Evaluating the Aspect-Category-Opinion-Sentiment task on a custom dataset*

Loris Di Quilio[1]

[1]DEc, University 'G. d'Annunzio', Chieti-Pescara, Italy

Abstract

In this work, we report the results of some experiments with Aspect Based Sentiment Analysis (ABSA) on a dataset consisting of user reviews of products of a manufacturing company operating in the beauty industry. We focus on one of the more challenging ABSA tasks, the Aspect Category Opinion Sentiment task, and compare the results obtained by using three different tools.

Keywords

Aspect-based Sentiment Analysis, Aspect Category Opinion Sentiment

1. Introduction

Sentiment analysis aims to determine and understand the opinion sentiment expressed in a text . The basic approach performs this analysis prediction at the sentence or document level, identifying the overall sentiment of the sentence or whole document. In this case, it is assumed that a single sentiment is associated with a single topic in the text, but that may not always be the case. For this reason, a fine-grained sentiment analysis named *Aspect Based Sentiment Analysis* (ABSA), has received increasing attention. In this task, the objective includes identifying which specific aspects or features the sentiments refer to.

Aspect-based Sentiment Analysis (ABSA) is a fine-grained sentiment analysis where the goal is to identify the aspects of given target entities and the sentiment expressed for each aspect [1, 2, 3]. Over the years research in ABSA has specialized in various sub-tasks based on the prediction characteristic of a single sentiment element or of several ones together[4, 5]. The components [1] of these tasks are the following:

- **category (c)**: is a pre-defined category related to a specific domain of interest. For example, AMBIENCE, PRICE, and FOOD are examples of categories for the *restaurant* domain.
- **aspect term (a)**: represents the specific opinion target explicitly mentioned in the provided text. For instance, in the sentence "*The pizza is delicious but the service is terrible*", the explicit aspects are "pizza" and "service". When this is implicit, as in the

AIxIA'23: 22nd International Conference of the Italian Association for Artificial Intelligence, November 06–09, 2023, Rome, Italy

*DEc, University 'G. d'Annunzio', Chieti Pescara, Italy

loris.diquilio@studenti.unich.it (L. D. Quilio)

[1]the nomenclature of the four components and tasks could be different in the various works.

sentence "*it's very reasonably priced*", when the subject is not explicitly named, the value of the aspect term is "NULL".

- **opinion term (o)**: is the word, or the words, used by opinion users to convey their sentiments or feelings about the target entity or aspect. For example, in the sentence "*The pizza is delicious but the service is terrible*", "*delicious*" and "*terrible*" are opinion terms, expressing a positive and negative sentiment toward the pizza and the service, respectively.
- **polarity (p)**: characterizes the sentiment orientation expressed towards an aspect category or an aspect term. Sentiment polarity can be positive, negative, or neutral indicating that the sentiment is favorable, unfavorable, or neither, respectively.

Among the tasks of Aspect-based Sentiment Analysis that aim to predict a single sentiment element, there are:

- **Aspect Term Extraction** (ATE);
- **Aspect Category Detection** (ACD);
- **Opinion Term Extraction** (OTE);
- **Aspect opinion co-extraction** (AOCE);
- **Aspect Sentiment Classification** (ASC).

The tasks where multiple sentiment elements are predicted include:

- **Aspect-Opinion Pair Extraction** (AOPE);
- **End-to-End ABSA** (E2E-ABSA);
- **Aspect Category Sentiment Analysis** (ACSA);
- **Aspect Sentiment Triplet Extraction** (ASTE);
- **Aspect Category Sentiment Detection** (ACSD);
- **Aspect Category Opinion Sentiment** (ACOS).

Following we show a summary of the tasks using the input sentence: "The pizza is delicious but the service is terrible".

Task	Input	Output
ATE	sentence	pizza (a), service (a)
ACD	sentence	food (c), service(c)
OTE	sentence	delicious (o), terrible (o)
ASC	sentence, pizza sentence, service	positive(p) negative (p)
AOPE	sentence	{pizza (a), delicious (o)}, {service (a), terrible (o)}
E2E ABSA	sentence	{pizza (a), positive p)}, {service (a), negative (p)}
ACSA	sentence	{food (c), positive (p)}, {service (c), negative (p)}
ASTE	sentence	{pizza (a), positive (p), delicious (o)}, {service (a), negative (p), terrible (o)}
ACSD	sentence	{food (c), pizza (a), positive (p)}, {service (c), service (a), negative (p)}
ACOS	sentence	{pizza (a), food (c), delicious (o), positive (p)}, {service (a), service (c), terrible (o), negative (p)}

In this paper, we will focus our attention on the ACOS task which aims at predicting all the sentiment information at once, namely category (c), aspect term (a), opinion term (o), and polarity (p). For the ACOS task, a relatively limited body of research and literature exists. Our primary objective is to establish an integrated framework that leverages multiple tools for efficient ACOS task execution.

