Deep Learning Brasil at ABSAPT 2022: Portuguese Transformer Ensemble Approaches

Juliana Resplande Sant'Anna Gomes[1,*,†], Eduardo Augusto Santos Garcia[1,†], Adalberto Ferreira Barbosa Junior[1], Ruan Chaves Rodrigues[2], Diogo Fernandes Costa Silva[1], Dyonnatan Ferreira Maia[1], Nádia Félix Felipe da Silva[1], Arlindo Rodrigues Galvão Filho[1] and Anderson da Silva Soares[1]

[1] *Institute of Informatics, Federal University of Goiás, Brazil*

[2] *Faculty of Informatics, University of the Basque Country (UPV/EHU), Spain*

Abstract

Aspect-based Sentiment Analysis (ABSA) is a task whose objective is to classify the individual sentiment polarity of all entities, called aspects, in a sentence. The task is composed of two subtasks: Aspect Term Extraction (ATE), identify all aspect terms in a sentence; and Sentiment Orientation Extraction (SOE), given a sentence and its aspect terms, the task is to determine the sentiment polarity of each aspect term (positive, negative or neutral). This article presents we present our participation in Aspect-Based Sentiment Analysis in Portuguese (ABSAPT) 2022 at IberLEF 2022. We submitted the best performing systems, achieving new state-of-the-art results on both subtasks.

Keywords

Aspect-Based Sentiment Analysis, Transformer architecture, Portuguese Natural Language Processing

1. Introduction

Aspect-based Sentiment Analysis (ABSA) can be defined as a fine-grained approach to sentiment analysis. In this task, instead of attempting to classify an entire sentence under a single sentiment polarity, we must classify the individual sentiment polarity of all tokens that make a significant contribution to the overall sentiment of the sentence. In the task terminology, these tokens are called the aspect terms of a sentence.

Inspired on the format of SemEval-2014 Task 4 [1], the Aspect-Based Sentiment Analysis in Portuguese (ABSAPT) 2022 at IberLEF 2022 features two subtasks. The first subtask is called Aspect Term Extraction (ATE): given a set of sentences, the task is to identify all aspect terms present in each sentence. In this article, we present our participation at the Aspect-Based

IberLEF 2022, September 2022, A Coruña, Spain.

[*] Corresponding author.

[†] These authors contributed equally.

juliana.resplande@discente.ufg.br (J. R. S. Gomes); edusantosgarcia@discente.ufg.br (E. A. S. Garcia); adalbertojunior@discente.ufg.br (A. F. B. Junior); ruanchaves93@gmail.com (R. C. Rodrigues); diogo_fernandes@discente.ufg.br (D. F. C. Silva); dyonnatan@discente.ufg.br (D. F. Maia); nadia@inf.ufg.br (N. F. F. d. Silva); arlindo@inf.ufg.br (A. R. G. Filho); anderson@inf.ufg.br (A. d. S. Soares)

0000-0001-6900-1931 (J. R. S. Gomes); 0000-0002-0549-8771 (E. A. S. Garcia)

Sentiment Analysis in Portuguese (ABSAPT) 2022 at IberLEF 2022. We submitted the best performing systems, achieving new state-of-the-art results on all subtask.

The second subtask is called Sentiment Orientation Extraction (SOE): given a set of sentences that have already been annotated for their aspect terms, the task is to determine the sentiment polarity of each aspect term (positive, negative or neutral). This subtask is also known as Aspect Term Polarity Classification (ATP) [1, 2] or Aspect Sentiment Analysis (ASA) [3].

The remainder of this article is structured as follows: In the next section, **Related Work**, we go into the previous research that supported our approach; Next, under **Dataset**, we present a detailed analysis of the training data provided for both subtasks. Moreover, under **Methodology**, **Experimental setup** and **Results**, we articulate our strategy to address each subtask; Finally, under **Conclusion**, we bring an overview of the results and a proposal for future work.

2. Related Work

ATE ABSITA [4] was the EVALITA 2020 [5] shared task on Aspect Term Extraction and Aspect-Based Sentiment Analysis. Both the first-ranked team [6] and the second-ranked team [7] had the approach of simply framing Aspect Term Extraction as a Named Entity Recognition task, and then fine-tuning state-of-the-art Transformer models on the training data for the task. We followed a similar approach during our participation at ABSAPT 2022.

ATE ABSITA [4] also features a subtask similar to Sentiment Orientation Extraction (SOE) at ABSAPT 2022. The first-ranked team [6] framed it as a problem of text classification, under the premise that the portion of the text that surrounds each aspect should have the same overall sentiment as the aspect itself.

Although we also experimented with this approach, our best performing system at the Sentiment Orientation Extraction subtask framed it as a text generation problem, similar to what was done by Zhang et al. [8] and Chebolu et al. [9].

