An Unsupervised Approach for Aspect Category Detection Using Soft Cosine Similarity Measure

Preprint · December 2018

4 authors, including:

Sajad Movahedi
University of Tehran
8 PUBLICATIONS **15** CITATIONS
SEE PROFILE

Heshaam Faili
University of Tehran
110 PUBLICATIONS **688** CITATIONS
SEE PROFILE

An Unsupervised Approach for Aspect Category Detection Using Soft Cosine Similarity Measure

Erfan Ghadery[*], **Sajad Movahedi**[*], **Heshaam Faili, Azadeh Shakery**
University of Tehran
Tehran, Iran
{erfan.ghadery, s.movahedi, hfaili, shakery}@ut.ac.ir

Abstract

Aspect category detection is one of the important and challenging subtasks of aspect-based sentiment analysis. Given a set of pre-defined categories, this task aims to detect categories which are indicated implicitly or explicitly in a given review sentence. Supervised machine learning approaches perform well to accomplish this subtask. Note that, the performance of these methods depends on the availability of labeled train data, which is often difficult and costly to obtain. Besides, most of these supervised methods require feature engineering to perform well. In this paper, we propose an unsupervised method to address aspect category detection task without the need for any feature engineering. Our method utilizes clusters of unlabeled reviews and soft cosine similarity measure to accomplish aspect category detection task. [1] Experimental results on SemEval-2014 restaurant dataset shows that proposed unsupervised approach outperforms several baselines by a substantial margin.

1 Introduction

User-generated reviews are valuable resources for both companies and potential customers. These reviews can help online retail companies in recognizing their weaknesses and strengths, and facilitate the decision process for consumers. However, it is impractical and tedious to read such a huge amount of reviews one by one manually. Therefore, the need for an automatic system that processes a huge amount of reviews and provides a summary of these different reviews in a suitable form seems necessary.

Given a list of pre-defined aspect categories (e.g. 'food' and 'price' in restaurant domain), aspect category detection aims to assign a subset of these categories to review sentences. SemEval 2014 Task 4 (Pontiki et al., 2014) tried to tackle this task by providing datasets for multiple domains, including restaurant and laptop reviews.

Previous works in the literature mostly are based on supervised machine learning approaches (Çetin et al., 2016)(Kiritchenko et al., 2014)(Castellucci et al., 2014). Although these methods perform well, they need annotated train data which are time-consuming and expensive to obtain. Thus, unsupervised approaches maybe be a good choice, especially for low resource languages.

Soft cosine measure is a similarity measure that assesses the similarity between two sentences, even when they have no words in common(Sidorov et al., 2014). For aspect category detection, our method utilizes soft cosine similarity. Firstly, we cluster a set of unlabeled review sentences into k cluster. Clustering is performed based on the Euclidean distance between the average of their word embeddings. Our motivation for using the cluster of sentences is based on the intuition that sentences in the same cluster share similar information about categories they belong to. The similarity between a given sentence and a pre-defined category is defined as the soft cosine similarity between sentence and a set of manually selected seed words corresponding to that category. So, the similarity values can give us information about categories that sentence belongs to. We also define similarity between a cluster and a category as averaging the similarity scores of the sentences in the cluster. Finally, given a test review sentence, scores obtained for the sentence and the nearest cluster to it are interpolated. These final scores are normalized and used to detect the categories mentioned in the sentence. If the similarity of a category surpasses a threshold, it is assigned to the sentence.

[*] Equal Contribution.

[1] We have made our code available at https://github.com/erfan-ghadery/Unsupervised-Aspect-Category-Detection.

We evaluate our method on SemEval-2014 restaurant dataset. Experimental results show that the proposed method outperforms several baselines by a high margin.

2 Related works

Early works for addressing aspect extraction relied on approaches such as identifying frequent nouns and noun phrases using association rule mining, dependency relations, and lexical patterns(Hu and Liu, 2004)(Qiu et al., 2011)(Popescu and Etzioni, 2007). The SemEval workshops, over the course of three years, has included aspect-based sentiment analysis in their competitions. One of the subtasks introduced during SemEval is aspect category detection, which our proposed method is going to address. Most of the supervised approaches proposed to address this subtask utilizing machine learning algorithms and train a set of the one-vs-all classifier using hand-crafted features (Kiritchenko et al., 2014), (Xenos et al., 2016), (Toh and Su, 2016).

