International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-2, December, 2019

Detection of Sentiment Analysis with Co-Occurrence Data using Supervised and Unsupervised Methods

R. Madhu Priya, J. Naga Muneiah

Abstract: ***With the rapid growth of user-generated content on the internet, sentiment analysis of online reviews has become a hot research topic recently, but due to variety and wide range of products and services, the supervised and unsupervised domain- specific models are often not practical. As the number of reviews expands, it is essential to develop an efficient sentiment analysis model that is capable of extracting product aspects and determining the sentiments for aspects. A text processing framework that can summarize reviews would therefore be desirable. A subtask to be performed by such a framework would be to find the general aspect categories addressed in review sentences, for which this paper presents two methods. In this paper, we propose an unsupervised model for detecting aspects in reviews. In this model, first a generalized method is proposed to learn multi-word aspects. Second, a set of heuristic rules is employed to take into account the influence of an opinion word on detecting the aspect. In contrast to most existing approaches, the first method presented is an unsupervised method that applies association rule mining on co-occurrence frequency data obtained from a corpus to find these aspect categories. The proposed unsupervised method performs better than several simple baselines, a similar but supervised method, and a supervised baseline; the proposed model does not require labeled training data and can be applicable to other languages or domains. We demonstrate the effectiveness of our model on a collection of product reviews dataset, where it outperforms other techniques.***

Keywords: ***Aspect category detection, consumer reviews, co-occurrence data, sentiment analysis, supervised, unsupervised.***

I. INTRODUCTION

In the past few years, with the rapid growth of user-generated content on the internet, sentiment analysis (or opinion mining) has attracted a great deal of attention from researchers of data mining and natural language processing. Sentiment analysis is a type of text analysis under the broad area of text mining and computational intelligence. Three fundamental problems in sentiment analysis are: aspect detection, opinion word detection and sentiment orientation identification [1-2].Aspects are topics on which opinion are expressed. In the field of sentiment analysis, other names for aspect are: features, product features or opinion targets [1-5]. Aspects are important because without knowing them, the opinions expressed in a sentence or a review are of limited use. For example, in the review sentence "after using it, I found the size to be perfect for carrying in a pocket", "size" is the aspect for which an opinion is expressed. Likewise aspect detection is critical to sentiment analysis, because its effectiveness dramatically affects the performance of opinion word detection and sentiment orientation identification. Therefore, in this study we concentrate on aspect detection for sentiment analysis. Retail companies such as Amazon and Bol have numerous reviews of the products they sell, which provide a wealth of information, and sites like Yelp offer detailed consumer reviews of local restaurants, hotels, and other businesses. Research has shown these reviews are con- sidered more valuable for consumers than market-generated information and editorial recommendations [4]–[6], and are increasingly used in purchase decision-making [7]. a supervised machine learn- ing approach to aspect category detection is feasible, yielding a high performance [11]. Many approaches to find aspect categories are supervised [11]–[14]. However, sometimes the flexibility inherent to an unsupervised method is desirable.

Existing aspect detection methods can broadly be classified into two major approaches: supervised and unsupervised. Supervised aspect detection approaches re- quire a set of pre-labeled training data. Although the supervised approaches can achieve reasonable effectiveness, building sufficient labeled data is often expensive and needs much human labor. Since unlabeled data are generally publicly available, it is desirable to develop a model that works with unlabeled data. Additionally due to variety and wide range of products and services being reviewed on the internet, supervised, domain-specific or language-dependent models are often not practical. Therefore the framework for the aspect detection must be robust and easily transferable between domains or languages.

This approach, however, only considered adjectives as opinion words which are not able to cover every opinion, yet the approach was capable of finding hidden links between product aspects and adjectives. Unfortunately, there were no quantitative experimental results reported, specifically for implicit aspect identification.

