Study of Semantic Models in Identifying Aspects in Customer Reviews

Mily Lal[1], Akanksha Goel[2], Amruta Chore[3]

[1, 2, 3]Padmashree Dr D.Y.Patil Institute of Engineering Management and Research, Akurdi, Pune, India

Abstract: ***With the outburst of ecommerce sentiment-rich resources such as online review sites and blogs, people actively use this information to understand what others think about a particular subject. This area of study helps to derive the opinion, sentiment or the outlook of a speaker mainly used when conducting market research. This paper does an evaluation of the systems such as LSA and PMI that set up semantic association between aspects and opinions found in customer reviews. PMI-IR is predicted to give better results as observed in a user study.***

Keywords: aspect extraction, opinion mining, opinion word, Sentiment analysis

1. Introduction

The rapid evolution of Web as well as its read-write nature has enabled more and more users to interact and share knowledge and information. Extracting opinions and sentiments from the web became more challenging, because deep understanding of the semantic structure of the natural language is required. This paper proposes to study aspect-based opinion mining methodology on customer reviews. It describes the comparison of two methods (LSA and PMI) [20][11]which finds the semantic association between aspects and opinions

2. Related Work

2.1 Deriving Aspects

Identification and extraction of explicit aspects are done initially as they are clearly precise in the reviews. This ultimately helps in the identification of implicit aspects that are hidden or implied in a review sentence. Nouns and noun phrases are extracted as they represent potential words that are frequently talked about.

Here POS (Parts Of Speech) tagger is used for separating various part of speech tokens [15][17].Opinion words are taken to be adjective, verbs and adverb phrases appearing in the review sentence [15][17].

The most likely aspect opinion pair occurring in an implicit sentence can be obtained using the association between opinion & explicit aspects Partial derivation of implicit aspects was done by Hu & Liu [6] by applying the same methods used for explicit aspect extraction.

Another approach used is clustering of opinion phrases which are then labeled with property names. Measures of Association are calculated on the opinion cluster and explicit feature using different methods like PMI (Pointwise Mutual Information), Likelihood ratio test (LRT), Cooccurance association rules (CoAR) [5][8].

2.2 Rule Generation

Hu and Liu [3][5][25] generated all strong association rules by applying apriori algorithm to extract all aspects expressed in reviews. The method was successful in identifying explicit aspects. The quantitative results for implicit feature identification are unknown [5][25].

2.3 Summarization

One of the simplest way to access the results is to produce a aspect based summary of opinions [3][25]. The relative frequency of feature can be seen by the application of Feature buzz summary. Organizations can know what their customers really care about [2]. Whereas Object buzz summary mentions the frequency of different products in competition. [28]. Lastly, the summarization of opinions can be obtained by producing a short textual summary based on multiple reviews or even a single review [28]

3. Proposed Techniques

3.1 Aspect Extraction

Nouns/noun phrases are frequently taken as aspects whereas adjective modifiers, adverb modifiers are taken to be opinion words. Aspects and Opinion are detected with the help of Stanford POS tagger.

The proposed paper is domain independent and unsupervised thus eliminates tedious and time consuming work for supervised methods. It is effective in medium size corpus. But for large corpora, this may result in precision drop. During extraction, adjectives like “whole” and “recent” that are extracted as opinion words can be associated with nouns/noun phrases, thus leading to extracting wrong aspects.

Every opinion extracted has some implicit polarity (positive, negative or neutral), associated with them which transforms the orientation of the aspects.

SentiWordNet is used to determine the polarity of each modifier. SentiWordNet (SWN) is designed to aid in opinion mining tasks [15]. Each synonymous set in SWN has a positive sentiment score, a negative sentiment score and an objectivity score. SWN requires word sense disambiguation to find the correct sense of a word and its associated scores. For example, “an unpredictable plot in the movie” is a positive phrase, while “an unpredictable steering wheel” is a negative one.

