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.
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 its 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 opinions 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 its 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](image)

**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, adverbs, 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.

![Sentiment Analyzer](image)

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. 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.

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


4. Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell and Melanie Martin, Learning Subjective Language. In the proceeding of Association for Computational Linguistics, 2004


17. Farah Benamara, Carmine Cesarono, Diego Reforgiato, Sentiment Analysis: Adjectives and Adverb are better than Adjectives Alone. In the proceeding of ICWSM, 2007, Boulder, CO, USA.