Revised Manuscript Received on December 30, 2019.
* Correspondence Author
Ms. R Madhu Priya*, M.Tech Dept. of CSE, Chadalawada Ramanamma Engineering , College, Tirupati, India. madhupriyacrec@gmail.com
Prof. J NagaMuneiah, Head Dept. of CSE, Chadalawada Ramanamma Engineering, College, Tirupati, India. nagamuni513@gmail.com

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

A two-phase co-occurrence association rule mining approach to identify implicit aspects [15].

In the first phase of rule generation, association rules are mined of the from [*opinion word explicit aspect*], from a co-occurrence matrix. Each entry in the co-occurrence matrix represents the frequency degree of a certain opinion-word co-occurring with a certain explicitly mentioned aspect. In the second phase, the rule consequents (i.e., the explicit aspects) are clustered to generate more robust rules for each opinion word. In [19], a semi-unsupervised method is proposed that can simultaneously extract both sentiment words and prod- uct/service aspects from review sentences. The method first extracts appraisal expression patterns (AEPs), which are representations of how people express opinions regarding products or services. The set of AEPs is obtained by selecting frequently occurring shortest dependency paths between two words in a dependency graph.

The model can easily be transform between domains or languages. In the remainder of this paper, detailed discussions of existing works on aspect detection will be describes the proposed aspect detection model for sentiment analysis, including the overall process and specific designs. Subsequently we describe our empirical evaluation and discuss important experimental results. Finally we conclude with a summary and some future research directions.

II. RELATED WORK

Several methods have been proposed, mainly in the context of product review mining [1-14]. The earliest attempt on aspect detection was based on the classic information extraction approach of using frequently occurring noun phrases presented by Hu and Liu [3]. Their work can be considered as the initiator work on aspect extraction from reviews. They use association rule mining (ARM) based on the Apriori algorithm to extract frequent itemsets as explicit product features, only in the form of noun phras- es. Their approach works well in detecting aspects that are strongly associated with a single noun, but are less useful when aspects encompass many low-frequency terms. The proposed model in our study works well with low-frequency terms and uses more POS patterns to extract the candidates for aspect. Wei et al. [4] proposed a semantic- based product aspect extraction (SPE) method. Their approach begins with preprocessing task, and then employs the association rule mining to identify candidate product aspects. An early work on implicit aspect detection is [17].

The authors propose to use semantic association analysis based on point-wise mutual information (PMI) to differentiate implicit aspects from single notional words. Unfortunately, there were no quantitative experimental results reported in their work, but intuitively the use of statistical semantic association analysis should allow for certain opinion words such as “large,” to estimate the associated aspect (“size”). Association rule mining is also employed in [20], where first the candidate aspect indicators are extracted based on word segmentation, part-of-speech (POS) tagging, and aspect clustering. After that, the co-occurrence degree between these candidate aspect indicators and aspect words are calculated, using five collocation extraction algorithms.Association rule mining is also the main technique in [21].

Unlike [15] and [20], no annotated explicit aspects are required, instead the double propagation algorithm from [22] is employed to identify the explicit aspects. An advantage of this double propagation method is that it links explicit aspects to opinion words. This is used later, to restrict the set of possi- ble implicit aspects in a sentence to just those that are linked to the opinion words present in that sentence. high performing supervised aspect category detec- tion is proposed in [11]. Instead of a MaxEnt classifier, five binary (one-versus-all) SVMs are employed, one for each aspect category. The SVMs use various types of n-grams (e.g., stemmed, character, etc.) and information from a word clustering and a lexicon, both learned from YELP data. The lexicon learned from YELP data significantly improved the F_1-score, which was reported to be 88.6% and ranked first among 21 submis- sions in SemEval-2014 workshop.

III. UNSUPERVISED IMETHOD

The proposed method is unsupervised uses co-occurrence association rule mining in a related way as [12], by learning relevant rules between notional words, defined as the words in the sentence after removing stop words and low frequency words, and the well thought-out categories. This enables the algorithm to mean a category based on the words in a sentence. Association rules are mined when a strong relation between a notional word and one of the aspect categories exists, with the strength of the relation being modeled using the co- occurrence frequency between category and notional word. In this function we focus on selecting some aspects from the candidates as seed set information. We introduce a new metric named A-Score, which selects the seed set in an unsupervised manner. This metric is employed to learn a small list of top aspects with a complete precision.

