This laptop has great coffee: Training a Dutch ABSA model from customer reviews

Fieke Middelraad and Lorenzo Gatti (*supervisor*)

University of Twente, Enschede 7500AE, The Netherlands

1 Introduction

One of the challenges of aspect-based sentiment analysis (ABSA) is the need for large amounts of annotated data. The data scarcity is even more prevalent for languages other than English, including Dutch. In my thesis [6], I aim at bypassing this issue by using customer reviews from the webshop Bol.com, and considering the "plus" and "minus" points provided by customers as aspects with a sentiment label. This research explores two methods: using a multi-task model and performing the tasks of aspect extraction and classification separately. The predictions are evaluated using both an automatic and a human evaluation.

Previous research in ABSA for Dutch [1, 2] uses pipeline approaches with multiple NLP components. Similarly to [5], this research is instead based on a multi-task GPT-2 [7] (and, in particular, a version adapted for Dutch [10]), but uses fortuitous data for fine-tuning. This approach is compared to a GPT-2 and BERT approach (specifically, BERTje [9]); BERT has also been used for ABSA, albeit in different setups [4,8] and, to the best of my knowledge, never for Dutch.

2 Method

Data. The dataset consists of 298,109 customer reviews about electrical appliances, retrieved from Bol.com. Each review contains a text and a list of positive and negative aspects ('pros' and 'cons'). For example, consider the following review: *This laptop is super fast, but it overheats quickly.*, with associated the pros *good quality*, and the cons *expensive* and *overheats quickly.* To create a dataset containing only aspects that are explicitly mentioned in the review (e.g., removing *good quality*), I use the pre-trained fastText Dutch embeddings [3]. I calculate the cosine similarity between the average vector of each aspect and the vector of each word in the review text. If the similarity is above the empirically found threshold of 0.83, I assume the aspect is mentioned in the review. Finally, I split the remaining 29,848 reviews into a training, validation and testing set.

Method. Method one, Multi-task GPT-2, uses a pre-trained Dutch GPT-2 model [10] for both aspect extraction and classification simultaneously. I fine-tune the model to output pros and cons when prompted with a review text.

For the second method, GPT-2 and BERT, the aspect extraction is almost identical to the first method. In this case, I fine-tune GPT-2 to output aspects

Table 1. Automatic and human evaluation results

	GPT-2 and BERT		Multi-task GPT-2	
	automatic	human	automatic	human
Aspect extraction (accuracy)	0.547	**0.632**	0.485	0.551
Sentiment classification (F1)	**0.987**	0.945	0.979	0.943

only, rather than specified pros and cons. Next, I fine-tune BERTje [9] to output a sentiment label for each aspect and corresponding review.

Evaluation. The automatic evaluation uses word embeddings and their cosine similarity to compare the model predictions to the plus and minus points given by customers. The human evaluation consists of an online survey which asks the respondents whether an aspect is positive, negative or not mentioned, given a review. The survey contains 100 reviews. Each review is annotated by taking the majority vote of at least three and on average six responses.

3 Discussion and Conclusion

The results are presented in Table 1. For the aspect extraction task, the GPT-2 and BERT method appears to have the best performance. The automatic evaluation is too strict in declaring aspects irrelevant, since both methods score higher on the human evaluation.

For the aspect sentiment classification task, both methods score similarly high. GPT-2 and BERT slightly outperforms multi-tasked GPT-2, although both methods achieve F1 scores very close to 1, meaning they labelled all sentiments almost perfectly. However, that does bring into question the validity of the evaluation. Since I can only evaluate generated aspects that are mentioned in the review according to the extraction evaluation, I can only evaluate a relatively small part of the testing data. This is the part that was mentioned explicitly in the review, which could mean it is also the easier part to find the sentiment of.

Overall, the GPT-2 and BERT method outperforms the multi-task GPT-2 method. This could be because the former uses two separate models for the two separate tasks, while the multi-task GPT-2 model has to divide its capabilities over two tasks. Hence, it makes sense that it performs a little less well.

In conclusion, both methods show that customer review data can be used to fine-tune a language model for an aspect-based sentiment analysis task. The combination of GPT-2 and BERT outperforms multi-task GPT-2. The performance on the aspect extraction task can be improved upon in future work.

References

1. De Clercq, O., Hoste, V.: Rude waiter but mouthwatering pastries! An exploratory study into Dutch Aspect-Based Sentiment Analysis. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). pp.

2910–2917. European Language Resources Association (ELRA), Portorož, Slovenia (5 2016), https://aclanthology.org/L16-1465

2. De Clercq, O., Lefever, E., Jacobs, G., Carpels, T., Hoste, V.: Towards an integrated pipeline for aspect-based sentiment analysis in various domains. In: Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. pp. 136–142. Association for Computational Linguistics, Copenhagen, Denmark (9 2017). https://doi.org/10.18653/v1/W17-5218, https://aclanthology.org/W17-5218
3. Grave, E., Bojanowski, P., Gupta, P., Joulin, A., Mikolov, T.: Learning Word Vectors for 157 Languages (2018). https://doi.org/10.48550/ARXIV.1802.06893, https://arxiv.org/abs/1802.06893
4. Hoang, M., Bihorac, O.A., Rouces, J.: Aspect-Based Sentiment Analysis using BERT. In: Proceedings of the 22nd Nordic Conference on Computational Linguistics. pp. 187–196. Linköping University Electronic Press, Turku, Finland (9 2019), https://aclanthology.org/W19-6120
5. Hosseini-Asl, E., Liu, W., Xiong, C.: A Generative Language Model for Few-shot Aspect-Based Sentiment Analysis (2022), https://arxiv.org/abs/2204.05356
6. Middelraad, F.: This laptop has great coffee: Training a Dutch ABSA model from customer reviews (2022), https://purl.utwente.nl/essays/91778
7. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language Models are Unsupervised Multitask Learners (2019), http://www.persagen.com/files/misc/radford2019language.pdf
8. Sun, C., Huang, L., Qiu, X.: Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence. CoRR **abs/1903.09588** (2019), http://arxiv.org/abs/1903.09588
9. de Vries, W., van Cranenburgh, A., Bisazza, A., Caselli, T., van Noord, G., Nissim, M.: BERTje: A Dutch BERT Model. CoRR **abs/1912.09582** (2019), http://arxiv.org/abs/1912.09582
10. de Vries, W., Nissim, M.: As Good as New. How to Successfully Recycle English GPT-2 to Make Models for Other Languages. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics (2021). https://doi.org/10.18653/v1/2021.findings-acl.74