SeAbOM: Semi-supervised Learning for Aspect-Based Opinion Mining

Sugandha C. Nandedkar and Jayantrao B. Patil

Abstract Opinion-rich information generated by social media users plays a vital role in today's market economy. The impact exists from understanding the current fashion trend to the product failure anatomy. It helps to decide the plans for the growth of an industry. To address this issue, business analysts want a detailed aspect-based analysis of user opinion. Many researchers either tried a supervised or unsupervised approach for the same. The literature review showed that the weak structuring of the corpus impacts the outcome of the state-of-the-art method. This paper illustrates the semi-supervised way to extract and summarize the aspect and sentiments associated with the user reviews. We proposed a mechanism to learn aspect-related terms (ARTs) for the seed aspect terms (ATs) from the corpus. We used the customer review dataset and SemEval Reviews-English to test the working and performance of our system. The results show that the proposed method achieves a recall upto 0.88 for the review corpus.

Keywords Opinion mining · Sentiment analysis · Aspect mining · Semi-supervised aspect mining

1 Introduction

The world of social media is full of opinion-rich information. The users from a variant domain are the cause of the development of this big data. The users' opinion on any product or topic is available with a single click in the big data world. Analyzing the users' opinions involves processing natural language text [1]. The application of opinion analysis ranges from understanding the current fashion trends to determine

S. C. Nandedkar (✉)
Department of Computer Science and Information Technology, Dr. B.A.M. University, Aurangabad, MS, India
e-mail: sugandhanandedkar@dietms.org

J. B. Patil
Department of Computer Engineering, R. C. Patel Institute of Technology, Shirpur, MS, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

M. Saraswat et al. (eds.), *Proceedings of International Conference on Data Science and Applications*, Lecture Notes in Networks and Systems 288,
https://doi.org/10.1007/978-981-16-5120-5_36

the reason for a product failure. The opinion-rich big data help to design business plans.

Despite being rich and valuable, the big data corpus is cursed by the characteristics such as unstructured, incomplete, grammatically weak, and highly webbed [2, 3]. Our goal of the research is to extract structured information from such unstructured text automatically. The opinion word (OW) and opinion target word (TW) can be easily extracted using the opinion sentence's syntactic dependency parser. Along with the extracted OW and TW as the information component, we are immersed in the aspect term (AT) associated with them. The aspect terms are helpful to generate a structured summary of the data [4]. The ATs associated with any topic or product are generally chosen when deciding the aim of text analysis. For example, while processing the review dataset in the restaurant domain, one may be interested in analyzing it concerning the food quality, ambiance, pricing, etc.

The proposed research work addresses extracting the aspect term (AT) associated with the opinion sentences. The result has many imbricate and challenges. Primarily, the TW used in the sentence may not be the exact opinion term or a synonym of it. For example, consider the following sentences:

Review sentence 1	Food was delicious
Review sentence 2	I liked the marvelous taste
Review sentence 3	The fish-curry was outstanding

A human reader can easily find that the words food, taste, and fish-curry are associated with the aspect term food quality. These words are the aspect-related terms (ARTs) used by the reviewer to express the opinion. We need to collect and congregate all such ARTs associated with each AT. Further, note that the ARTs may be a single word or a group of words, i.e., a phrase. Identifying the phrase of appropriate size is the other challenge. Once the ARTs and associated AT are identified, we need to extract the exact sentiment associated with it. The OW present in the opinion sentence is used to extract the same. In a single aspect sentence, it is easy to find the OW associated with the AT. For a multi-aspect sentence, it is difficult to pair an OW with the ART/AT. The task of aspect extraction associated with the review can be carried out either locally or globally. Annotating a document as a political, spiritual, sports, social, etc., is extracting AT globally. In our research work, we focused on local-level AT. It gives the most refined aspect orientation of the corpus.