A-Score Metric

Here we introduce a new metric, named A-score which uses inter-relation information between words to score them. We score each candidate aspect with A-score metric defined as:

$$A\text{—}Score(a) = f(a) \times \sum \log \left(\frac{f(a,bi) \times N + 1}{f(a) \times f(bi)} \right)$$

Where a is the current aspect, f(a) is the number of the sentences in the corpus which a is appeared, f(a, bi) is the frequency of co-occurrence of aspect a and biin each sentence. bi is ith aspect in the list of seed aspects, and N is number of sentences in the corpus. The A-Score metric is based on mutual information between an aspect and a list of aspects, in addition it considers frequency of each aspect. We apply the add-one smoothing to the metric, so all co-frequencies be non-zero. This metric helps to extract more informative aspects and more co-related ones.

Figure 1 gives an overview of the proposed model used for detecting aspects in sen- timent analysis. Below, we discuss each of the functions in aspect detection model in turn.

Model: Aspect Detection for Sentiment Analysis
Input: Reviews Dataset
Method:
Extract Review
Sentences FOR
each sentence
Use POS Tagging
Extract POS Tag Patterns as Candidates
for Aspects END FOR
FOR each
candidate
aspect Use
Stemming
Select
Multi-Word
Aspects Use a Set
of Heuristic
Rules
END FOR
Make Initial Seeds for Final Aspects
Use Iterative Bootstrapping for Detecting
Final Aspects Aspect Pruning
Output: Top Selected Aspects

Fig.1. The proposed model for aspect detection for sentiment analysis

IV. SUPERVISED METHOD

The supervised method employs co-occurrence association rule mining to detect categories. The following algorithm used as evaluation metric the F_1-score

Algorithm: Identify Category Set C and Compute Weight Matrix W

input : training set
input : occurrence threshold θ
output: category set C, Weight matrix W

C, X, Y
foreach *sentence* s *Training set* **do**
  foreach s_k $s_L, s_{D_1}, s_{D_2}, s_{D_3}$ **do**
    foreach *dependency forms/lemmas* j s_k **do**
      if j / Y **then**
        add j to Y
      end
      Y_j Y_j 1
      foreach *category* c s_C **do**
        if c / C **then**
          add c to C
        end
        if $(c, j) / X$ **then**
          add (c, j) to X
        end
        $X_{c,j}$ $X_{c,j}$ 1
      end
    end
  end
end
foreach (c,j) X **do**
  if $Y_j > \theta$ **then**
    $W_{c,j}$ $X_{c,j}/Y_j$
  end
end

V. EVALUATION

The proposed methods of evaluation, the training and test data from SemEval-2014 [10] are used. It contains 3000 training sentences and 800 test sentences taken from restaurant reviews. that each sentence has at least one category and that approximately 20% of the sentences have multiple categories. With 20% of the sentences having multiple categories, a method would benefit from being able to predict multiple categories. The association rule mining is useful in this scenario as multiple rules can apply to a single sentence. Because both unsupervised and supervised method work best for well-defined aspect categories, the last category in this data set, anecdotes/miscellaneous poses a challenge. It is unclear what exactly belongs in this category, and its concept is rather abstract. For that reason, we have chosen not to assign this category using any of the actual algorithms, but instead, this category is assigned when no other category is assigned by the algorithm. This was as expected, since dependency indicators consists of more than one component, which makes it harder to find rules that generalize well to unseen data, and, in addition, they also rely on the grammatical correctness of the sentence.

Using dependency indicators, in addition to lemma indicators, do seem to result into finding more categories, even though it is less precise in doing so. This is especially the case for the category food. The main reason for this is that food is by far the largest category, resulting in more available training data for this category, which makes it easier to find rules.

