Aspect-Based Sentiment Analysis using Auxiliary Questions: A Simplified Approach

Natural Language Processing assignment for the course Intelligent Systems at Universidad Politécnica de Madrid

Sabine Warringa

January 26, 2022

Contents

1 Motivation and background

A well-researched topic in the domain of Natural Language Processing is *sentiment analysis*. Essentially, this means assigning a sentiment score or category (usually *negative*, *neutral*, or *positive*) to a piece of writing. This has many use cases in different domains. A concept that has been less thoroughly researched, but is just as useful, is *Aspect-Based Sentiment Analysis* (ABSA). This type of sentiment analysis involves assigning a sentiment score/label to a piece of writing, for each category (or, topic, or aspect) independently. In this project, the categories are predefined and do not need to be extracted from the texts. The most predominant use case for this kind of sentiment analysis is transforming customer reviews into actionable insights. Several tools exist that provide for instance dashboarding that serves this purpose, for example the services offered by Chattermill, which was the inspiration behind this project. The goal of this project is twofold:
T1: Provide a simplified approach for Aspect Based Sentiment Analysis with predefined categories
T2: Provide a prototype for a tool that can be used by businesses to get insights on customer reviews

The remainder of this report is structured as follows. First, the methodology and data used for task T1 is introduced in Section 2. Then, the results are discussed in Section 3. Section 4 introduces the prototype tool as in task T2. Finally, Section 5 gives some limitations and suggestions for improvement.

2 Data and Methodology

In ABSA, the categories can either be predefined by the researcher, or as well need to be extracted from the text. In this project, solely the former is used. Essentially, this then boils down to Subtask 3 (Aspect Category Detection) and Subtask 4 (Aspect Category polarity) in SemEval-2014 Task 4 (Pontiki et al., 2014). Many well-performing methods have been proposed, which are mostly out of the scope of this assignment. In this project, a simplified version of the approach by Sun et al. (2019) is implemented. On a high level, Sun et al. (2019) propose to convert the categories into auxiliary sentences (*What do you think about the food?*). These sentences are then fed to a Question and Answering model. Then, sentiment analysis is performed on the answer (*The food is great!* $\rightarrow$ *positive*). Such a two-phase approach is also used for the models in this paper, as graphically described in Figure 1.

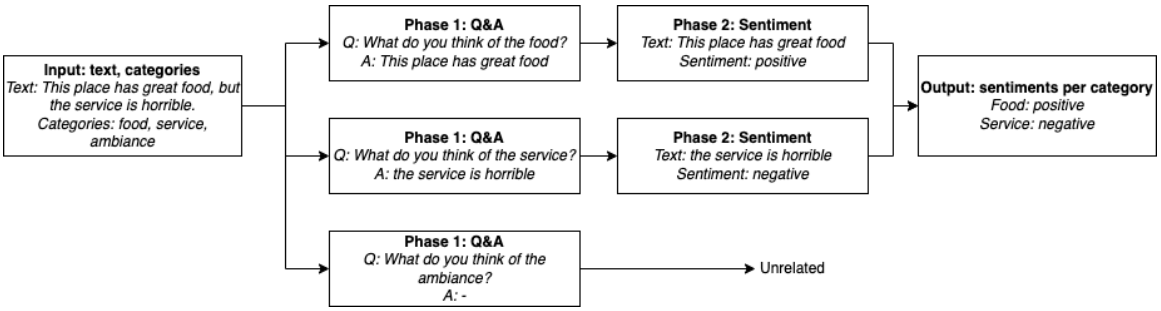


Figure 1: High level overview of model

Phase 1: Q&A uses BERT for Question and Answering (Devlin et al., 2018), as these have shown promising results in the past (Sun et al., 2019). Four different pretrained BERT models have been explored: both a cased and uncased BERT large model and both a cased and uncased DistilBERT model. These models are implemented in Python using the transformers library, Tensorflow, and

HuggingFace (Wolf et al., 2019). Only if the probability score that is assigned by transformers library is greater than 0.01, the answer is accepted. Otherwise, it is assumed that the category is unrelated. This is for instance the case for category *ambiance* in Figure 1.

Phase 2: Sentiment uses the TextBlob library in Python (Loria, 2018). TextBlob provides a polarity score between -1 and 1. This score is then mapped to the categories *negative* if $score \in [-1, -0.01]$, *neutral* if $score \in (-0.01, 0.01]$, and *positive* if $score \in (0.01, 1]$.

The approach as in Figure 1 has been applied to three different ABSA datasets, for each of the four (distil)BERT models:

RESTAURANT (Pontiki et al., 2014) consists of restaurant reviews in English, along with the polarity of several categories: *service* (38.8% negative/3.6% neutral/57.6% positive), *food* (17.9/7.7/74.4%), *price* (37.8/3.3/58.9%), *ambience* (25.5/6.0/68.5%), and *anecdotes/ miscellaneous*. Entries with the last category are deleted, as well as entries with the sentiment *conflict*, as these are not recognisable by TextBlob. The dataset then consists of 1918 unique reviews with in total 2416 review-category combinations. On average each review is hence assigned 1.3 categories.