The next section illustrates the literature review of aspect mining. The literature review pointed the gap present in the existing system. We proposed the semi-supervised algorithm to extract aspect-related terms. The algorithm and its working are explained in Sect. 3. Performance of the algorithm is analyzed using two datasets as explained in Sect. 4. The last section focuses on the conclusion of the research work findings and future scope.

2 Literature Review

Aspect-based opinion mining provides a detailed summary concerning the aspect terms of interest. There are four significant approaches followed [5],

- Aspect extraction based on frequent nouns and noun phrases
- Aspect extraction by exploiting opinion and target relations
- Aspect extraction using supervised learning
- Aspect extraction using topic modeling.

2.1 Aspect Extraction Based on Frequent Nouns and Noun Phrases

This method focuses on the POS tagging and occurrence frequency of the terms. Frequently appearing noun and noun phrases are considered for evaluation. The rest of the terms are discarded. The technique is straightforward and quite effective in depicting highly topic-related terms [6]. Popescu and Etzioni introduced the use of point-wise mutual information (PMI) score to determine the aspect and the discriminator terms [7]. The work proposed by Ramos [8] focused on term frequency-inverse term frequency (TF-IDF) for extracting aspect terms present in the corpus. TF score highlights the frequent terms, whereas IDF score helps to down the general terms and scales up the rare terms. Blair-Goldensohn et al. [9] refined the term frequency count by considering only the opinioned sentences. Their work proposed a simple sentiment analysis where aspect terms are collapsed at the word stem level. Still, it lost the information component present in implicit format. A manually tuned weighted sum ranks the aspects discovered using the proposed system. It relies on the frequency of sentiment-bearing sentences and the type of sentiment phrases/patterns. Cheng et al. [10] proposed a probabilistic aspect mining model (PAMM) to extract and analyze drug reviews. It focuses on finding the terms related to one specific category at a time. The proposed PAMM claims to depict better ARTs than the PMI method. Kim and Chung [11] presented TF-C-IDF as TF-IDF of core terms using core corpus-based weights.

2.2 Aspect Extraction by Exploiting Opinion and Target Relations

The syntactic dependency present in the sentence can depict the aspect terms. Supervised learning techniques, such as dependency parser, double propagation method [12], etc., are used for the same. Rana and Cheah used the Normalized Google Distance (NGD) and sequential patterns for extracting both the explicit and implicit aspects [13]. They further used particle swarm optimization (PSO) along with NGD

for grouping synonyms. The author Asghar et al. used expanded heuristic patterns for aspect extraction. They claimed that the system outperforms the other methods [14]. Vo et al. [15] proposed using a key–value pair window as a primary semantic unit for indicating terms and their relationships in the corpus. A term with its POS tag plays the role of value; the dependency relationship plays the role of the key of the window. The proposed improvement in results by using the dependence parser (DP), co-relationship resolution (CR), and Named Entity Recognition (NER) tools of NLP.

2.3 Aspect Extraction Using Supervised Learning

The mechanism used relabeled data by a human annotator for model training. Hidden Markov model (HMM), conditional random field (CRF), C-NC method, bidirectional long short-term memory (BiLSTM), etc., are used for supervised aspect learning [16]. Laddha and Mukherjee proposed a hybrid generative-discriminative framework for extracting aspect-specific opinion expression. The model uses CRF for discriminative sequence modeling, and relevant terms are gathered using generative models [17]. The researcher claims a significant performance across all the domains. Zuo et al. used the complementary aspect-based mining model (CAMEL) to extract popular and specific aspects through sets while retaining all related opinions for contrastive analysis [18].

2.4 Aspect Extraction Using Topic Modeling

Aspect extraction using topic modeling follows the unsupervised clustering technique to group the documents of a similar type. Poria et al. established a 7-layer deep convolution neural network for the same. The author claims for improved performance with the use of the word embedding model [19]. García-Pablos et al. implemented the word to vector LDA (W2VLDA) model combined with minimal configuration and unsupervised methods. It performs aspect/opinion-word separation, aspect/category classification, and polarity classification of feelings for any given domain and language [20]. Dragoni et al. developed the SMACk framework for extracting the online reviews' opinions. It comprises three main elements, namely the argument module, sentiment module, and visualization module. The integration behavior for a formal theory in argumentation resulted in improved performance. Further, it also helped consumers in making more informed decisions in online sales [21].