VI. EXPERIMENTAL RESULTS

In this section we discuss the experimental results for the proposed model and pre- sented algorithms. We employed datasets of customer reviews for five products for our evaluation purposeThis dataset focus on electronic products: Apex AD2600 Progressive-scan DVD player, Canon G3, Creative Labs Nomad Jukebox Zen Xtra 40 GB, Nikon Coolpix 4300, and Nokia 6610. Table 2 shows the number of manually tagged product aspects and the number of reviews for each product in thedataset.

Table-I: Summary of customer review dataset

DataSet	Number of reviews	No. of manual aspects
Canon	45	100
Nikon	34	74
Nokia	41	109

Creative	95	180
Apex	99	110

VII. CONCLUSION

This paper proposed a model for the task of detection for sentiment analysis for co-occurrence data aspects in reviews. two methods for detecting aspect categories, the first one is unsupervised method, which is used for spreading the creation ended a graph built from word co-occurrence data, enabling the use of both direct and indirect relations between words. The second, supervised, method uses a rather straight- forward co-occurrence method where the co-occurrence frequency between annotated aspect categories and both lem- mas and dependencies is used to calculate conditional prob- abilities. If the maximum conditional probability is higher than the associated, trained, threshold, the category is assigned to that sentence. We plan to employ clustering methods in conjunction with the model to extract implicit and explicit aspects together to summarize output based on the opinions that have been expressed on them by using machine learning techniques.