2. Dataset and annotation

Concerning the annotations, there are differences with the datasets available in the literature due to the many implicit aspects referring to packaging and opinion terms often composed from multiple words.

The dataset is composed of a training and test set, and each sentence can have multiple annotations.

	Train	Test	Total
Sentences	623	133	756
Annotations	881	157	1038

Table 1
Number of sentences and annotations in the training and test datasets

The composition appears balanced in terms of positive and negative polarity (*p*), with neutral sentiment not being of interest. As regards the categories, thirteen classes were identified, mostly balanced, except for the category belonging to ”general satisfaction of the final consumer”.

For this work, a custom template in *Label Studio* was built, which allows all elements of interest to be annotated for each review. In Figure 1 we show an example of a sentence annotated with this annotation tool: the explicitly mentioned aspect and opinion elements can be directly selected in the text, while the polarity and the category, which is not shown, can be chosen from the predefined ones. A translation module has been developed to convert the JSON encoding of the dataset exported from Label Studio to other formats.

Figure 1: Example of a sentence annotated with Label Studio

3. Experimental evaluation

In this section, we present the details of the experimental evaluation we performed on our dataset using some tools that have been specifically built for the ACOS task. We have selected three tools that have stemmed from significant studies in this field and for which the source code is publicly available online. All the selected tools leverage the fine-tuning of pre-trained models, specifically T5 [6, 7] and BERT[8], as a crucial component of their functionality:

- **Paraphrase modeling** [9]: the model's objective is to generate a sequence of words, denoted as y, from an input sentence x. The sequence y should contain all the desired sentiment elements. Once the sequence y is generated, it's possible to recover the so-called "sentiment quads" $Q = (a, c, o, p)$. This approach aims to fully leverage the semantics of the sentiment elements represented by Q by generating them in natural language form within the sequence y. The pre-trained language model used is ***T5-base***. This is the only tool among those we have considered that does not support implicit opinion terms;
- **Extract Classify-ACOS** [10]: This tool first performs aspect-opinion co-extraction, then predicts category-sentiment given the extracted aspect-opinion pairs. The tool uses the BERT model with AdamW optimizer[2] [11], so the data is transformed into a format suitable for it by inserting the token CLS[3] at the beginning and at the end of each sentence;
- **PyABSA** [12]: this tool is a variation of the original one, made for aspect-opinions pair extraction. There is no documentation about quadruple extraction because this feature is still experimental. The format of this tool was taken as a reference to transform the data once exported from the annotation tool. Also in this case *T5-base* is used as the pre-trained model.

In the table, we show the settings we used for the experiments for each tool.

Tool	batch-size	learning rate	epochs
Paraphrase modeling	16	3e-4	20
Extract Classify-ACOS	32{*a, o*}, 16(*p*), 8(*c*)	2e-5{*a, o*}, 3e5(*p*),(*c*)	20
PyABSA	16	5e-5	20

Table 2
Tool settings used for training

The second experiment is motivated by the fact that sentences in our domain often contain implicit opinions, frequently composed of multiple words rather than single terms. So we established a relaxed correctness criterion for considering a prediction correct when it matches the gold standard in terms of aspect, category, and polarity, and when the similarity between the predicted opinion term and the real one is at least 70%. For computing string similarity we used the SequenceMatcher[4] Python function, which compares pairs of sequences by finding the longest common subsequence while excluding uninteresting elements, with a quadratic time

[2]AdamW optimizer: is a stochastic gradient descent method that is based on adaptive estimation of first-order and second-order moments with an added method to decay weights.
[3]CLS: this token utilized for BERT stands for classification
[4]https://docs.python.org/3/library/difflib.html

complexity for the worst case. In this way, for instance, the prediction of the opinion "super practical to slip into my bag" can be considered correct even if the real opinion is "practical to slip into my bag".

3.1. Results

The performance of the models was evaluated using *precision*, *recall*, and *F1-Score*. *Precision* is the ratio of relevant instances retrieved to all instances retrieved. *Recall* is the ratio of relevant instances retrieved to all relevant instances. *F1-Score* is the harmonic mean of precision and recall. The results are shown in Table 3. A quadruple prediction is deemed correct only if it matches the gold standard in all four components, except for the last tool in the table.