To include negation information, the negation word and negated word are joined with a hyphen. For example “not good” is replaced with not-good. The sentiment score of this new “word” is the negative of the sentiment score of the negated word [17]. Negation can occur in a variety of often subtle ways, thus can lead to poor results.

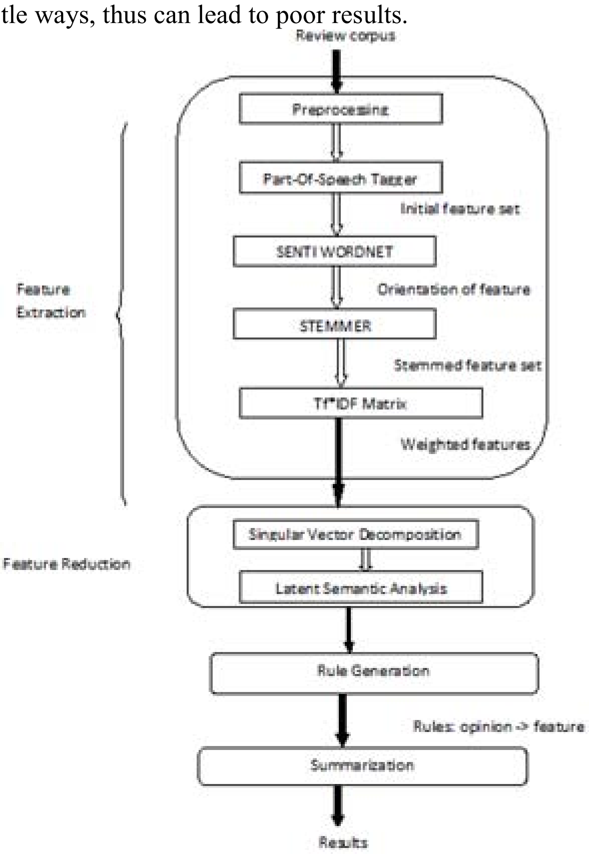


Figure 1: System Model using LSA

3.2 Finding Hidden Association

This paper proposes to do analysis between LSA and PMI-IR in finding associations between aspects and opinions in customer reviews.

Latent Semantic Analysis is used to reduce the dimension of vector representations of textual data [20]. LSA is the only available method for text content based similarity inference. LSA is a mathematical and statistical approach, claiming that semantic information can be derived from a word-document co-occurrence matrix and words and documents can be represented as points in a (high-dimensional) Euclidean space.

LSA requires comparatively high computational performance and memory. The challenge in using this methodology is the difficulty in resolving the optimal number of dimensions to use for performing the SVD. The number of dimensions that can be used is restricted by the size and nature of the document collection.

PMI-IR is also a statistical approach but not based on the concept of semantic spaces. It uses the results of information retrieval to compute associativity in terms of the words mutual *information*. The degree of relatedness is measured by the probability of co-occurrence versus independent occurrence of terms [11]. PMI gave better results than LSA on synonym tests [16]. PMI-IR has also performed effectively than other computational methods for quantifying similarity [12].

4.3 Rule Generation

The proposed approach can be viewed as an elaborate extension of Hu and Liu’s method [3][5][25].

The proposed work is designed specifically to identify aspects that do not occur explicitly in review sentences. Secondly, the approach discriminates between opinion words and aspect words i.e opinion words can only occur in the rule antecedents, while rule consequents must be opinion aspects [18]. Thirdly association rules are generated directly from the LSA matrix / PMI matrix of opinions and aspects. Large number of incorrect rules may be generated which are caused by the incorrect identification of opinion words or explicit aspect words by the previous modules. However it helps in generating quite reasonable rules due to the LSA/PMI that helps to measure semantic associations between the objects [20][11].

