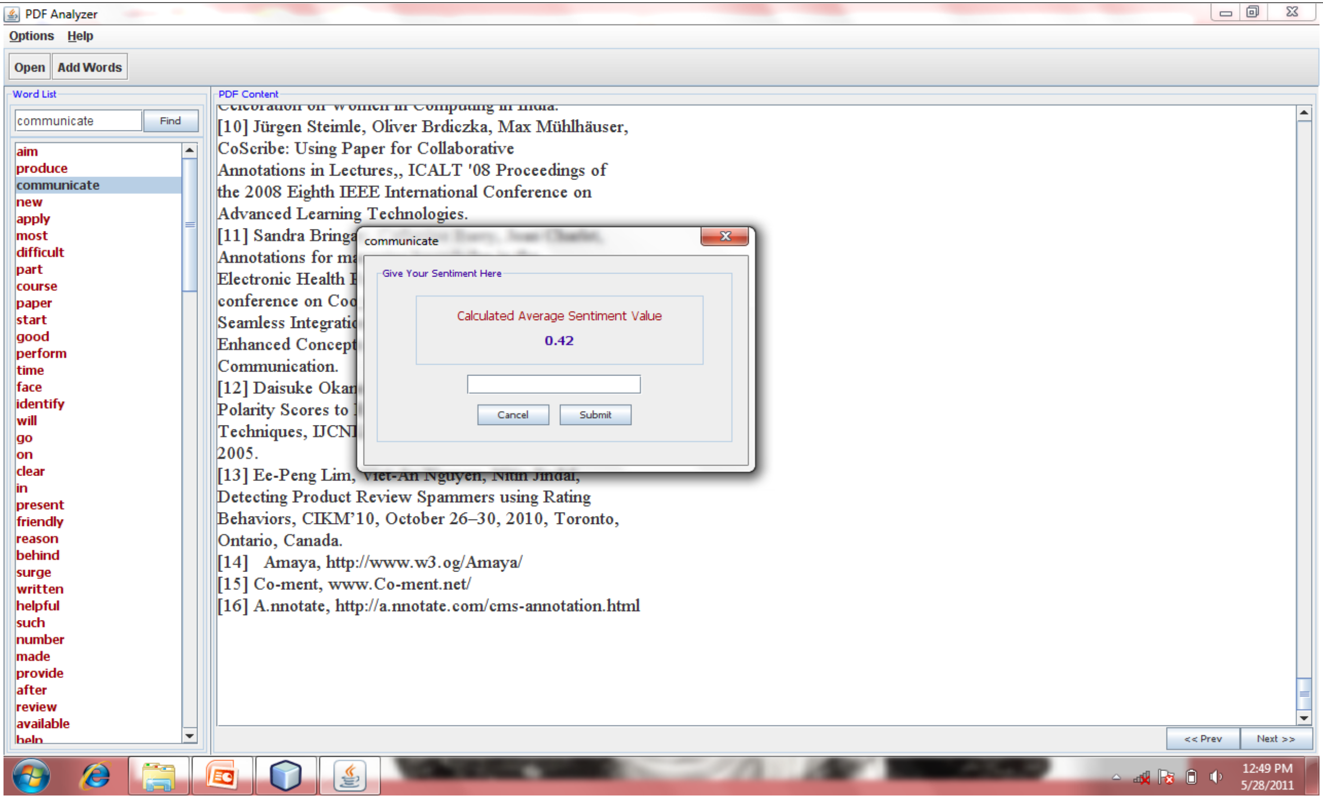
aim
produce
communicate
new
apply
most
difficult
part
course
paper
start
good
perform
time
face
identify
will
go
on
clear
in
present
friendly
reason
behind
surge
written
helpful
such
number
made
provide
after
review
available
help

communicate new knowledge and to apply innovative applications of existing knowledge which makes a significant impact at national or international level. The most difficult part of a masters or doctoral degree course arguably understands the research paper while identifying their problem area. Then it's vital that students start with brainstorming interesting research paper ideas and finding a good research paper related to their work. Students of these programs are required to perform the research activity. But most of the time researchers face difficulty to identify their problem area. Defending the research will go on smoothly if the paper is clear in its aim of solving an argument.

In this paper, I present an application which provides a user friendly interface based on the context of research academic degree program for research activity. One reason behind this surge is that viewpoints, summaries, notes, observation written by authors on the PDF document are often helpful to readers. My application extracts the metadata such as Title, Keywords, Date and Time, Author, Summary etc and Annotations from the PDF document automatically and also classifies the PDF document either on the basis of the number of comments or on the basis of number of authors made their comments on it. My application also provide facility to classifies the PDF document based on feedback in terms of scores given by research students in between the range of [1-10] after review the comments available on the PDF document. This help research student in decision-making about the relevance of the PDF document or to judge

Fig.2. Sentiment Analyzer

Manual inspection is carried out by three authors to verify whether the word is sentiment word or not. Our system discard those words which were not identifies as sentiment word. It gives average score for the words which have already assigned a sentiment score by other users.