3. Dataset

The dataset was taken from TripAdvisor reviews, specifically from the hospitality industry, consisting of hotel and room reviews. Table 1 is an example of a positive polarity from the train dataset. The structure of the data is as follows: id, review, polarity, aspect, start position and end position.

Table 1
Sentiment Orientation Extraction example. Start and End positions are abbreviated into start pos. and end pos.

id	review	polarity	aspect	start pos.	end pos.
2414	Hospedei-me em maio nesse hotel pela terceira vez ...	1	hotel	26	31

For the Aspect Term Extraction task, we build a NER dataset converting the original train data to the BIO/IOB format (Inside, Outside, Beginning), a common tagging format where each token can be classified with the prefixes *B*, *I* and *O*. The *B* prefix indicates the begging of a new classification chunk, the *I* prefix indicates that the token is inside a previous chunk and the *O* tag indicates that the token doesn't belong to any class or chunk. An example of the annotation is shown in Table 2.

Table 2
Example of an Aspect Term Extraction sentence converted to the BIO tagging format.

A	estrutura	do	hotel	é	muito	boa.	A	piscina	é	excelente	e	os	quartos	também.
O	O	O	B-ASPECT	O	O	O	O	B-ASPECT	O	O	O	O	B-ASPECT	O

Analyzing the provided dataset in search of imbalanced data that could be exploited, we found some noteworthy cases. Polarity, for example, is unbalanced towards 1 (positive), representing about 68% of the dataset. Furthermore, the distributed polarity with respect to the aspects is also unbalanced, as seen in Figure 1.

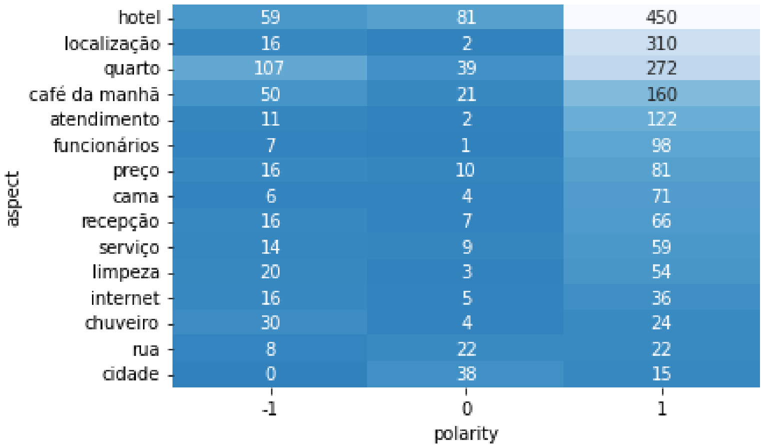

Figure 1: Aspect polarity heat map from top 15 ocorrences.

Another interesting point is the aspect term occurrences. Although the dataset contains 77 unique aspects, the top 15 represent 79% of the data, as shown in Figure 2. Furthermore, the aspect term position is not evenly distributed across the dataset, as its occurrence is more common at the beginning of a review.

Concerning the Aspect Term Extraction, we found that the distribution of aspects in the reviews has a mean of 3.7 and a standard deviation of 1.7, As seen in Figure 3. In contrast, the distribution of words has a mean of 68 and a standard distribution of 21. The low amount of aspects per review can provide an extra challenge for a model to learn accurately, since most of the words in a given review will not be an aspect.

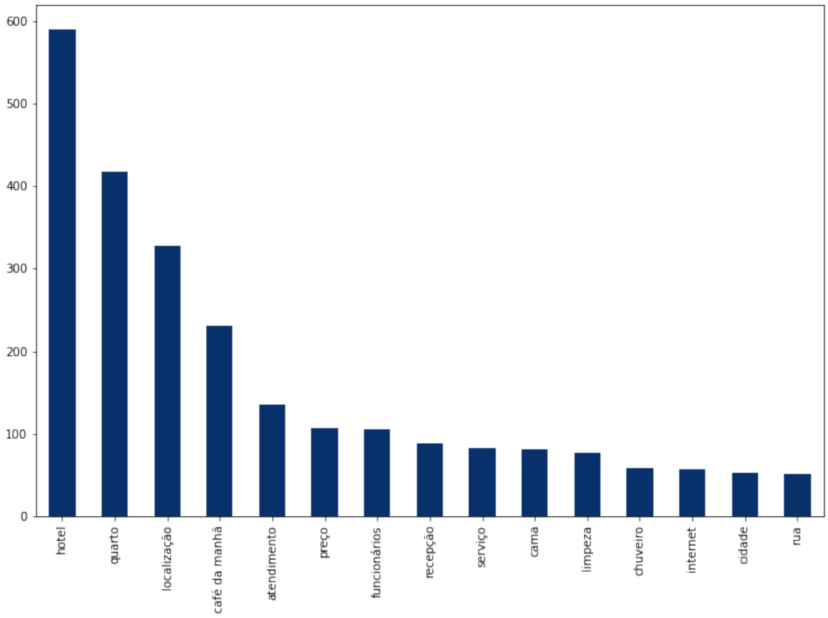

Figure 2: The number of occurrences of the 15 most common aspect occurrences in the training dataset.

