ISSN:0975-9646

# A Text Polarity Analysis Using Sentiwordnet Based an Algorithm

Deepak Singh Tomar
*VITM,Gwalior(M.P.)*

Pankaj Sharma
*Asst.Professor,VITM,Gwalior*

*CSE Department,*
*RGPV University, India*

***Abstract-*** **sentiment analysis aim to get the underlying viewpoint of text which could be opinion, online review, movie rating comments etc. the aim of this project is to offer better sentiment text analysis strategy which recognize the polarity of text message including positive, negative, and neutral using sentiwordnet. The contribution of this paper is use POS(parts of speech) tagger to examine specific prior polarity of text. Polarity analysis has been an important subtask in sentiment analysis but detecting correct polarity has been a major issues. Polarity enhancers and negations detection effect the overall polarity in an abnormal way. So we cannot depend on polarity of a particular word alone for accurate results. This paper discuss all the possible approaches available for accurate sentiment analysis and also the issues in detecting correct polarity of a sentence.**

**Keywords-sentiment analysis, sentiwordnet, POS tagger, Polarity**

## I. INTRODUCTION

Natural language processing (NLP) is the computerized approach to analyzing text that is based on both a set of theories and a set of technologies. In other words NLP is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring text at one or more levels of linguistics analysis for the purpose of achieving human-like language processing for a range of tasks or application. in NLP [13] sentiment analysis is an important technique to recognize the polarity of text opinion (positive, negative, neutral, both). A typical approach to sentiment analysis is to start with a lexicon of positive or negative words and phrases. in these lexicons, entries are tagged with their priori prior polarity. Opinion polarity can be understood with the help of example.

1.Asian observer “*generally approved*$^{+}$” of his victory while European government “*denounced*$^{-}$” it.

2. We “*don’t hate*$^{+}$ ” the sinner he says, but we “*hate*$^{-}$“ sin.

In this paper we are focusing on polarity of text with the help sentiwordnet based algorithm. Use pos tagger like adjectives adverb etc. to recognize each word polarity. Contribution of each word polarity for analysis can be produce variation in result so we will do a complete sentence polarity test with the help of sentiwordnet algorithm. Polarity can be represent by the amount of numeric value.

*Various features related to sentiment analysis are:*

**Intensity:** It refers to the strength of the private state that is being expressed, in other words, how strong is an emotion or a conviction of belief. As language users, we intuitively perceive distinctions in the intensity levels of different private states. For example, outraged and extremely annoyed are more intensely negative than irritated. Recognizing intensity includes not only identifying private states of different intensity, but also detecting the absence of private states. Thus, recognizing intensity subsumes the task of distinguish between subjective and objective language.

**Polarity:** The term polarity has a number of different uses, but in this dissertation it is used primarily to refer to the positive or negative sentiment being expressed by a word. However, there is an important distinction between the prior polarity of a word and its contextual polarity [10]. The prior polarity of a word refers to whether a word typically evokes something positive or something negative when taken out of context. For example, the word beautiful has a positive prior polarity, and the word horrid has a negative prior polarity. The contextual polarity of a word is the polarity of the expression in which the word appears, considering the context of the sentence and the discourse. Although words often do have the same prior and contextual polarity, many times the word’s prior and contextual polarities differ. Words with a positive prior polarity may have a negative contextual polarity, or vice versa For example, in sentence 1 the word “*denounced*” has negative and approved has a positive polarity while in sentence 2 don’t hate has positive and hate has negative polarity according to sentiwordnet based algorithm.

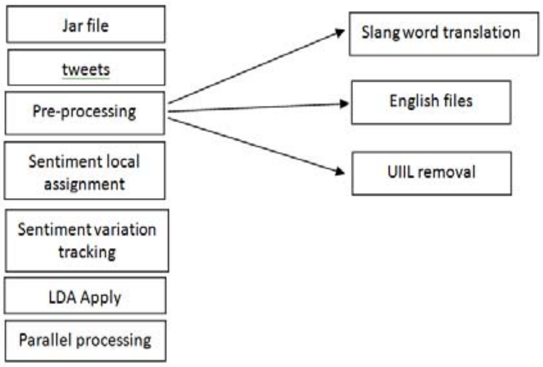
Opinion mining [11] is necessary part of sentiment analysis which facilitate to identify the nature of sentence like positive or negative polarity of text . in modern days several social sites are running , review’s[9] and feedback are given by user on sites in the form of comment, pos etc. these comments are sense positive or negative text. This sense will help to find out the accurate result with the help of sentiment analysis. This analysis can be done by English word phrases, algorithm and application. LDA approach is useefull to identify the contextual polarity text.