communicate

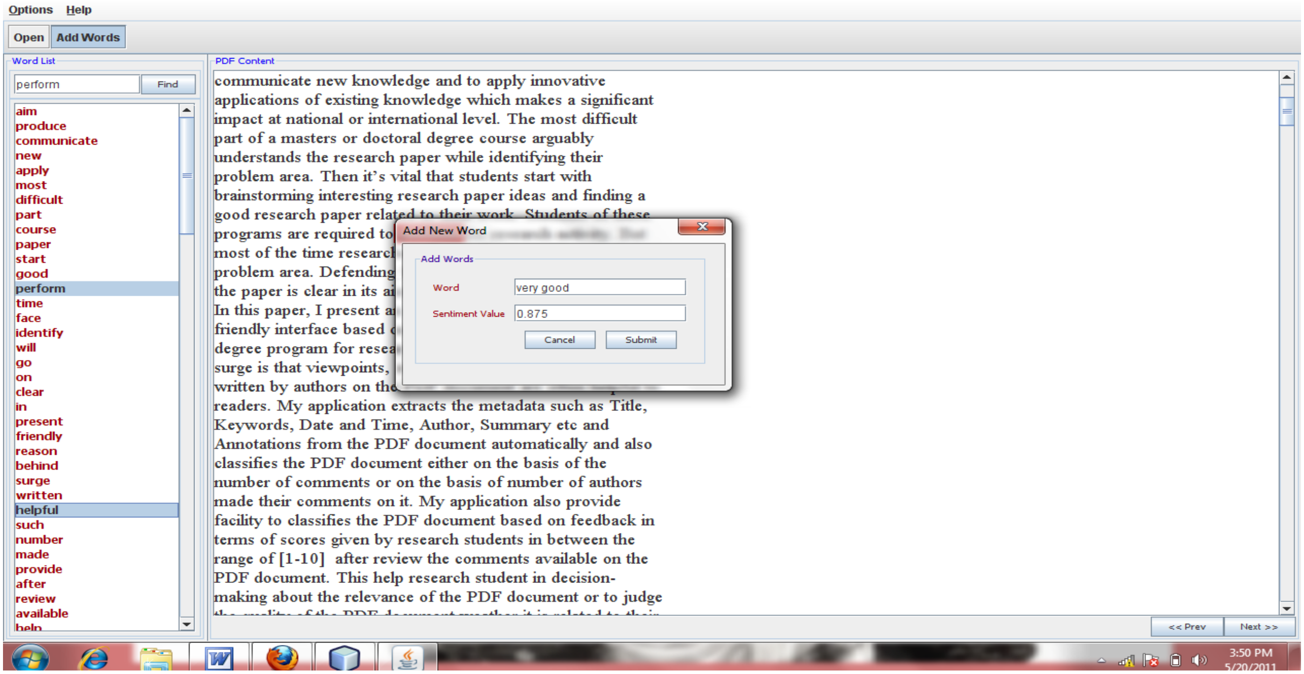
Calculated Average Sentiment Value

0.42

[10] Jürgen Steimle, Oliver Brdiczka, Max Mühlhäuser, CoScribe: Using Paper for Collaborative Annotations in Lectures,, ICALT '08 Proceedings of the 2008 Eighth IEEE International Conference on Advanced Learning Technologies.
[11] Sandra Bringa
Annotations for ma
Electronic Health I
conference on Coo
Seamless Integratio
Enhanced Concept
Communication.
[12] Daisuke Okan
Polarity Scores to
Techniques, IJCNI
2005.
[13] Ee-Peng Lim,
Detecting Product Review Spammers using Rating Behaviors, CIKM'10, October 26–30, 2010, Toronto, Ontario, Canada.
[14] Amaya, http://www.w3.og/Amaya/
[15] Co-ment, www.Co-ment.net/
[16] A.nnotate, http://a.nnotate.com/cms-annotation.html

Fig.3. Average Score of Sentiment Word

Our system provides user friendly interface to enable any member of the community to contribute in opinion lexicon generation by providing sentiment words and their polarity values. It also show the average value of the sentiment score if in the next iteration , user give their polarity value on which already values are given. The information is also supplemented by searching the World Wide Web for additional information regarding the sentiment words, polarity values and their orientation. The iterative process stops when no more new words are found.



perform

Word: very good

Sentiment Value: 0.875

communicate new knowledge and to apply innovative applications of existing knowledge which makes a significant impact at national or international level. The most difficult part of a masters or doctoral degree course arguably understands the research paper while identifying their problem area. Then it's vital that students start with brainstorming interesting research paper ideas and finding a good research paper related to their work. Students of these programs are required to
most of the time researc
problem area. Defending
the paper is clear in its ai
In this paper, I present a
friendly interface based
degree program for resea
surge is that viewpoints,
written by authors on the
readers. My application extracts the metadata such as Title, Keywords, Date and Time, Author, Summary etc and Annotations from the PDF document automatically and also classifies the PDF document either on the basis of the number of comments or on the basis of number of authors made their comments on it. My application also provide facility to classifies the PDF document based on feedback in terms of scores given by research students in between the range of [1-10] after review the comments available on the PDF document. This help research student in decision-making about the relevance of the PDF document or to judge