3.4 Summarization

Review summarization intends at producing a sentiment summary, which consists of sentences from a document that capture the author’s opinion. The summary may be either a single paragraph as in [21] or a structured sentence list as in [3][9][25]. The former is produced by selecting some sentences or a whole paragraph which the author expresses his or her opinion(s). The latter is generated by the auto mined aspects that the author comments on. The proposed method used is more relevant to the method used in [3][25][9] i.e. aspect based summary of opinions on an object or multiple competing objects.

4. Implementation

Statistical Opinion Mining is used which tackles sentiment analysis in terms of data mining and is based on statistical methods. The customer review corpus was collected from www.amazon.com. Amazon as the source of reviews, which includes user reviews for cell phones. The corpus contains 300 reviews. Products in this site have a large number of reviews. Each of the review includes a text review. Additional information available but not used in this project includes date, time, author name, location and ratings. Reviews based on cell phones were manually collected. A typical review contains free text summary about a product. All reviews are plain text. First a set of standard preprocessing steps are carried out, viz., tokenizing and stemming.

Stanford POS automatically classified words into categories of nouns and noun phrases based on the following pattern

NN or NNS , NN NN or NNS Example:- In the sentence, "*This camera produces beautiful pictures*", "pictures" and "camera" will be extracted as it satisfies the first pattern. In the sentence, "*This is a simple cell phone"*, "cell phone" will be extracted as it satisfies the second pattern. be used with the SentiWordNet database. SentiWordNet_3.0.0.identified words polarity as negative or positive based on th given rule: score less than 0 are as negative and scores greater than 0 is taken as positive. The rest were considered neutral. While using Latent Semantic Analysis, values close to 1 represent very synonyms. A threshold was set to reduce the number of terms collected. Association rule were used to mine the aspect->opinion rules from the resulting LSA matrix. Support was considered to be 1% [3][5]. A distinguished set of rules were obtained. The summarization phase was conducted as follows. The sentences related to the query were collected. Semantic polarity of the opinions found in each sentence is identified from the previously identified set. This organized sentence list were shown as the summary [3][25][10]. For implicit aspects, a matched list of rules is collected by searching the rules antecedents. Example:- *"it's also a good tool for entertainment"* Opinions -> good Implicit aspect = nokia-n95 -> [nokia-n95, memory, sound, memory-card]. If the implicit aspect is identical to the queried aspect then the sentence is structured in the summary.

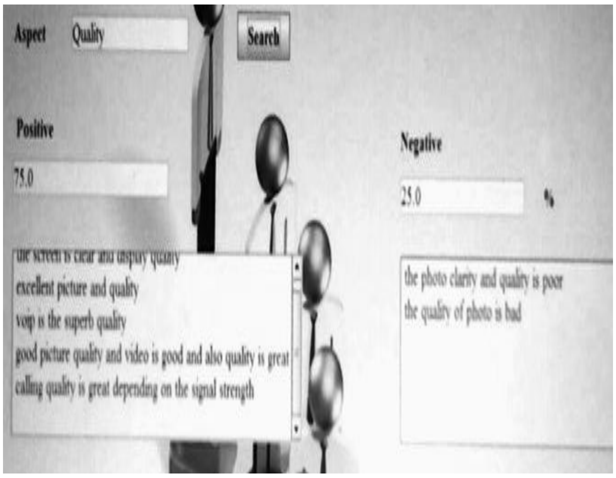
The Comparison model used PMI instead of LSA methodology. The point-wise mutual information (PMI) Mi(w)[aspect] between the word w and the class i [opinion] is defined on the based on the level of co-occurrence between the class i and word w. On the basis of mutual independence, the expected co-occurrence is given by Pi*F(w), and the true co-occurrence is given by F(w)* pi(w). The word w is positively correlated Mi(w) is greater than 0 and negative when less than 0.

5. Observations

As observed the number of positive sentiments in LSA model is more than PMI model.

