International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

Finding Semantic Orientation of Reviews Using Unsupervised PMI Algorithm

Sneha M Nakade[1], Sachin N Deshmukh[2]

[1]Department of Computer Science & Information Technology,
Dr. Babasaheb Ambedkar Marathwada University,
Aurangabad (MS) India

[2]Department of Computer Science & Information Technology,
Dr. Babasaheb Ambedkar Marathwada University,
Aurangabad (MS) India

Abstract: ***Recent years have shown quick expansion of the social web over the Internet, where individuals can express their opinion on various things, for example, products, persons, subjects, and discussion etc. As e-commerce is quickly developing, item audits on the Web have turned into a critical data hotspot for clients' choice making when they want to purchase items on the web.Sentiment classification of such reviews of individuals generally requires lot amount of training data but availability of labeled data for different domains is generally time consuming and tedious task. This paper present simple unsupervised learning algorithm called Pointwise Mutual Information (PMI) followed by Semantic Orientation (SO). Averaging the semantic orientation of phrase does the classification of user reviews. Phrase with positive semantic orientation is associated with positive sentiment and negative semantic orientation is associated with negative sentiment. If the average semantic orientation of phrases is positive then the review is classified as Positive otherwise Negative.***

Keywords: Pointwise Mutual Information, Unsupervised Algorithm, Semantic Orientation, Sentiment Analysis

1. Introduction

If we want to take a decision, we first prefer to seek others opinion, we evaluate opinions and take decision. Same thing is applied to organizations when they introduce new product or on the way to introduce it; organizations take opinions of its customers in the form of reviews of product on official websites of organization, social media sites such as Facebook, Twitter, Blogs or online shopping sites. Customer also wants to know opinions of existing users before they use service or purchase a product. These reviews help organizations and its customers to evaluate the response or love among people about product or service.

Sentiment analysis, also called opinion mining, is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes. It represents a large problem space. There are also many names and slightly different tasks, e.g., sentiment analysis, opinion mining, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, review mining, etc. [1].

Conventional ways to deal with content order require a lot of training data. Procurement of such data can be exorbitant and tedious. Because of the exceedingly space particular nature of the conclusion grouping assignment, moving starting with one area then onto the next normally requires the securing of another labeled data. For this reason, unsupervised or very weakly supervised methods for sentiment classification are especially desirable [2].

This paper presents an unsupervised learning algorithm for classifying reviews. This algorithm takes text review as input and gives output as whether the review is positive or negative. Firstly it assigns POS tagging to eachterms of review for identifying the adverbs and adjectives. As adverbs and adjectives are descriptors of another wordand modify the meaning of word. These extracted words are called phrases.The second step is to assign semanticorientation of the extracted phrases. The phrase with positivesemantic orientation has good association and the phrase with negative semantic orientation has bad association. Third step is to find whether the review is positive or negative. If the average semantic orientation of phrases is positive then the review is classified as positive review and if the average semantic orientation of the phrases is negative then the review is classified as negative review [3].

2. Related Work

In [11], author presents a basic calculation using unsupervised learning of semantic orientation from great degree huge corpora. A positive semantic orientation suggests attractive quality (e.g., "legit", "fearless") and a negative semantic orientation infers undesirability (e.g., "aggravating", "unnecessary"). The strategy includes issuing inquiries to a web crawler and utilizing pointwise mutual information data to dissect the outcomes. The calculation is exactly assessed utilizing a preparation corpus of around one hundred billion words the subset of the Web that is filed by the picked web crawler. Tried with 3,596 words (1,614 positive and 1,982 negative), the calculation accomplishes an exactness of 80%. The 3,596 test words incorporate descriptive words, intensifiers, things, and verbs.

In [12], author shows a basic unsupervised learning calculation for perceiving equivalent words, in view of measurable information obtained by questioning a web index. The calculation, called PMI-IR, utilizes Pointwise Mutual Information (PMI) and Information Retrieval (IR) to

gauge the similitude of sets of words. PMI-IR is exactly assessed utilizing 80 equivalent word test questions from the Test of English as a Foreign Language (TOEFL) and 50 equivalent word test questions from a gathering of tests forunderstudies of English as a Second Language (ESL). On both tests, the calculation acquires a score of 74%. PMI-IR is appeared differently in relation to Latent Semantic Analysis (LSA), which accomplishes a score of 64% on the same 80 TOEFL questions. The paper talks about potential utilizations of the new unsupervised learning calculation and a few ramifications of the outcomes for LSA and LSI (Latent Semantic Indexing).

3. Proposed Algorithm

3.1 Selection of Bigrams and Feature Phrase Extraction

The very first step of classification using unsupervised PMI learning is to extract the phrases containing adverbs and adjectives in the review. The work of [4], [5], [6] showed the adverbs and adjectives are the good indicators of subjectivity.

Single word adjective and adverbs may have different meaning in different context and they modify the meaning of other word quickly. For Example, the word "great" may have positive orientation in the movie review such as "great acting" and negative orientation in another movie review for "great loss".

So rather than selecting single word adjective or adverb we selected bigrams containing adjective and adverb.Firstly POS tagging is applied to each word of review. Table 1 shows POS tags and their meaning. Two consecutive words are extracted from the review is any pattern matches of Table 2.