There are only a few unsupervised approaches to address the aspect category detection subtask in the literature. In (He et al., 2017) authors trained a network similar to auto-encoder with attention mechanism to attend to aspect-related words. (Pablos et al., 2014) propose to use a double propagation technique to mine rules based on dependency relations for finding aspect terms of each category. Also, irrelevant aspect terms were pruned using stop words and the PageRank algorithm on a graph-based approach. (Schouten et al., 2018) proposed an unsupervised method called 'spreading activation' that performs association rule mining, using a set of seed words and a co-occurrence matrix between words to form a co-occurrence digraph to detect aspect categories. One of the shortcomings of this method is the need for tuning multiple parameters. Unlike this method, our proposed method only requires a small number of parameter (a threshold, the number of clusters, and the interpolation coefficient).

3 Proposed Method

This section describes our proposed unsupervised method. In the following subsections, we will discuss each of the main components in detail.

3.1 Manual Selection of Category Seed Words

We manually select a set of 5 seed words for each category (20 in total) to represent the category. Because the anecdotes/miscellaneous category is very unspecific and abstract, we didn't choose any seed words for it(Schouten et al., 2018). For this aspect, following (Schouten et al., 2018), we would assign this category only to sentences that were not assigned any other category. The set of seed words for each category can be seen in Table 1.

Table 1: List of the seed words for each category.

Category	Seed Words
food	food, delicious, menu, fresh, tasty
service	service, staff, friendly, attentive, manager
price	price, cheap, expensive, money, affordable
ambience	ambience, atmosphere, decor, romantic, loud

3.2 Sentence similarity

In order to find the similarity score of a given sentence compares to a category, we utilize soft cosine measure. For each category, we define the similarity of the given sentence to that category as the average of soft cosine similarity values between the sentence and each of the seed words belonging to that category. Let x be a given sentence and a_i be the i-th category. We define the $sentSim_{a_i}(x)$ to be the similarity value between a_i and x as seen in equation 1.

$$sentSim_{a_i}(x) = \frac{\sum_{i=1}^{|s|} softcossim(x, s_i)}{|s|} \quad (1)$$

As stated in (Zamani and Croft, 2016), the sigmoid function can have a discriminating effect on the similarity values obtained from similarity measures like cosine similarity. To make the similarity values more discriminating, we transferred the similarity values obtained in the previous step via the sigmoid function as seen in equation 2.

$$sentScore_{a_i}(x) = \frac{e^{sentSim_{a_i}(x)}}{1 + e^{sentSim_{a_i}(x)}} \quad (2)$$

Now, for each sentence we have a vector $sentScore \in \mathbb{R}^c$, which c is the number of categories and each element represents the similarity score between the sentence and a pre-defined category.

3.3 Cluster similarity

A set of unlabeled sentences are acquired from the Yelp dataset challenge [2]. In order to decrease noise samples and since the precision is a more important factor than recall in acquiring true sentences, only sentences that contain at least one of the category names (eg. 'food', 'service') are selected. Using k-means clustering algorithm, these unlabeled sentences are clustered into k clusters. Clustering is done based on the Euclidean distance between the average of word embedding of sentence words, where word embeddings are trained on the Yelp dataset using Continuous Bag of Words (CBOW) algorithm. For each cluster, a similarity value per category is calculated as shown in equation 3:

$$clustSim_{a_i}(c_k) = \frac{\sum_{x \in c_k} sentSim_{a_i}(x)}{|c_k|} \quad (3)$$

where c_k is the k-th cluster. Similar to sentence similarity, we also utilize sigmoid function in here for it's discriminating effect.

$$clustScore_{a_i}(x) = \frac{e^{clustSim_{a_i}(x)}}{1 + e^{clustSim_{a_i}(x)}} \quad (4)$$

Therefore, for each cluster we have a vector $clustScore \in \mathbb{R}^c$, which c is the number of categories and each element represents the similarity between cluster samples and one of the predefined categories.