## II. LITERATURE SURVEY

Researchers have taken a keen interest in sentiment analysis for the last few years. It has attracted a great deal of attention because of its challenging research problems and the wide range of applications for both academia and industry. It needs a computational study for extracting knowledge from the people’s opinions, appraisals and emotions toward entities, events and their attributes. In today's international global world market and highly growing internet usage, people prefer online shopping, banking, ticket reservation, hotel booking, etc. So sentiment analysis [12] from online customer reviews is becoming a requirement of an organization, customer and also manufacturer. Different researchers have been working on different aspects of this area. The existing work on sentiment analysis can be categorized into document, sentence and word/feature level classification.

In [1] sentence level opinion polarity prediction is done by assigning lexical polarities and deriving sentence polarity from these with the use of contextual valence shifters. A methodology for iterative failure analysis is developed and used to refine our lexicon and identify new contextual shifters. To evaluate the patterns, they explore the idea of measuring their consistency and strength using metrics of polarity like the log-likelihood and SO-A equations. One idea for doing this is to use the SO-A equation on a series of n-grams constructed by taking a given pattern and substituting in different lexicon words. By doing this we would like to be able to rank the patterns by consistency (how consistently they shift the polarity in the same direction) and strength (how far they shift the polarity). The drawback of this approach is they have not included but-conjunction, adverbs of excess and sufficiency, and hedging in more detail also more complex rhetorical devices such as idiom and irony are not included in their work.

## III. Previous Work



So much research has been done in different areas over the past decade. Classification and summarization are used part of opinion mining. LDA based approach used to contribute accurate result of text opinion. To find contextual polarity one by one word analyze in the form of adjective, noun, adverb etc and perform preprocessing . in opinion also used slang word which are also handled by the preprocessing stage of LDA based model. Feature-based Sentiment classification on the other hand considers the opinions on features of certain objects. For example, in reviews related to laptops classifying the Sentiments only on the basis screen quality. In one of the pioneer work [2], the authors present a method Of subjective identification for sentiment analysis based on Minimum cuts. This is important because the irrelevant data from the reviews could be eliminated. The problem is viewed as a classification task and different types of supervised learning techniques have been used in this field. Some of the most common ones are naive Bayes classifier, Support Vector Machine [3], Maximum Entropy [4] etc. Even some graph based techniques [5] are also used.

Sentiments are the words or sentences that represent view or opinion that is held or expressed that can be positive, negative or neutral. To analyze public sentiment variations and find possible reasons behind these variations, for this propose Latent Dirichlet Allocation (LDA) based models: Foreground and Background LDA (FB-LDA).To extract tweets associated with the target, it will undergo the complete dataset and extract all the tweets that contain the keywords of the target. Compared with regular text documents, tweets square measure usually less formal and infrequently written in manner. Sentiment analysis tools applied on raw tweets typically come through terribly poor performance in most cases. Therefore, preprocessing techniques on tweets square measure necessary for getting satisfactory results on sentiment analysis. This paper uses Thayer's Model of human emotion [5] to Classify text. This two dimensional approach adopts the Theory that human emotion can be obtained by: Stress (negative polarity/positive polarity) and Energy(low intensity/high intensity), and divides it into four broad classes: Satisfied, Sad, Exuberant and Angry. Two binary classifiers were trained. First was trained to get The polarity (positive or negative) of the text. While the Second was trained on intensity (low or high) of the text. Figure illustrates the approach. Polarity Some existing lexicon resources like Sentiwordnet 3.0[6] and General Inquirer [7] were used to extract some features from the text. These features were trained using support vector machine, to predict the binary class label.