2.5 Gap Analysis

Using the term frequency for aspect determination helps to extract the preliminary information from the corpus. For the more detailed summarized analysis, it does not prove better. Similarly, the aspect extraction performance by exploiting opinion and target relations is limited by the dependency parser's performance. Supervised learning requires pre-labeled data annotated by a human annotator. Preparing labeled data for supervised model training is a tedious task. To train a model to extract aspects from any domain, we need labeled data of that domain. We need to mention the aspect terms and non-aspect terms explicitly. Pre-labeling of samples required for supervised learning is a costly process. We need a machine learning engineer or a data scientist for the manual annotation task. On the other hand, the 'limited application spectrum' is the disadvantage related to unsupervised learning.

In our proposed method, we are using the semi-supervised approach for aspect mining. The semi-supervised learning method uses a handful of pre-labeled data, and the rest of the data is unlabeled. The program iterates through the corpus to collect the terms associated with pre-labeled terms. Thus, the set of labeled terms goes on increasing gradually. We adopted this as a bootstrapping method.

3 Proposed Method of Aspect Extraction Using Semi-supervised Learning

As mentioned above, we are using the semi-supervised approach for aspect mining. In the pre-work stage, we divide the entire corpus (C) into documents (d_i). Further, divide each document into sentences (S_{ij}) representing jth sentence in ith document and divide each sentence into individual word tokens as W_{ijk}. Find POS tag for each word token as (P_{ijk}). Using the dependency parser, find out syntactic dependencies among the different word tokens present in each sentence. We need to consider the fact that the dependency parser does not interpret all sentences. For example, the parser cannot attend the grammatically weak sentences or incomplete sentences. Still, we have its tokens' POS tag. The dependency relationship is helpful for further analysis.

The nouns and adjectives are tentatively considered as features, and opinion words, respectively. To extract an aspect associated with a sentence, use the feature related to the sentence. To fetch the feature word's sentiment orientation in the sentence using the opinion word associated with the extracted feature word, we need to co-extract these word pairs from each sentence. To grab valid word pairs, Wang and Wang [3] proposed the Revised Mutual Information (RMI) score for the word pairs represented by Eq. 1. Note that they have used only the noun (NN*) and adjective (JJ*) word pairs. They missed out on a few vital pieces of information as few noun pairs frequently occur together. Similarly, few adjective word pairs frequently occur together. Including them as a feature–opinion (FO) pair, feature–feature (FF) pair, and

Table 1 Contingency table

	W_2	$\sim W_2$
W_1	A	B
$\sim W_1$	C	D

opinion–opinion (OO) pair can help to mine valuable information from the corpus.

$$\text{RMI}(W_1, W_2) = A \times \log \frac{N \times A}{(A + B) \times (A + C)} \tag{1}$$

Equation 1 shows that the RMI score is calculated based on five factors A, B, C, D, and N. These factors are computed using the contingency table mentioned in Table 1.

The contingency table is calculated for each word pair (W_1, W_2) of FO, FF, and OO.

A: Co-occurrence frequency of W_1 and W_2 is a number of sentences where W_1 and W_2 have appeared together.
B: Number of sentences where W_1 is present, but W_2 does not occur.
C: Opposite to B, that is a number of sentences where W_2 is current, but W_1 does not appear.
D: Number of sentences where both W_1 and W_2 have not occurred.
N: Is the summation of all A, B, C, and D.

Initially, collect all the nouns and adjectives present in the corpus as candidate features (Cand_Feature) and candidate opinions (Cand_Opinion), respectively. The finalized set of features and opinion lexicons is named (Final_Feature) and (Final_Opinion). Following a semi-supervised approach for aspect mining, we pick up a few terms as seed terms. Initially, a few aspects of the entity are considered as initial values in Final_Feature. We can also use some general adjective terms as initial values in Final_Opinion.