LSA was the supposed difficulty in determining the optimal number of dimensions to use for performing the SVD was determined to be "6" as the optimal number of dimensions to use for performing the SVD as recall as 66.67% and precision as 71.42%. Too few dimensions and important patterns are left out, too many and noise caused by random word choices will creep back in.

PMI estimates for each pair of words, the aspects and the opinions were calculated by dividing the number of times that aspect and opinion has co-occurred within a single document with the product of the respective frequencies of aspect and opinion in the entire corpus. PMI is easy to compute even on large corpus and requires low memory. It dose automatic approximations of semantic similarity. But it is observed to be bad with sparse data. Even if the opinion-aspect pair occurs together once, high PMI score is obtained.



Aspect Quality

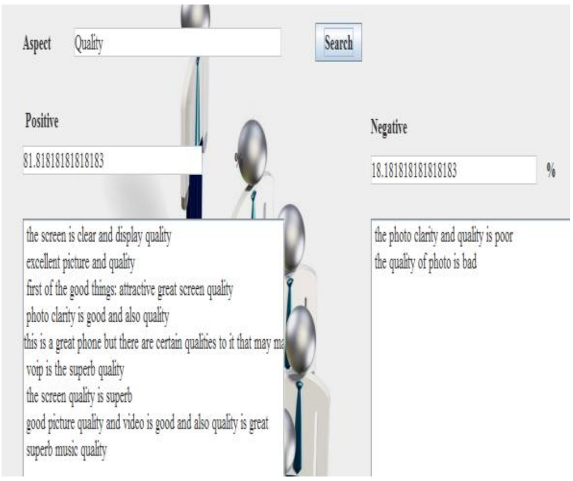
Positive 75.0

Negative 25.0 %

the screen is clear and display quality
excellent picture and quality
voip is the superb quality
good picture quality and video is good and also quality is great
calling quality is great depending on the signal strength

the photo clarity and quality is poor
the quality of photo is bad

Figure 2: Summary generated by PMI model



Aspect Quality

Positive 81.81818181818183

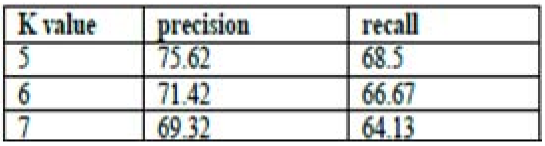
Negative 18.181818181818183 %

the screen is clear and display quality
excellent picture and quality
first of the good things: attractive great screen quality
photo clarity is good and also quality
this is a great phone but there are certain qualities to it that may ma
voip is the superb quality
the screen quality is superb
good picture quality and video is good and also quality is great
superb music quality

the photo clarity and quality is poor
the quality of photo is bad

Figure 2: Summary generated by LSA model

Table 1: Precision & Recall of LSA



K value	precision	recall
5	75.62	68.5
6	71.42	66.67
7	69.32	64.13

6. Conclusion and Future Work

This paper compares two semantic systems, LSA and PMI-IR, for predicting the associations between aspect and opinion in a text review corpus.

The both methodologies used here is quite reasonable in identifying aspects and their semantic association but some undesirable errors still persists in the results. This might be due to incorrect identification of aspects - opinion pairs or due to the error in segmentation and parsing.

An important issue related to this domain trustworthiness of online opinions which is not considered throughout this work. There is not much study reported on evaluating the authenticity of product reviews. In future, we will determine novel and effective technique for detecting spam reviews which will help customer in selecting best buying option.

References

[1] Turney, p. (2002). Thumbs up or thumbs down? Semantic orientation applied tounsupervised classification of reviews, In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, Pennsylvania.

[2] Liu, B., "Sentiment Analysis and Subjectivity" Handbook of Natural Language Processing, Second Edition, (editors: N. Indurkhya and F. J. Damerau), 2010.

[3] Hu, M. and Liu, B. 2004. Mining and summarizing customer reviews. International Conference on Knowledge Discovery and Data Mining (ICDM).