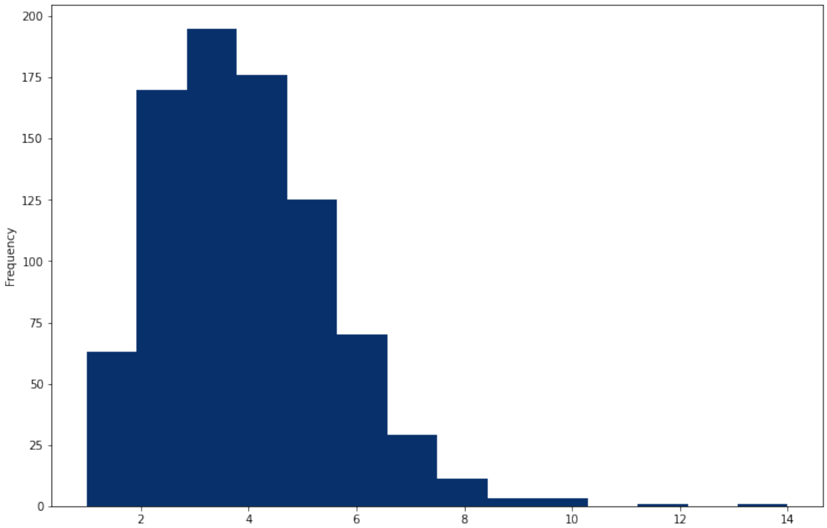

Figure 3: The frequency distribution of the total number of aspects for each unique sentence on the training dataset.

4. Methodology

We adopt different methodologies in the ATE and SOE tasks. For ATE, we treat it primarily as a single sentence tagging task [10]. For SOE, we test two distinct strategies: treating SOE as sentence pair classification or as conditional text generation.

4.1. Task 1 - Aspect Term Extraction

For Aspect Term Extraction, we assess four training considerations: training strategy, Transformer model, dataset preprocessing and dataset configuration, in which we explain in the next subsections.

4.1.1. Training strategy

For the objective of Aspect Term Extraction, we transform the dataset to behave as a Named Entity Recognition task (NER) [6], where we classify chunks of the sentence with the objective to identify which group of tokens represents an aspect. The reviews are tokenized and each aspect is tagged with the BIO format, we train a transformer model using the Hugging Face transformers library [11].

4.1.2. Transformer model

We consider evaluating the following Tranformer-based models with Portuguese support: Bertimbau base ; mDeBERTa v3 base; and our own RoBERTa base in Portuguese trained on BrWaC and the Portuguese portion of OSCAR corpora. Also an mDeBERTA pretrained model on Evalita, MAMs and Semeval datasets, which can be considered as mDEBERTA trained previously on additional data.

4.1.3. Dataset preprocessing

The original competition dataset is separated by aspect, for the conversion to a NER dataset we join all the aspects of reviews with the same text and each unique review it's used as a single training example. We use the NTLK library [12] to tokenize and generate the BIO annotated dataset for a model's training.

4.1.4. Dataset configuration

In addition to the dataset in Portuguese of the ABSAPT 2022 we used external ABSA datasets such as Evalita, MAMs, Semeval 2014, 2015, and 2016 tasks.

4.2. Task 2 - Sentiment Orientation Extraction

For Sentiment Orientation Extraction, we assess four training considerations: training strategy, review preprocessing, Transformer model and dataset configuration, in which we explain in the next subsections.

4.2.1. Training strategy

SOE can be targeted with masked language models or autoregressive language models [8, 9].

If SOE is targeted with a masked language model, following Devin et. al. [10], the review and the aspect term are concatenated with a sentence separator token ([SEP]), and then the resulting sequence of tokens is assigned a sentiment polarity.

However, if SOE is targeted with an autoregressive language model, it will generate sentiment polarity labels starting from a prompt in the format of "Review: [review content] Aspect: [aspect term] Polarity:", which should be completed with either "positive", "negative" or "neutral". This is done either in a zero-shot fashion, in the case of GPT-3, or after fine-tuning a sequence-to-sequence Transformer model to examples given in this format, such as PTT5 [13].