## IV. Proposed Technique

To sense the opinion nature and find out the accurate contextual polarity of text , we have to use some systematic technique so that each processing or analysis can be done step by step . this research paper try to evaluate better polarity test of opinion or sentence by the sentiwordnet based an algorithm which facilitate to identify the opinion nature (positive or negative). Intensity and pos tagger are performed important role to find out the accurate polarity.these pos tagger are adjective, adverb, noun etc which have their own specific abbreviation so each word of sentence are specified by their abbreviation according to its tagger and follow the Boolean concept{(+ve)(-ve)=(-ve), (+ve)(+ve)=(+ve),(-ve)(-ve)=(+ve)}to find out the contextual polarity of sentence. In this technique if all pos tagger are in even count then it will treated as +ve and odd treated as a –ve. Both are summarized to evaluate the whole sentence polarity.

In this research work we try to better evaluate the polarity of a sentence by the use of "Sentiwordnet", a lexical resource for sentiment analysis. We take the help of

Stanford POS (part of speech) tagger to tokenize our sentence. We then select only that part of speech which could effect the polarity of a sentence or in other words polar words. We then propose our algorithm to find the overall polarity of the sentence also including those words which could improve, inverse or decrease the polarity of the corresponding word.

## POS Tagger:

A Part-Of-Speech Tagger (POS Tagger) is a piece of software that reads text in some language and assigns parts of speech to each word (and other token), such as noun, verb, adjective, etc., although generally computational applications use more fine-grained POS tags like 'noun-plural'. The tagger was originally written by Kristina Toutanova. Since that time, Dan Klein, Christopher Manning, William Morgan, Anna Rafferty, Michel Galley, and John Bauer have improved its speed, performance, usability, and support for other languages.

The system requires Java 1.6+ to be installed. Depending on whether you're running 32 or 64 bit Java and the complexity of the tagger model, you'll need somewhere between 60 and 200 MB of memory to run a trained tagger (i.e., you may need to give java an option like java -mx200m). Plenty of memory is needed to train a tagger. It again depends on the complexity of the model but at least 1GB is usually needed, often more.

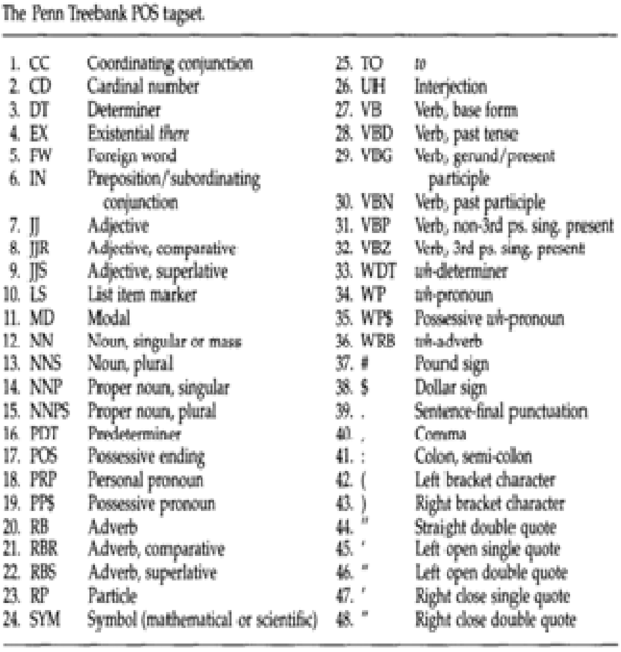
Several downloads are available. The basic download contains two trained tagger models for English. The full download contains three trained English tagger models, an Arabic tagger model, a Chinese tagger model, and a German tagger model. Both versions include the same source and other required files. The tagger can be retrained on any language, given POS-annotated training text for the language.

**Sentiwordnet** : SENTIWORDNET[8] is the result of the automatic annotation of all the synsets of WORDNET according to the notions of “positivity”, “negativity”, and “neutrality”. Each synset s is associated to three numerical scores Pos(s), Neg(s), and Obj(s) which indicate how positive, negative, and “objective” (i.e., neutral) the terms contained in the synset are. Different senses of the same term may thus have different opinion-related properties. For example, in SENTIWORDNET 1.0 the synset [estimable(J,3)] corresponding to the sense “may be computed or estimated” of the adjective estimable, has an Obj score of 1:0 (and Pos and Neg scores of 0.0), while the synset [estimable(J,1)] corresponding to the sense “deserving of respect or high regard” has a P os score of 0:75, a Neg score of 0:0, and an Obj score of 0:25. Each of the three scores ranges in the interval [0:0; 1:0], and their sum is 1:0 for each synset. This means that a synset may have nonzero scores for all the three categories, which would indicate that the corresponding terms have, in the sense indicated by the synset, each of the three opinions related properties to a certain degree.