LAPTOP (Pontiki et al., 2014) consists of laptop reviews in English, along with 1042 unique aspects and their associated sentiments. The categories considered in this project are the ten most frequent aspects in the dataset, where *battery life* is replaced with *battery*: *battery* (48.4/7.5/44.1%), *use* (5.7/1.9/92.5%), *features* (12.5/18.8/68.8%), *screen* (41.1/8.9/50.0%), *software* (50.0/21.9/28.1%), *price* (13.2/3.8/83.0%), *warranty* (41.9/41.9/16.1%), *keyboard* (36.6/19.5/43.9%), *hard drive* (53.3/36.7/10.0%), and *programs* (22.2/25.0/52.8%). Entries with other aspects are deleted, as well as entries with sentiment *conflict*. The dataset now consists of 408 reviews with in total 457 review-category combinations. On average each review is hence assigned 1.1 categories.

MAMS dataset was introduced by Jiang et al. (2019) as they observed that most ABSA research had been conducted using datasets containing reviews with a low number of categories assigned, like for instance RESTAURANT and LAPTOP. The MultiAspect Multi-Sentiment (MAMS) dataset contains restaurant review in English, along with the sentiment per category. The categories included are: *food* (11.1/56.3/32.7%), *place* (20.0/62.0/18.0%), *staff* (66.7/9.3/24.0%), *service* (52.1/20.3/27.6%), *price* (35.4/42.2/22.4%), *menu* (8.2/78.3/13.5%), and *ambience* (27.8/16.4/55.9%). The dataset consists of 3149 unique reviews with in total 6136 review-category combinations. On average each review is hence assigned 1.9 categories, which is significantly more than in the RESTAURANT and LAPTOP datasets.

The implementation of these models can be found on `https://github.com/sabinewar/absa_auxsentence`.

3 Results and Conclusion

The results of applying all four (distil)BERT models to the three datasets as described in Section 2 are summarized in Table 1 and Figure 2. Note that *Found* is the fraction of review-category pairs for which the model in Phase 1 found a relative answer. The reported *F1* score is the macro-F1. The reported F1 and Accuracy metrics are calculated only over the entries for which an answer was found in Phase 1. For example, the accuracy and F1 are calculated over the 75.6% of the review-category combinations for which destilBERT cased found an answer in the review.

Table 1: Percentage of categories found in Phase 1 QA, accuracy and macro-F1 scores

	RESTAURANT			**LAPTOP**			**MAMS**		
Model	*Found*	*Accuracy*	*F1*	*Found*	*Accuracy*	*F1*	*Found*	*Accuracy*	*F1*
distilBERT cased	75.6%	35.2%	0.303	61.3%	30.0%	0.272	34.0%	41.7%	0.347
distilBERT uncased	66.4%	40.8%	0.344	48.8%	40.4%	0.376	23.9%	42.3%	0.376
BERT cased	100%	**61.3%**	**0.462**	100%	**55.6%**	0.501	100%	**46.9%**	**0.459**
BERT uncased	100%	60.5%	0.456	100%	**55.6%**	**0.505**	100%	46.9%	0.453