Tool	Precision	Recall	F1-score
Paraphrase modeling (T5-base)	0.373	0.382	0.377
Extract Classify-ACOS (BERT)	0.384	0.205	0.268
PyABSA (T5-base)	0.323	0.310	0.316
PyABSA (T5-large)	0.414	0.409	0.411
PyABSA (T5-large with similarity)	**0.538**	**0.526**	**0.528**

Table 3
Results of the experiments on the ACOS task with different tools

Among the tools with base pre-trained models (T5-base and BERT), the Paraphrase modeling tools seem to be the overall best, but the support for the implicit opinion, lacking from this tool, could be important for some application domains. The Extract Classify-ACOS tool seems to be slightly better than Paraphrase modeling in terms of precision but has a significantly lower value for recall. The last tool we considered, PyABSA, is not the best in terms of performance but it turned out to be very well designed, allowing us to customize it for performing further experiments using a larger pre-trained model (T5-Large) and employing a similarity criterion for one of the components. By using the larger model the precision increased from about 32% to 41% using the standard correctness criterion, and to 54% using the relaxed correctness criterion based on similarity.

4. Conclusion and future work

We benchmarked three ACOS systems from existing literature on a new domain using our custom dataset.

The research aims to create a unified framework for executing various ABSA tasks using different tools on the same dataset. Adapters would handle data translation into the correct format. The framework should allow defining various experiments and exploring different scenarios through automatic and controlled selection of test and train data, based on data categories and polarities.

We foresee an integrated framework where these tools' predictions are used to automatically or semi-automatically enhance and expand the training data, thereby improving the efficiency and overall quality of the sentiment analysis models.

References

[1] W. Zhang, X. Li, Y. Deng, L. Bing, W. Lam, A survey on aspect-based sentiment analysis: Tasks, methods, and challenges, CoRR abs/2203.01054 (2023). doi:10.48550/arXiv.2203.01054.

[2] M. P. et al., Semeval-2014 task 4: Aspect based sentiment analysis, in: P. Nakov, T. Zesch (Eds.), Proc. 8th International Workshop on Semantic Evaluation, SemEval@COLING 2014, Dublin, Ireland, August 23-24, 2014, The Association for Computer Linguistics, 2014, pp. 27–35. doi:10.3115/v1/s14-2004.

[3] M. P. et al., Semeval-2016 task 5: Aspect based sentiment analysis, in: S. B. et al. (Ed.), Proc. 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016, The Association for Computer Linguistics, 2016, pp. 19–30. doi:10.18653/v1/s16-1002.

[4] M. M. Trusca, F. Frasincar, Survey on aspect detection for aspect-based sentiment analysis, Artificial Intelligence Review 56 (2023) 3797–3846. doi:10.1007/s10462-022-10252-y.

[5] G. Brauwers, F. Frasincar, A survey on aspect-based sentiment classification, ACM Computing Surveys 55 (2023) 65:1–65:37. doi:10.1145/3503044.

[6] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of Machine Learning Research 21 (2020) 140:1–140:67.

[7] S. V. et al., Instruction tuning for few-shot aspect-based sentiment analysis, in: J. Barnes, O. D. Clercq, R. Klinger (Eds.), Proc. 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, WASSA@ACL 2023, Toronto, Canada, July 14, 2023, Association for Computational Linguistics, 2023, pp. 19–27.

[8] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis MN, USA, June 2-7, 2019, Vol 1, Association for Computational Linguistics, 2019, pp. 4171–4186.

[9] W. Z. et al., Aspect sentiment quad prediction as paraphrase generation, in: M. M. et al. (Ed.), Proc. of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Punta Cana, Dominican Republic, 7-11 November, 2021, Association for Computational Linguistics, 2021, pp. 9209–9219. doi:10.18653/v1/2021.emnlp-main.726.

[10] H. Cai, R. Xia, J. Yu, Aspect-category-opinion-sentiment quadruple extraction with implicit aspects and opinions, in: C. Z. et al. (Ed.), Proc. 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, Vol 1, August 1-6, 2021, Association for Computational Linguistics, 2021, pp. 340–350. doi:10.18653/v1/2021.acl-long.29.

[11] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: Y. Bengio, Y. LeCun (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.

[12] H. Yang, K. Li, A modularized framework for reproducible aspect-based sentiment analysis, CoRR abs/2208.01368 (2022). doi:10.48550/arXiv.2208.01368.