[4] Popescu, A. and Etzioni, O. (2005). Extracting product features and opinions from reviews, In Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP). Potts, C. (2007) The Expressive Dimension. Theoretical Linguistics 33:165-198.

[5] Ding, X., Liu, B. and Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining, In Proceedings of the Conference on Web Search and Web Data Mining (WSDM).

[6] Qi Su,Kun Xiang,Houfeng Wang,Bin Sun and Shiwen Yu(2006).Using Pointwise Mutual Information to Identify Implicit Features in Customer Reviews.ICCPOL ,LNAI 4285 ,pp.22-30,Springer(2006).

[7] Zhu J.,H.Wang,M,Zhu and B.K.Tsou.2011.Aspect based opinion polling from customer reviews. IEEE Transactions on Affective Computing,2(1):37-49.

[8] Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of ACL-04, 42nd Meeting of the Association for Computational Linguistics, pages 271–278, Barcelona, ES.

[9] Zhuang, L., F. Jing, X.-Yan Zhu, and L. Zhang. Movie review mining and summarization. In Proceedings of ACM International Conference on Information and Knowledge Management (CIKM-2006), 2006

[10] Qiu, G., B. Liu, J. Bu, and C. Chen. Opinion word expansion and target extraction through double propagation. Computational Linguistics, 2011.

[11] Church, K, Gale, W., Hanks, P., Hindle, D. (1991)Using Statistics in Lexical Analysis, in Zernik (ed.)*Lexical Acquisition: Exploiting OnLine Resources toBuild a Lexicon*, 115-164, Lawrence Erlbaum Associates Publishers.

[12] Terra, E., & Clarke, C. L. A. (2003). Frequency Estimates for Statistical Word Similarity Measures. *Proceedings of Human Language Technology Conference*. North American chapter of the Association for Computational Linguistics annual meeting, 244-251.1.

[13] Jin, W. and H. Ho. A novel lexicalized HMM-based learning framework for web opinion mining. In Proceedings of International Conference on Machine Learning (ICML-2009), 2009a.

[14] Jin, W. and H. Ho. OpinionMiner: a novel machine learning system for web opinion mining and extraction. In Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2009), 2009b.

[15] Esuli, A., and F. Sebastiani. "Determining term subjectivity and term orientation for opinion mining." In Proceedings of Annual Conference of the European Chapter of the Association of Computational Linguistics (EACL-2006), 2006.

[16] Turney, P. (2001). Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. *Proceedings of the Twelfth European Conference on Machine Learning*, 491-502

[17] Shitanshu Verma, Pushpak Bhattacharyya."Incorporating Semantic Knowledge for Sentiment Analysis". Proceedings of ICON-2008: 6th International Conference on Natural Language Processing Macmillan Publishers, India.

[18] Zhen Hai, Kuiyu Chang and Jung-jae Kim. Implicit Feature Identification via Co-occurrence Association rule Mining.Springer(2011)

[19] Deerwester, Dumais, Furnas, Lanouauer, and Harshman,Indexing by latent semantic analysis, Journal of the American Society for Information Science, 41 (1990), pp. 391-407.

[20] Akshat Bakliwal,Piyush Arora , Vasudeva Varma. Entity Centric Opinion Mining from Blogs. Preceedings of the 2nd Workshop on Sentiment Analysis where AI meets Psychology(SAAIP 2012),pages 53-64, COLING 2012,December 2012

[21] J. Wiebe, "Learning subjective adjectives from corpora,". Proceedings of AAAI, 2000.

[22] Bing Liu. Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, May 2012.

[23] Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In Proceedings of HLT-NAACL 2003, pp. 252-259.

[24] Liu, B., Hu, M., And Cheng, J. 2005. Opinion observer: analyzing and comparing opinions on the web. In WWW '05: Proceedings of the 14th international conference on World Wide Web. ACM, New York, NY, USA, 342–351.