Each candidate feature present in Cand_Feature is paired with each opinion term present in Final_Opinion. Depict the RMI score for each pair. If the pair's RMI score is above the pre-specified threshold value, then include that feature word into ResFeaLex and exclude it from Cand_Feature. Similarly, each opinion term present in Cand_Opinion is paired with each term present in Final_Feature. Depict RMI score for each such pair. If the RMI score of a pair is above the pre-specified threshold, then include that opinion term in the Final_Opinion and exclude it from Cand_Opinion. Continue this process until all terms in the Cand_Feature and Cand_Opinon are not empty. While iterating, if neither a new feature nor an opinion term is learned, break the loop. Thus, we end up with a quality feature and opinion terms learned from the corpus by following a semi-supervised approach. As we have to feed the algorithm a few entity-related aspect terms and corresponding opinion words, the resulting feature and opinion terms are related to the specified entity. The pseudocode for the algorithm is as follows.

Pseudocode for SeAbOM: Semi-Supervised Learning for Aspect based Opinion Mining

```
Semi_Supervised_Aspect_OM(R, KB, RMI, Fea_th, Op_th)
/* Input customers' review dataset as R, precomputed RMI
matrix. candidate feature – opinion terms, KnowledgeBase
(KB) representing seed terms. Fea_th and Op_th are the
thresholds for feature terms and opinion terms. */
/*Output is the set of finalized features and opinion
terms.*/

Preprocess and parse the review dataset R
Set all unique noun terms as Cand_Fea and all unique ad-
jective terms as Cand_Op sets
Set the terms from Knowledgebase as seed terms in Fi-
nal_Fea
While Cand_Fea and Cand_Op are not empty
  For each cand_op_term in Cand_Op
       Score(cand_op_term) = (∑featureϵFinal_Fea RMI(feature,
       cand_op_term))/|Final_Fea|
       If (Score(cand_op_term)) > Op_th then
         Remove cand_op_term from Cand_op and
         Insert cand_op_term in Final_Op
  For each cand_fea_term in Cand_Fea
       Score(cand_fea_term) = (∑opinionϵFinal_Op
         RMI(cand_fea_term, opinion))/ |Final_Op|
       If (Score(cand_fea_term)) > Fea_th then
         Remove cand_fea_term from Cand_Fea and
         Insert cand_fea_term in Final_Op
  If no new term is learned in the present iteration,
  then break
Return Final_Fea and Final_Op as learned terms
```

4 Experimental Setup and Performance Analysis

We used the customer review dataset[1] and SemEval 2016 Reviews-English[2] to test our system's working and performance. Dataset 1 contains annotated reviews of five products: cellular phone, MP3 player, DVD player, and two digital cameras. Manually identified sentiment polarity and feature words are tagged with each sentence. Dataset 2 contains a gold standard labeled textual dataset reviews for laptops and restaurants.

Table 2 describes the statistics of the datasets. It includes a number of reviews, sentences, and polarity distribution of the same. Table 3 describes dependency

[1] https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#datasets.

[2] http://alt.qcri.org/semeval2016/task5/index.php?id=data-and-tools.

Table 2 Dataset description

	Dataset 1					Dataset 2	
	Digital camera 1	Digital camera 2	Cellular phone	MP3 player	DVD player	Laptops	Restaurant
No. of reviews	45	34	41	95	100	396	350
No. of sentences	597	346	546	1716	739	2373	2000
No. of positive sentences	184	130	192	430	151	1049	1114
No. of negative sentences	55	30	73	290	193	628	516
No. of neutral sentences	358	186	281	996	395	696	370