**Table 1.1 POS type**

| Pos_name | Pos_abbreviation | Sentiwordnet_Abr |
|---|---|---|
| Noun | NN | n |
| Adjective | JJ | a |
| Verb | VB | v |
| Adverb | RB | r |
| Noun | NNS | n |
| Adjectives | JJS | a |

**Table 1.2**



The Penn Treebank POS tagset.

| | | | | | |
|---|---|---|---|---|---|
| 1. | CC | Coordinating conjunction | 25. | TO | *to* |
| 2. | CD | Cardinal number | 26. | UH | Interjection |
| 3. | DT | Determiner | 27. | VB | Verb, base form |
| 4. | EX | Existential *there* | 28. | VBD | Verb, past tense |
| 5. | FW | Foreign word | 29. | VBG | Verb, gerund/present participle |
| 6. | IN | Preposition/subordinating conjunction | 30. | VBN | Verb, past participle |
| 7. | JJ | Adjective | 31. | VBP | Verb, non-3rd ps. sing. present |
| 8. | JJR | Adjective, comparative | 32. | VBZ | Verb, 3rd ps. sing. present |
| 9. | JJS | Adjective, superlative | 33. | WDT | *wh*-determiner |
| 10. | LS | List item marker | 34. | WP | *wh*-pronoun |
| 11. | MD | Modal | 35. | WP$ | Possessive *wh*-pronoun |
| 12. | NN | Noun, singular or mass | 36. | WRB | *wh*-adverb |
| 13. | NNS | Noun, plural | 37. | # | Pound sign |
| 14. | NNP | Proper noun, singular | 38. | $ | Dollar sign |
| 15. | NNPS | Proper noun, plural | 39. | . | Sentence-final punctuation |
| 16. | PDT | Predeterminer | 40. | , | Comma |
| 17. | POS | Possessive ending | 41. | : | Colon, semi-colon |
| 18. | PRP | Personal pronoun | 42. | ( | Left bracket character |
| 19. | PP$ | Possessive pronoun | 43. | ) | Right bracket character |
| 20. | RB | Adverb | 44. | " | Straight double quote |
| 21. | RBR | Adverb, comparative | 45. | ‘ | Left open single quote |
| 22. | RBS | Adverb, superlative | 46. | “ | Left open double quote |
| 23. | RP | Particle | 47. | ’ | Right close single quote |
| 24. | SYM | Symbol (mathematical or scientific) | 48. | ” | Right close double quote |

We propose following algorithm which also uses part of speech tagger(developed by Stanford university) to first tag the whole sentence and then selecting only those candidates which are polar words, or words with possible polarity.

## Proposed Algorithm

1) Create an input file "sample-input.txt" containing 1 or more sentences to check the polarities.
2) Read the file with a file reader object.
3) Parse each sentence token by token with the help of POS tagger.
4) POS tagger will assign a tag to each token.
5) Check the tag of each token if the tag is "JJ" or "JJS" (i.e the tagged token is an adjective/opinion word) then pass this word in SentiWordnet to check the score as well as polarity of that particular word.
6) SentiWordnet will return the sentiment-type of that word (egpositive,weak_positive,strong_positive,negative,strong_negative,neutral etc based on score).
7) Count the no of positive(pos_count) and no of negative (neg_count) adjectives for each sentence.
8) If the neg_count is an ODD number then the sentence is considered "Negative" as a whole.(-)+(+)=(-) else goto step 9.
9) If the neg_count is an EVEN number(consider zero as even) then the sentence is considered "positive" as a whole (-)+(-)=(+) or (+)+(+)=(+).

## V. Result Analysis

Result of review can be evaluated using the sentiwordnet based algorithm and collect customer review data as an opinion. These collection of opinion can be large no. of sentence which have large no. of positive, negative or neutral dataset. So result can be find out using sentiwordnet only on selective dataset.