Table 1: POS Tags and Meaning

POS Tag	*Meaning*
JJ	Adjective
RB, RBR, or RBS	Adverb
NN or NNS	Noun
VB, VBD, VBN, or VBG	Verb

Table 2: Tag Patterns to Extract Two-Word Phrase fromReview

First Word	*Second Word*	*Third Word*
JJ	NN or NNS	Anything
RB, RBS or RBR	JJ	Not NN or NNS
JJ	JJ	Not NN or NNS
RB, RBR, or RBS	VB, VBD or VBN	Anything

3.2 Selection of Seed Words

To identify the orientation (i.e. positive or negative) of extracted phrase we need to pass some words to our algorithm, called as seed words with positive orientation and negative orientation. If the extracted phrase comes with positive seed words then phrase is associated with positive orientation and if phrase comes with negative seed words then the phrase is associated with negative orientation.

Table 3: List of Positive and Negative Seed Words:

Positive Seed Words	*Excellent, Good*
Negative Seed Words	Poor, Bad, Hate, Suck, Horrible, Terrible

3.3 Finding PMI and SO between Extracted Phrase and Seed Word

The Pointwise Mutual Information measure of extracted phrase with seed words is calculated by (1) given as follows.

$$\mathrm{PMI}(\mathrm{term1},\mathrm{term2}) = \log \frac{\mathrm{P}(\mathrm{term\ 1} * \mathrm{term\ 2})}{\mathrm{P}(\mathrm{term\ 2})\mathrm{P}(\mathrm{term\ 2})} \qquad (1)$$

Here, P(term1,term2) is the co-occurrence probability of term1 and term1, and P(term1)P(term2) gives the probability that the two terms co-occur if they are statistically independent. The ratio between P(term1,term2) and P(term1) P(term2) is thus a measure of the degree of statistical dependence between them [8]. The log of this ratio is the amount of information that we acquire about the presence of one of the word when we observe other [9]. Here the term1 means extracted two-word phrase from Table 2 and term2 means the seed words form Table 3, there is no special reason to choose these words only. We calculate PMI of phrase with respect to both categories of seed words.

After calculation of PMI of phrase we calculate the Semantic Orientation of two-word phrase as given in (2). To find the Semantic Orientation measure of a phrase is calculated as follows:

$$\mathrm{SO}(\mathrm{phrase}) = \mathrm{PMI}(\mathrm{phrase}, \{\mathrm{Positive\ Seed\ Word}\}) - \mathrm{PMI}(\mathrm{phrase}, \{\mathrm{Negative\ Seed\ Word}\}) \qquad (2)$$

If the phrase is associated with any of the Positive Seed Word then Semantic Orientation is Positive, if phrase is associated with any of the negative word then the Semantic Orientation is negative.

3.4 Classification of Reviews

The third step of analyzing the orientation of review is to take average of Semantic Orientation of each phrase of review. If the average Semantic Orientation is Positive then review is classified as positive and if the average semantic orientation is negative then the review is classified as negative.

4. Experiments

Experiments are done on 667 reviews from Multidomain Dataset [10]. There are 381 negative reviews and 286 positive reviews. Table IV shows number of reviews from different domains such as Book, DVDs, Kitchen Appliances, and Electronics Appliances. There is variation among the accuracy of reviews among domain. The classification accuracy checked against the five star ratings given by the author of reviews.

Table 4: Accuracy of Proposed Algorithm

Domain	*No. of Reviews*	*Accuracy*
DVDs(Movie Review)	245	64.08%
Electronics	101	81%
Books	139	73.38%
Kitchen Appliances	182	79.12%
All	667	74.39

5. Discussion on Results

If we look at result, DVDs i.e. movie review accuracy is comparatively less than other domains. The question is why movie review domain is having less accuracy. The movie review consists of many factors such as Camera Direction, Story, Acting, and Sound Quality etc. The reviewer reviews about many things of movie. Movie a good movie may have unpleasant things like wise bad movies may have some pleasant things. Table 5 shows some misclassified example form DVDs (Movie Review) domain.

Table 5: Misclassified Reviews from DVD (Movie Reviews) Domain

Movie	*Better Than Chocolate*	*Chappelle's Show*
Authors Rating	5	1
Average SO	-0.01036	0.4968
Sample Phrase	Terribly wrong	Good Reason
SO of Sample Phrase	-3.2623	2.3532
Context of Sample Phrase	I was terribly wrong yet most pleasantly surprised.	I agree, these were not used in the first two seasons for a good reason, they are not that good

In [9] Peter D. Turney uses PMI-IR algorithm which achieves an average accuracy of 74% when evaluated on 410 reviews from Epinions, sampled from four different domains (reviews of automobiles, banks, movies, and travel destinations). The accuracy ranges from 84% for automobile reviews to 66% for movie reviews.