Table 3 Relationships extracted from the datasets

	Dataset 1					Dataset 2	
	Digital camera 1	Digital camera 2	Cellular phone	MP3 player	DVD player	Laptops	Restaurant
No. of relationships extracted	11,191	6163	8830	30,128	11,602	9483	43,199
No. of amod relationships	776	435	527	1587	624	771	2893
No. of nsubj relationships	1096	619	938	3140	1224	1148	4573
No. of avmod relationships	721	390	615	2083	810	657	3114
No. of dobj relationships	620	376	454	1722	764	573	1629
No. of other relationships	7978	4343	6296	21,596	8180	6334	30,990

relationships extracted from the sentences. It is observed that dependencies like the nominal subject (nsubj), adjective modifier (avmod), direct object (dobj), and adverb modifier (avmod), etc. are more useful as compared to other dependencies for feature–opinion pair extraction.

We evaluate our system's performance from the perspectives effectiveness of feature–opinion word pair co-extraction and effectiveness of extracting aspect category associated with the feature word. We used spaCy library for NLP in Python. We used the seed word mentioned in [20]. The various results we obtained are described

in Tables 4, 5, and 6. We used a confusion matrix from scikit-learn library to evaluate the performance of our system. It consists of four types of values: true positive, true negative, false positive, and false negative. Using the confusion matrix, we can calculate the various parameters like accuracy, precision, recall, and f-measure. For our work, we have focused only on precision and recall values. Table 4 shows the precision of obtained results for both datasets. Whereas Table 5 shows the recall for the same. We noticed that the selection of threshold values impacts the performance of the system. Table 6 shows our proposed system's performance analysis with the other state-of-the-art method mentioned in [15]. Note that the performance of our approach is compared with the available results only.

Table 4 Precision of semi-supervised aspect learner

Threshold	Dataset 1					Dataset 2	
	Digital camera 1	Digital camera 2	Cellular phone	MP3 player	DVD player	Laptops	Restaurant
F_th > 10, O_th > 10	0.59	0.56	0.52	0.44	0.46	0.65	0.57
F_th > 1, O_th > 0.2	0.64	0.59	0.45	0.36	0.43	0.61	0.49
F_th > 5, O_th > 1	0.67	0.62	0.48	0.41	0.61	0.69	0.65

Table 5 Recall of semi-supervised aspect learner

Threshold	Dataset 1					Dataset 2	
	Digital camera 1	Digital camera 2	Cellular phone	MP3 player	DVD player	Laptops	Restaurant
F_th > 10, O_th > 10	0.74	0.69	0.72	0.62	0.65	0.71	0.73
F_th > 1, O_th > 0.2	0.85	0.78	0.9	0.93	0.72	0.84	0.71
F_th > 5, O_th > 1	0.88	0.83	0.87	0.91	0.72	0.87	0.73

Table 6 Performance analysis of semi-supervised aspect learner

Product name	Opinion–aspect relation		Semi-supervised aspect learning	
	Precision	Recall	Precision	Recall
Digital camera	0.63	0.82	0.67	0.88
Laptop	0.77	0.83	0.69	0.87

5 Conclusion and Future Scope

In this paper, we proposed a semi-supervised approach for aspect learning. The performance of unsupervised learning is limited. It is dependent upon either the occurrence frequency or the syntactic relationships among the tokens. Thus, their performance is highly affected by the lacunas associated with the unstructured text corpus. In comparison, the supervised approach required a labeled data for model training. As mentioned above, we need a team of human annotators for the same. Our proposed method requires just a handful of seed terms. We named them aspect terms. By using the bootstrapping approach mentioned in the semi-supervised learning algorithm, our model learned more ARTs related to each AT. The experimental results showed that semi-supervised learner works more efficiently as compared to other methods.

The obtained results enlightened us with many future research directions and uncovered problems. For instance, there are a limited set of standard datasets available for aspect-based opinion mining task. Also, most of the datasets are either in English or Chinese. Standard datasets of other languages are not available. Other issues associated with opinion mining lies with sarcastic statements and fake statements. Hence, these findings show that a gold standard dataset in various domains is needed along with the attention on hybrid learning techniques.