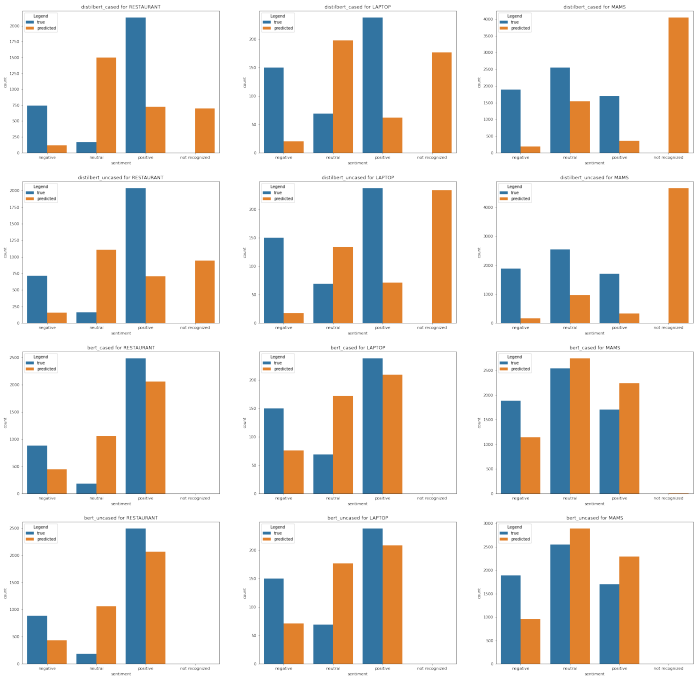


Figure 2: Results of different models on different datasets

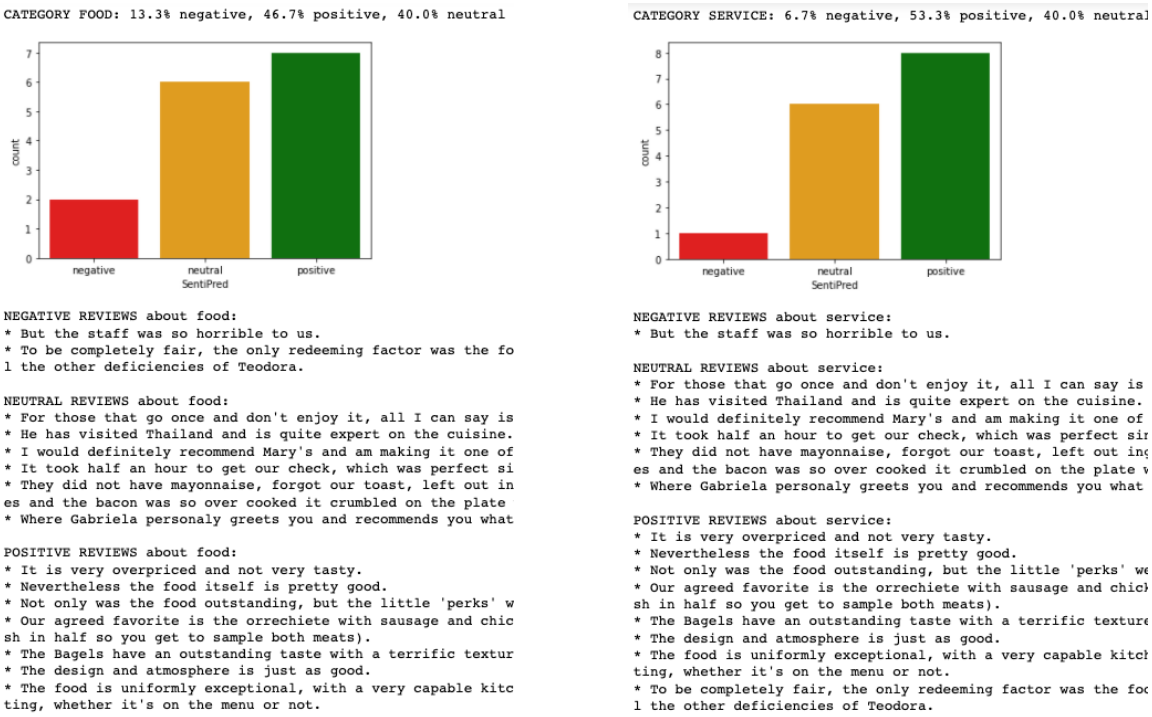
The runtime for varies from 2 minutes (LAPTOP, destilBERT cased) to 117 minutes (MAMS, BERT uncased) on a MacOS 11.1 machine (Intel Core i5 @ 1.8GHz, 8GB RAM).

In general, all four models underperform compared to Sun et al. (2019), but still manage to capture some variance. All macro-F1 scores are low, meaning that the model performs well on common classes but no much on the less represented classes. Figure 2 shows that in almost all the cases the category *neutral* is overpredicted, while *negative* is underpredicted. This might be an indication that TextBlob in general overestimates the polarity, or that the mapping from polarity score to sentiment category is incorrect. The BERT models in general outperform the destilBERT

models in terms of all reported metrics. This is expected as none of the models are finetuned (trained further) for these specific tasks, which is strongly recommended especially for destilBERT models (Wolf et al., 2019). The BERT cased model generally performs best. Hence, this model will be used in the prototype tool related to Task T2.

4 Tool prototype

Task T2 of this project aims at providing a prototype for a tool that can be used by businesses to get insights on customer reviews. As mentioned in Section 3, this tool will use the BERT cased model. In its current version, users upload a .csv file containing customer reviews and a set of categories. Then, ABSA is performed as described in Section 2. The user receives a summary of the results. The tool is illustrated in Figure 3.



(a) Category *food* (b) Category *service*

Figure 3: Illustration of tool prototype

5 Limitations and suggestions

As outlined in Section 2, the simplified approach presented in this project underperforms heavily. This is not necessarily unexpected, as it is a very simplified approach. Moreover, the (destil)BERT and TextBlob models chosen are pretrained, but not finetuned specific for this task and domain. However, the choice of hyperparameters heavily impacts the predictions, for example the interval bounds for converting polarity to sentiment category in Phase 2 and the choice of threshold for an answer in Phase 1 to be accepted. Hence, a suggestion for further work would be to train the (destil)BERT and TextBlob models and perform hyperparameter tuning.

References

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jiang, Q., Chen, L., Xu, R., Ao, X., and Yang, M. (2019). A challenge dataset and effective models for aspect-based sentiment analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6280–6285.

Loria, S. (2018). TextBlob documentation. *Release 0.15*, 2:269.

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., and Manandhar, S. (2014). SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Sun, C., Huang, L., and Qiu, X. (2019). Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence. *arXiv preprint arXiv:1903.09588*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.