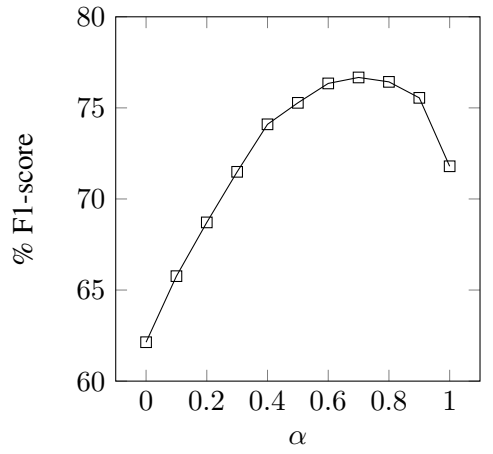
3.4 Assign aspect categories

Given a test review sentence, first, we calculate the sentScore vector for the sentence, then, this vector is interpolated with the clustScore corresponding to the nearest cluster to the given sentence, as shown in equation 5. The nearest cluster is found by finding the closest centroid to the sentence based on the Euclidean distance between the average of the word embeddings of the sentence and the cluster centroid.

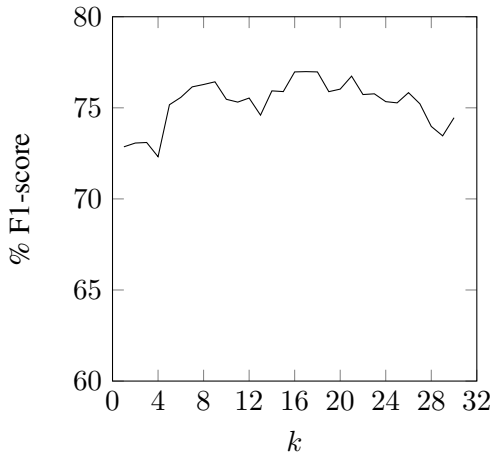
$$score = \alpha sentScore_N + (1 - \alpha) clustScore_N \quad (5)$$

where $sentScore_N$ and $clustScore_N$ are L2-Normalized vectors of sentence and its cluster respectively. Categories that exceed a threshold are assigned to the sentence. We find the optimal threshold by a simple linear search.

[2] https://www.yelp.com/dataset/challenge



(a) Interpolation coefficient α sensitivity



(b) The effect of number of clusters k.

Figure 1: The effect of hyperparameters.

4 Experiments

We train word embeddings on the Yelp challenge unlabeled dataset. Word embedding size is set to 300 and CBOW model (Mikolov et al., 2013) with default parameters is used for training the embedding vectors. Removing stop-words and tokenizing sentences are performed as the pre-processing step using NLTK package (Bird et al., 2009). For implementing soft cosine similarity measure and word embedding training, we used gensim package (Řehůřek and Sojka, 2010). A simple linear search is performed to find the optimal hyperparameters α and K - the interpolation coefficient between sentence score and its cluster score, and the number of clusters for k-means clustering respectively. We set the parameter α to 0.7 and the number of clusters to 17 to obtain the best result in our method.

4.1 Results and Analysis

Micro F1-score, precision, and recall of all the category labels are used as evaluation metrics and performance of the proposed method evaluated on test data from SemEval-2014(Pontiki et al., 2014). The test data contains 800 test review sentence in the restaurant domain.

The proposed method is compared to the following baselines. **Random** baseline assigns a random category based on the frequency of the categories appearing in the train data. **Majority** baseline assigns the two most common categories in the train data ('food' and 'anecdotes/miscellaneous') to each of the test sentences. **COMMIT-P1WP3** (Schouten et al., 2014) baseline is a co-occurrence based method. After constructing a co-occurrence matrix between the words and the pre-defined categories, this method calculates the probability of a given sentence belonging to each category based on this matrix. **V3** (Pablos et al., 2014) baseline utilizes WordNet similarity to compare the detected aspect terms to representative terms of each category. **SemEval-2014 baseline** (Pontiki et al., 2014) is the baseline provided by the SemEval 2014 which uses a simple k-nearest neighbor classifier to detect aspect categories. In the **Spreading Activation** baseline (Schouten et al., 2018), using a set of seed words and a co-occurrence matrix similar to (Schouten et al., 2014), a set of association rules are mined, associating categories to terms appearing in the data. The aspect category detection task is then performed using these association rules.

Table 2 shows that among baseline methods, Spread Activation, an unsupervised approach, performs the best with F1-score 67%. Our unsupervised method outperforms Spread Activation by 9.98% in terms of F1-score and similarly improves over SemEval-2014 baseline, V3, COMMIT-P1WP3, Majority, and Random baselines by 13.08%, 16.78%, 17.68%, 27.31% and 47.30% respectively. These results clearly demonstrate the effectiveness of the proposed unsupervised method in aspect category detection task.