Fig.4. Shows addition of new Sentiment Word

VI. Conclusion

We have developed an application using java server programming language to generate a domain specific opinion lexicon for the purpose of sentiment analysis of the document which provide assistant to research scholars during various activities such as literature survey task, writing their own research articles etc. Our tool also provides facility to research community to query knowledge base to get the sentiment of the document whether it is positive or negative. We believe that it is helpful to research community.

References

1. Lun-Wei Ku, Yu-Ting Liang and Hsin-His Chen , Opinion Extraction, Summarization and Tracking in News and Blogs Corpora. In Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs (2006).
2. Jiang YANG, Min HOU, Ning WANG, “ Recognizing Sentiment Polarity in Chinese Reviews Based on Topic Sentiment Sentences”, Natural Language Processing and Knowledge Engineering (NLP-KE), 2010 International Conference, Beijing.
3. Sara Owsley , Sanjay Sood , Kristian J. Hammond . Domain Specific Affective Classification of Documents. In Proceeding of American Association of Artificial Intelligence, AAAICAAW’06.
4. Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell and Melanie Martin, Learning Subjective Language. In the proceeding of Association for Computational Lingusitics, 2004
5. Weibe, J.M. 2000, Learning Subjective Adjectives from Corpora. Proceedings of the 17th National Conference on Artificial Intelligence. Menlo Park, CA: AAAI Press.
6. Jeonghee Yi, Tetsuya Nasukawa, Razvan Bunescu and Wayne Niblack, Sentiment Analyzer: Extracting sentiments about a Given Topic using Natural Language processing Techniques. Third International Conference on Data Mining ICDM, 2003
7. Kamps J., & Marx,M. 2002. Word with Attitude. Proceeding of the first International Conference on Global WordNet, CIIL, Mysore, India, 332-341.
8. Kamps, J., marx, M., R.J., & de Rijke, M. (2004) Using WordNet to Measure Semantic Orientation of Adjectives. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04*, Vol. IV (2004), pp. 1115-1118.
9. Kim, S-M., & Hovy, E.H. (2004). Determining the Sentiments of Opinion. Proceedings of 20th International Conference on Computational Linguistics, pp. 1367-1373.
10. Janyce Wiebe, Ellen Riloff. Creating Subjective and Objective Sentence Classifiers from Unannotated Text. In the proceeding of CICLing (2005), pp. 486-497.
11. Kim, S-M., & Hovy, E.H. (2004). Determining the Sentiments of Opinion. Proceedings of 20th International Conference on Computational Linguistics, pp. 1367-1373.
12. Hu, M. and B. Liu, 2004. Mining and Summarizing Customer Reviews. Proceeding of the Tenth ACM SIGKDD Internatational Conference of Knowledge Discovery and Data , Aug. 22-25, ACM Press, Washington, USA., pp: 168-177
13. Hatzivassiloglou, V., & McKeown, K.R. 1997. Predicting the Semantic Orientation of Adjectives. Proceedings of the 35th Annual Meeting of the ACL and the 8th Conference of the European Chapter of the ACL (pp- 174-181). New Brunswick, NJ: ACL
14. Robert P. Schumaker, Yulei Zhang, Chun-Neng Huang, Sentiment Analysis of Financial News Articles.
15. Maite Taboada, Caroline Anthony and Kimberly Voll , Methods for Creating Semantic orientation Dictionaries. In Proceedings of Fifth International Conference on Language Resources and Evaluation (LREC 2006). Genoa, Italy. pp. 427-432.
16. Li Zhuang, Feng Jing, Xiao-Yan Zhu, Movie Review Mining and Summarization. In Proceedings of the 15th ACM International Conference on Information and Knowledge Management (2006), pp. 43-50.
17. Farah Benamara, Carmine Cesarano, Diego Reforgiato, Sentiment Analysis: Adjectives and Adverb are better than Adjectives Alone. In the proceeding of ICWSM, 2007, Boulder, CO, USA.
18. Li Shi , Yang Jun-zuo , Li Yi jun and Ye Qiang , An Experimental Research on Sentiment classification of Chinese Reviews by Semantic orientation Method. In the procedding of Control and Decision Conference, 2008. CCDC 2008. Chinese, 2008. CCDC 2008, pp. 3999 – 4004.
19. Sentiment classification of movie and product review using contextual valence shifters.
20. Pang B., Lee L., & Vaithyanathan.S, 2002. Thumps up? Sentiment Classification using Machine Learning Technique. Proceeding of the 2002 Conference on Empirical Methods in Natural language Processing. 79-86
21. http://sentiwordnet.isti.cnr.it/