References

1. C. Christian, I. Weismayer, Pezenka, *Aspect-Based Sentiment Detection: Comparing Human Versus Automated Classifications of TripAdvisor Reviews* (Springer International Publishing, 2018)
2. S. Nandedkar, J. Patil, Co-extracting feature and opinion pairs from customer reviews using hybrid approach, in *IEEE I2CT* (2018), pp. 769–773
3. B. Wang, H. Wang, Bootstrapping both product features and opinion words from Chinese customer reviews with cross-inducing, in *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence WI 2007*, no. 60675035 (2007), pp. 259–262
4. M. Tubishat, N. Idris, M.A.M. Abushariah, Implicit aspect extraction in sentiment analysis: review, taxonomy, opportunities, and open challenges. Inf. Process. Manag. **54**, 545–563 (2018)
5. B. Liu, Sentiment analysis and opinion mining. Synth. Lect. Hum. Lang. Technol. (2012)
6. B. He, I. Ounis, A study of the Dirichlet priors for term frequency normalisation, in *SIGIR 2005—Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (2005), pp. 465–471
7. A.M. Popescu, O. Etzioni, Extracting product features and opinions from reviews, in *HLT/EMNLP 2005—Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference* (2005), pp. 339–346
8. J. Ramos, Using tf-idf to determine word relevance in document queries. *Proceedings of the First Instructional Conference on Machine Learning* (2003), Vol. 242. No. 1
9. S. Blair-Goldensohn et al., Building a sentiment summarizer for local service reviews, in *NLPIX* (2008)
10. V.C. Cheng, C.H.C. Leung, J. Liu, A. Milani, Probabilistic aspect mining model for drug reviews. IEEE Trans. Knowl. Data Eng. **26**, 2002–2013 (2014)

11. J.C. Kim, K. Chung, Associative feature information extraction using text mining from health big data. Wirel. Pers. Commun. **105**, 691–707 (2019)
12. G. Qiu, B. Liu, J. Bu, C. Chen, Opinion word expansion and target extraction through double propagation. Comput. Linguist. **37**(1), 9–11 (2011)
13. T.A. Rana, Y.-N. Cheah, Hybrid rule-based approach for aspect extraction and categorization from customer reviews, in *9th International Conference on IT in Asia (CITA), Proceedings of the Conference* (2015)
14. M.Z. Asghar, A. Khan, S.R. Zahra, S. Ahmad, F.M. Kundi, Aspect-based opinion mining framework using heuristic patterns. Clust. Comput. **22**, 7181–7199 (2019)
15. A.D. Vo, Q.P. Nguyen, C.Y. Ock, Opinion-aspect relations in cognizing customer feelings via reviews. IEEE Access **6**, 5415–5426 (2018)
16. E. Riloff, R. Jones, Learning dictionaries for information extraction by multi-level bootstrapping, in *American Association for Artificial Intelligence (AAAI-99) Proceedings* (1999)
17. A. Laddha, A. Mukherjee, Extracting aspect specific opinion expressions, in *EMNLP 2016—Conference on Empirical Methods in Natural Language Processing, Proceedings* (2016), pp. 627–637
18. Y. Zuo et al., Complementary aspect-based opinion mining across asymmetric collections, in *Proceedings—IEEE International Conference on Data Mining, ICDM 2016*, Jan 2016, pp. 669–678
19. S. Poria, E. Cambria, A. Gelbukh, Aspect extraction for opinion mining with a deep convolutional neural network. Knowl.-Based Syst. **108**, 42–49 (2016)
20. A. García-Pablos, M. Cuadros, G. Rigau, W2VLDA: almost unsupervised system for aspect based sentiment analysis. Expert Syst. Appl. **91**, 127–137 (2018)
21. M. Dragoni, C. Da Costa Pereira, A.G.B. Tettamanzi, S. Villata, Combining argumentation and aspect-based opinion mining: the SMACk system. AI Commun. **31**, 75–95 (2018)