REFERENCES

1. Qiu, G., Liu, B., Bu, J., Chen, C.: Opinion word expansion and target extraction through double propagation. Computational Linguistics 37(1), 9–27 (2011)
2. 2. Thet, T.T., Na, J.C., Khoo, C.S.G.: Aspect-Based Sentiment Analysis of Movie Reviews on Discussion Boards. Journal of Information Science 36(6), 823–848 (2010)
3. Hu, M., Liu, B.: Mining opinion features in customer reviews. In: American Association for Artificial Intelligence (AAAI) Conference, pp. 755–760 (2004)
4. Wei, C.P., Chen, Y.M., Yang, C.S., Yang, C.C.: Understanding what concerns consumers: A semantic approach to product feature extraction from consumer reviews. Information Systems and E-Business Management 8(2), 149–167 (2010)
5. Brody, S., Elhadad, N.: An unsupervised aspect-sentiment model for online reviews. In: 2010 Annual Conference of the North American Chapter of the Association for Computa- tional Linguistics, Los Angeles, California, pp. 804–812 (2010)
6. Popescu, A., Etzioni, O.: Extracting product features and opinions from reviews. In: Con- ference on Human Language Technology and Empirical Methods in Natural Language Processing, Vancouver, pp. 339–346 (2005)
7. Yi, J., Nasukawa, T., Bunescu, R., Niblack, W.: Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In: 3rd IEEE Interna- tional Conference on Data Mining (ICDM 2003), Melbourne, FL, pp. 427–434 (2003)
8. Somprasertsri, G., Lalitrojwong, P.: Automatic product feature extraction from online product reviews using maximum entropy with lexical and syntactic features. In: IEEE In- ternational Conference on Information Reuse and Integration, pp. 250–255 (2008)
9. Zhu, J., Wang, H., Zhu, M., Tsou, B.K.: Aspect-based opinion polling from customer re- views. IEEE Transactions on Affective Computing 2(1), 37–49 (2011)
10. Zhai, Z., Liu, B., Xu, H., Jia, P.: Constrained LDA for Grouping Product Features in Opi- nion Mining. In: Huang, J.Z., Cao, L., Srivastava, J. (eds.) PAKDD 2011, Part I. LNCS, vol. 6634, pp. 448–459. Springer, Heidelberg (2011)
11. Su, Q., Xu, X., Guo, H., Guo, Z., Wu, X., Zhang, X., Su, Z.: Hidden sentiment association in chinese web opinion mining. In: 17th International Conference on World Wide Web, Beijing, China, pp. 959–968 (2008)
12. D. Smith, S. Menon, and K. Sivakumar, "Online peer and editorial recommendations, trust, and choice in virtual markets," J. Interact. Marketing, vol. 19, no. 3, pp. 15–37, 2005.
13. M. Trusov, R. E. Bucklin, and K. Pauwels, "Effects of word-of-mouth versus traditional marketing: Findings from an Internet social networking site," J. Marketing, vol. 73, no. 5, pp. 90–102, 2009.
14. M. T. Adjei, S. M. Noble, and C. H. Noble, "The influence of C2C communications in online brand communities on customer pur- chase behavior," J. Acad. Marketing Sci., vol. 38, no. 5, pp. 634–653, 2010.
15. B. Pang and L. Lee, "Opinion mining and sentiment analysis," Found. Trends Inf. Retrieval, vol. 2, nos. 1–2, pp. 1–135, 2008.
16. C.-L. Liu, W.-H. Hsaio, C.-H. Lee, G.-C. Lu, and E. Jou, "Movie rating and review summarization in mobile environment," IEEE Trans. Syst., Man, Cybern. C, Appl. Rev., vol. 42, no. 3, pp. 397–407, May 2012.
17. M. Pontiki et al., "SemEval-2014 Task 4: Aspect based sentiment anal- ysis," in Proc. 8th Int. Workshop Semantic Eval. (SemEval), Dublin, Ireland, 2014, pp. 27–35.
18. S. Kiritchenko, X. Zhu, C. Cherry, and S. M. Mohammad, "NRC-Cananda-2014: Detecting aspects and sentiment in customer reviews," in Proc. 8th Int. Workshop Semantic Eval. (SemEval), Dublin, Ireland, 2014, pp. 437–442.
19. Y. Zhang and W. Zhu, "Extracting implicit features in Online customer reviews for opinion mining," in Proc. 22nd Int. Conf. World Wide Web Companion (WWW Companion), 2013, pp. 103–104.
20. G. Qiu, B. Liu, J. Bu, and C. Chen, "Opinion word expansion and tar- get extraction through double propagation," Comput. Linguist., vol. 37, no. 1, pp. 9–27, 2011.
21. K. Schouten, F. Frasincar, and F. de Jong, "COMMIT-P1WP3: A co-occurrence based approach to aspect-level sentiment analysis," in Proc. 8th Int. Workshop Semantic Eval. (SemEval), Dublin, Ireland, 2014, pp. 203–207.
22. A. Garcia-Pablos, M. Cuadros, S. Gaines, and G. Rigau, "V3: Unsupervised generation of domain aspect terms for aspect based sentiment analysis," in Proc. 8th Int. Workshop Semantic Eval. (SemEval), Dublin, Ireland, 2014, pp. 833–837.
23. Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in Proc. 32nd Annu. Meeting Assoc. Comput. Linquistics, Las Cruces, NM, USA, 1994, pp. 133–138.
24. F. Crestani, "Application of spreading activation techniques in informa- tion retrieval," Artif. Intell. Rev., vol. 11, no. 6, pp. 453–482, 1997.
25. S. Bagchi, G. Biswas, and K. Kawamura, "Task planning under uncer- tainty using a spreading activation network," IEEE Trans. Syst., Man, Cybern. A, Syst., Humans, vol. 30, no. 6, pp. 639–650, Nov. 2000.
26. A. Katifori, C. Vassilakis, and A. Dix, "Ontologies and the brain: Using spreading activation through ontologies to support personal interaction," Cognitive Syst. Res., vol. 11, no. 1, pp. 25–41, 2010.
27. C. D. Manning et al., "The Stanford CoreNLP natural language processing toolkit," in Proc. 52nd Annu. Meeting Assoc. Comput. Linguist. Syst. Demonstrations, 2014, pp. 55–60. [Online]. Available: http://www.aclweb.org/anthology/P/P14/P14-5010
28. M.-C. de Marneffe and C. D. Manning, "Stanford typed depen- dencies manual," Stanford NLP Group, Stanford University, Stanford, CA, USA, Tech. Rep., Sep. 2008. [Online]. Available: https://nlp.stanford.edu/software/dependencies_manual.pdf
29. P. F. Bone, "Word-of-mouth effects on short-term and long-term product judgments," J. Bus. Res., vol. 32, no. 3, pp. 213–223, 1995.
30. R. Feldman, "Techniques and applications for sentiment analysis,"
31. Commun. ACM, vol. 56, no. 4, pp. 82–89, 2013.
32. S. Sen and D. Lerman, "Why are you telling me this? An examination into negative consumer reviews on the Web," J. Interact. Marketing, vol. 21, no. 4, pp. 76–94, 2007.
33. B. Bickart and R. M. Shindler, "Internet forums as influential sources of consumer information," J. Consum. Res., vol. 15, no. 3, pp. 31–40, 2001.
34. D. Smith, S. Menon, and K. Sivakumar, "Online peer and editorial recommendations, trust, and choice in virtual markets," J. Interact. Marketing, vol. 19, no. 3, pp. 15–37, 2005.