4.2 Hyperparameter tuning

Figure 1 (a) plots the sensitivity of our system to the interpolation coefficient α in equation 5. According to the curve, to achieve the best performance a higher weight should be given to sentence scores itself, which indicates that it plays the main role in finding corresponding categories. The results of the experiments prove our intuition that interpolating the cluster scores and the sentence scores improves the classification performance. The reason behind this improvement can be the sentences that contain a seed word or a semantically close word to the seed words of a category. Such sentences will improve the scores of other sentences in the cluster for that category. The best result was obtained at $\alpha = 0.7$. We also conducted experiments on the number of clusters in the k-means clustering algorithm. Figure 1 (b) provides the results of these experiments. We found the optimum value for k to be 17. As can be seen in the figure, the result of our method tends to deteriorate with k larger than 17. This can be due to the lowering consistency of clusters with larger k. Also, for the k smaller than 9, the same thing can happen.

Table 2: The result of baseline methods compared to our method. The COMMIT-P1EP3 method and the SemEval 2014 method are supervised methods (S) and the V3 and Spreading Activation method are unsupervised methods (U).

Method	precision	recall	F_1
Random	34%	26.34%	29.68%
Majority	40.75%	63.60%	49.67%
COMMIT-P1WP3 ^{S}	63.3%	55.8%	59.3%
V3 ^{U}	63.3%	56.9%	60.2%
SemEval-2014 baseline ^{S}	-	-	63.9%
Spreading Activation ^{U}	69.5%	64.7%	67.0%
Our method	**82.97%**	**71.80%**	**76.98%**

Conclusions

In this paper, we propose an unsupervised approach for aspect category detection that utilizes soft cosine similarity and the well-known k-means clustering to detect categories belonging to a review sentence. Experimental results on SemEval-2014 benchmark dataset show the effectiveness of our method compared to several supervised and unsupervised baseline methods.

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

Giuseppe Castellucci, Simone Filice, Danilo Croce, and Roberto Basili. 2014. Unitor: Aspect based

sentiment analysis with structured learning. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 761–767.

Fatih Samet Çetin, Ezgi Yıldırım, Can Özbey, and Gülşen Eryiğit. 2016. Tgb at semeval-2016 task 5: Multi-lingual constraint system for aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 337–341.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2017. An unsupervised neural attention model for aspect extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 388–397.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

Svetlana Kiritchenko, Xiaodan Zhu, Colin Cherry, and Saif Mohammad. 2014. Nrc-canada-2014: Detecting aspects and sentiment in customer reviews. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 437–442.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Aitor García Pablos, Montse Cuadros, and German Rigau. 2014. V3: Unsupervised generation of domain aspect terms for aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 833–837.

Maria Pontiki, Dimitrios Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. pages 27–35.

Ana-Maria Popescu and Orena Etzioni. 2007. Extracting product features and opinions from reviews. In *Natural language processing and text mining*, pages 9–28. Springer.

Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. 2011. Opinion word expansion and target extraction through double propagation. *Computational linguistics*, 37(1):9–27.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

Kim Schouten, Flavius Frasincar, and Franciska De Jong. 2014. Commit-p1wp3: A co-occurrence based approach to aspect-level sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 203–207.

Kim Schouten, Onne van der Weijde, Flavius Frasincar, and Rommert Dekker. 2018. Supervised and unsupervised aspect category detection for sentiment analysis with co-occurrence data. *IEEE transactions on cybernetics*, 48(4):1263–1275.

Grigori Sidorov, Alexander Gelbukh, Helena Gómez-Adorno, and David Pinto. 2014. Soft similarity and soft cosine measure: Similarity of features in vector space model. *Computación y Sistemas*, 18(3):491–504.

Zhiqiang Toh and Jian Su. 2016. Nlangp at semeval-2016 task 5: Improving aspect based sentiment analysis using neural network features. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 282–288.

Dionysios Xenos, Panagiotis Theodorakakos, John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2016. Aueb-absa at semeval-2016 task 5: Ensembles of classifiers and embeddings for aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 312–317.

Hamed Zamani and W Bruce Croft. 2016. Embedding-based query language models. In *Proceedings of the 2016 ACM international conference on the theory of information retrieval*, pages 147–156. ACM.