PMI-IR estimates PMI by issuing queries to a search engine (hence the IR in PMI-IR) and noting the number of hits (matching documents). Experiments used the AltaVista Advanced Search engine, which indexes approximately 350 million web pages (counting only those pages that are in English). They chose AltaVista because it has a NEAR operator. The AltaVista NEAR operator constrains the search to documents that contain the words within ten words of one another, in either order. Previous work has shown that NEAR performs better than AND when measuring the strength of semantic association between words [9]. The Semantic Orientation of phrases extracted using AltaVista's NEAR operator can be calculated as follows:

$$SO(\text{phrase}) = \log\left(\frac{\text{hits}(\text{phrase NEAR "excellent "})\text{hits}(\text{"poor"})}{\text{hits}(\text{phrase NEAR poor})\text{hits}(\text{excellent})}\right) (3)$$

Following Table 6 summarizes comparative analysis of proposed algorithm and work done by Peter D. Turney [9].

Table 6: Summarization of Proposed and P. Turney Algorithm

Feature	*Proposed Algorithm*	*PMI-IR Algorithm (Turney 2002)*
Accuracy	74%	74%
Seed Words	Positive Words: Excellent, Good Negative Words: Poor, Bad, Hate, Suck, Terrible, Horrible	Positive: Excellent Negative: Poor
No. of Reviews to be classify	667	410
Size of Corpus	8000 Reviews	350 Million Webpages
Time Required To process	1 Hour The query performed on local machine so the required time is less.	30 Hours Time required to send query to AltaVista Search Engine and to search in 350 millions web pages.
Movie Review Domain Results	64% accuracy The accuracy is less because the reviewer reviews about many things of a movie.	66% accuracy The accuracy is less because the reviewer reviews about many things of a movie.
Electronics/Automobile Appliances	81% The accuracy is higher because parts are good then it add up good product.	84% The accuracy is higher because good automotive parts usually do add up to a good automobile.

6. Future Work

The PMI-IR algorithm gives phrases with semantic orientation values, which can be further used as bag-of-words fashion for classification of reviews. Bag-of-words fashion is the domain specific model. For this reason we need to firstly create bag-of-words using the unsupervised PMI algorithm which can be further used as classification of more reviews.

The semantic orientation of phrases can also be used for summarizing the review. The sentence with highest phrase value can be used as summarization of the review. The extension of this work can also be used for opinion holder extraction and feature extraction of review. This further can be used for comparative analysis of two products and summarization of its features

7. Conclusion

This paper implements a simple unsupervised PMI algorithm for sentiment classification of product review. The PMI algorithm has three simple steps: first is to extracts two-word phrase containing adjective and adverbs. Second is to find Sematic Orientation of phrases and third is to take average of all SO and assign sentiment as positive or negative. The experiments are done on 667 reviews, gives the accuracy of

74%. Experiments done on 4 domain and DVD i.e. movie reviews has less accuracy of 64%. As the movies are mixture of many things and reviewer reviews on each of them whereas electronics review has more accuracy of 81%.The limitation of algorithm is that it can't classify the reviews containing review about many things. To overcome this we need to deep drive features of review and opinion holder of that feature.

References

[1] Bing Liu, "Sentiment Analysis and Opinion Mining," Synthesis Lectures on Human Language Technologies, no.1, pp. 1-167, 2005
[2] Aue, A., & Gamon, M., "Automatic Identification of Sentiment Vocabulary: Exploiting Low Association With Known Terms," Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in NLP, pp. 57–64, June 2005.
[3] S. Bindra, P. Karmarkar, A. Verma, L. Grover, " Social Media Mining for Opinion Analysis," International Journal of Engineering and Advanced Technology (IJEAT), vol. 5(1), 2015
[4] V. Hatzivassiloglou, K. McKeown, "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. 1997.
[5] V. Hatzivassiloglou, & J. Wiebe,"Effects of adjective orientation and gradability on sentence subjectivity, "Proceedings of 18th International Conference on Computational Linguistics. New Brunswick, NJ: ACL, 2000.
[6] M. Hearst, "Direction-based text interpretation as an information access refinement," In P. Jacobs (Ed.), Text-Based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval. Mahwah, NJ: Lawrence Erlbaum Associates. 1992
[7] E. Brill,"Some advances in transformation-based part of speech tagging,"Proceedings of the Twelfth National Conference on Artificial Intelligence, pp. 722-727, Menlo Park, CA: AAAI Press,1994
[8] B. Liu, "Sentiment Analysis and Subjectivity. Hand Book of Natural Language Processing, Chicago, 2010.
[9] P. Turney,"Thumbs Up. Thumbs Down. Semantic Orientation Applied To Unsupervised Classification of Reviews,"Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia,pp. 417-424, 2002.
[10] "Multi-Domain Sentiment Dataset," cs.jhu.edu, Feb. 26, 2016. [Online]. Available: https://www.cs.jhu.edu/~mdredze/datasets/sentiment/
[11] Peter D. Turney, M. L. Littman. "Unsupervised learning of semantic orientation from a hundred-billion- word corpus," Technical Report ERC-1094 (NRC 44929), National Research Council of Canada, 2001.
[12] P. Turney,"Mining the Web for synonyms: PMI-IR versus LSA on TOEFL". Proceedings of the Twelfth European Conference on Machine Learning, pp. 491-502, 2001