35. M. Trusov, R. E. Bucklin, and K. Pauwels, "Effects of word-of-mouth versus traditional marketing: Findings from an Internet social networking site," J. Marketing, vol. 73, no. 5, pp. 90–102, 2009.
36. G. O. Young, "Synthetic structure of industrial plastics (Book style with paper title and editor)," in *Plastics*, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15–64.
37. W.-K. Chen, *Linear Networks and Systems* (Book style). Belmont, CA: Wadsworth, 1993, pp. 123–135.
38. H. Poor, *An Introduction to Signal Detection and Estimation*. New York: Springer-Verlag, 1985, ch. 4.
39. B. Smith, "An approach to graphs of linear forms (Unpublished work style)," unpublished.
40. E. H. Miller, "A note on reflector arrays (Periodical style—Accepted for publication)," *IEEE Trans. Antennas Propagat.*, to be published.
41. J. Wang, "Fundamentals of erbium-doped fiber amplifiers arrays (Periodical style—Submitted for publication)," *IEEE J. Quantum Electron.*, submitted for publication.
42. C. J. Kaufman, Rocky Mountain Research Lab., Boulder, CO, private communication, May 1995.
43. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interfaces(Translation Journals style)," *IEEE Transl. J. Magn.Jpn.*, vol. 2, Aug. 1987, pp. 740–741 [*Dig. 9th Annu. Conf. Magnetics* Japan, 1982, p. 301].
44. M. Young, *The Techincal Writers Handbook.* Mill Valley, CA: University Science, 1989.
45. (Basic Book/Monograph Online Sources) J. K. Author. (year, month, day). *Title* (edition) [Type of medium]. Volume(issue). Available: http://www.(URL)
46. J. Jones. (1991, May 10). Networks (2nd ed.) [Online]. Available: http://www.atm.com

AUTHORS PROFILE

Ms. R Madhu Priya[1], M.Tech Student, Dept. of CSE, Chadalawada Ramanamma Engineering College, Tirupati, India. madhupriyacrec@gmail.com

Prof. Janapati NagaMuneiah received the B.Tech (CSE) from Jawaharlal Nehru Technological University, Hyderabad, India in 2001 and M.Tech in CSE from Sri Venkateswara University, Tirupati, India in 2010. He is pursuing his Ph.D in Jawaharlal Nehru Technological University, Kakinada, India in Computer Science and Engineering faculty. He has got 17 years of teaching experience. Presently he is working as Professor and Head of Department of CSE in Chadalawada Ramanamma Engineering college, Tirupati, A.P, India. His areas of interests include Data Mining, Data Warehousing, Big Data, Data Structures and Algorithms. He has guided 11 M. Tech theses. He has published more than 10 papers in International journals and some of them are published in SCIE and SCOPUS indexed journals.