Evaluation of our proposed method is done as follows, we collected two types of online customer reviews datasets to check the system performance (1) popular publicly available corpus from movie-review polarity dataset v2.0 IMDB movie reviews (http://www.cs.cornell.edu/people/pabo/movie-review-data/) .The data set consists of 1000 positive and 1000 negative reviews in individual text files, and also the sentences polarity dataset (includes 100 positive and 100 negative processed sentences and 50 neutral sentences as a subset of whole database). We performed our experiments on the dataset of about 200 hotel reviews downloaded, which is collected from trip advisor (http://www.tripadvisor.com/) that is one of the popular review sites about hotels and travelling.

**Table 1.3:** Opinion orientation for positive, negative and neutral

| Datasets | Sentiment Orientation | Sentence level accuracy |
|---|---|---|
| Movie Reviews | Positive | 73.2% |
| | Negative | 72.1% |
| | Neutral | 66% |
| Hotel Reviews | Positive | 73% |
| | Negative | 69.8% |
| | Neutral | 60% |

## VI. Conclusion & Future Work

We have proposed a sentiwordnet based algorithm to more efficiently find the polarity of a given sentence. The use of part of speech tagger to tag words and search only those words with polarity (adjectives and adverbs) has increased the performance and we don’t need to remove stop words. Sentiwordnet on the other hand is the main tool for calculating the score of a particular word in our sentence, besides that we also have to consider those words that in a way effect the polarity(inverse or enhance) of a particular. We have added in our algorithm these words. The algorithm perform quite good on normal input sentences chosen at random We evaluate our work few types of customer review datasets. From the results, it is clear that the proposed system achieved an average accuracy of 69.1% at the sentence level.

Future work for this can be an addition of module to check for spelling mistakes beforehand, also the sentiwordnet itself is not enough. Many words are not listed in the database so we need a new lexical resource or create or own as an extension to this. Online reviews also add smiley’s to them which is a language of its own. A single smiley can easily convey the polarity of the whole sentence without bothering, so we can also add a module for inclusion of smiley evaluation.

## References

1. Adam Longton- An empirical analysis of lexical polarity and contextual valence shifters for opinion classification.2003
2. Bo Pang and Lillian Lee. A sentimental education: Sen- timent analysis using subjectivity summarization based on minimum cuts. In Proceedings of the 42nd annual meeting on Association for Computational Linguistics, page 271. Association for Computational Linguistics, 2004.
3. Puneet Singh, Ashutosh Kapoor, Vishal Kaushik, and Hima Bindu Maringanti. Architecture for auto- mated tagging and clustering of song _les according to mood. International Journal of Computer Science Is- sues (IJCSI), 7(4), 2010.
4. Wei Jin. Mining hidden associations in text corpora through concept chain and graph queries. ProQuest, 2008.
5. Robert E Thayer. The biopsychology of mood and arousal. Oxford University Press, 1989.
6. Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In LREC, volume 10, pages 2200{2204, 2010.
7. Philip J Stone, Dexter C Dunphy, and Marshall S Smith. The general inquirer: A computer approach to content analysis. 1966.
8. Aurangzeb Khan, Baharum Baharudin, and Khairullah Khan - Sentiment Classification from Online Customer Reviews Using Lexical Contextual Sentence Structure-, J.M. Zainet al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 317–331, 2011.© Springer-Verlag Berlin Heidelberg 2011
9. Hang Cuiet all-Comparative Experiments on Sentiment Classification for Online Product Reviews, Copyright 2006, American Association for Artificial Intelligence (www.aaai.org). All rights reserved.
10. Recognizing Contextual Polarity: An exploration of features for phrase-level sentiment analysis- Theresa Wilson
11. Dey, L., Haque, S.M.: Opinion mining from noisy text data. Int. J. Document Anal. Recognition 12, 205–226 (2009), http://www.springerlink.com/content/1265305p655l2357/10.1007/s10032-009-0090-z
12. Neviarouskaya, A., Prendinger, H., Ishizuka, M.: Semantically distinct verb classes involved in sentiment analysis. In: Proceedings of the International Conference on Applied Computing (AC 2009), Japan, pp. 27–34 (2009).
13. Choi, Y., Cardie, C.: Learning with compositional semantics as structural inference for subsentential sentiment analysis. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2008), USA, pp. 793–801 (2008), http://portal.acm.org/citation.cfm?id=1613715.1613816