4.2.2. Review preprocessing

We tested two review preprocessing setups for the SOE task. In the first one, the entire review is concatenated with the aspect term, just as described under our training strategy. In the second one, only the portion of the review that is relevant to the aspect term is concatenated with it. Following Rosa et al. [6], we implement this by concatenating with the aspect term only the sentence in which the aspect term is located.

4.2.3. Transformer model

In order to determine the Transformer model to be used, we consider Tranformer models with Portuguese support: Bertimbau base; Bertimbau large; PPT5 base; PTT5 large; mDeBERTa base; XLM-RoBERTa base; LaBSE, Canine-c, RemBERT.

4.2.4. Dataset configuration

We test whether to use external data using B2W dataset and ABSA datasets, such as Evalita, MAMs, Semeval 2014, 2015, and 2016 tasks, by extending pseudo-training subsets on 5.2.

We additionally tested mDeBERTA pretrained model on MNLI and XNLI provided by `MoritzLaurer/mDeBERTa-v3-base-mnli-xnli`, which can be considered as mDEBERTA trained previously on additional data: MNLI and XNLI.

Moreover, we attempt to perform data augmentation over the original task dataset throuh target swap, according to what has been suggested by Liesting et al. [14]. This approach consists of artificially increasing the amount of sentences on the dataset by replacing aspect terms by others belonging to the same category. As there are no annotated categories to the reviews on the ABSAPT 2022 dataset, we tried to approximate such categories through topic modeling.

5. Experimental setup

In experiments, two V100 GPUs (32 GB) were employed, one for each task.

5.1. Task 1 - Aspect Term Extraction

To evaluate the quality of the models we created a pseudo test-training splits of 30/70 of training set where we make sure that each split has unique sentences, this avoids leakage of data between the pseudo-train and pseudo-test subsets. Including the external data, we create two subsets for the training of the ATE task:

- **Portuguese subset**: 70% random split of the original training-set of the competition dataset.
- **multilingual subset**: All external data from Evalita, MAMs, Semeval, plus the Portuguese subset.

All models are fine-tuned in using the HuggingFace Transformer library with a batch-size of 8, learning-rate of 3e-5 (BERTimbau, RoBERTa) or 4e-5 (mDeBERTa) for 8 epochs. The evaluation occurred in the 30% random split of the original training-set of the competition dataset.

5.2. Task 2 - Sentiment Orientation Extraction

We evaluate the approaches mentioned under subsection 4.2 by splitting the training set for the shared task into a new training set and validation set for evaluation purposes. We experiment with three different ways of splitting it:

- **subset 1**: Random approach. The training set for the shared task is arbitrarily split into a new training and validation set.
- **subset 2**: We are careful to keep the same proportion of reviews of each polarity on each split.
- **subset 3**: Besides polarity, we also try to keep the same ratio of aspect terms between the splits.

On all subsets, we attribute 70% of the reviews to the new training set, and 30% of the reviews to the validation set.

As shown in section 6, the best results for the SOE task in our experiments were achieved by PTT5 Large with the conditional text generation training strategy, while taking the entire review and the aspect term as an input to the model, and without using any external data.

We take this model and fine-tune it under four distinct combinations of learning rate and random seed: $\{3e{-}4, 7\}$, $\{1e{-}4, 5\}$ and $\{5e{-}5, 8\}$. Afterward, we produce the final submission through a majority voting ensemble of the predictions of the four fine-tuned models.

Ensemble training code is available on https://github.com/ju-resplande/dlb_absapt2022.

6. Results

6.1. Task 1 - Aspect Term Extraction

We evaluate the training strategies in 4.1, in terms of the following metrics: accuracy (acc.), precision-macro (precision), recall-macro (recall) and f1-macro (f1) for the pseudo-test split of the competition dataset.

The results of the internal evaluation in the pseudo-test splits of each combination of model and dataset tested can be found in Table 3.

The final submission was created using an ensemble of the 3 best models, using a simple median of the label probabilities output by the models for each token in a review, the competition results for the ATE task can be found in table 4.

Table 3
Best results on the ATE task. The symbol {MAMs, Evalita, Semeval}* refers to mDeBERTa previously trained on external datasets, explained in 4.1.2.

model	external data	acc.	precision	recall	f1
BERTimbau base	-	98.2	78.1	87.8	82.6
RoBERTa PT base	-	98.4	80.8	90.7	85.5
mDeBERTa base	MAMs, Evalita, Semeval	98.4	79.1	**94.0**	**85.9**
mDeBERTa base	{MAMs, Evalita, Semeval}*	**98.5**	**81.4**	90.1	85.5

Table 4
Competition final results for the Task 1 (ATE).

team_name	acc
TeamDeepLearningBrasil	**67.1448**
Teampiln	65.4974
TeamUFSCAR	59.3715
TeamPeAm	33.8243
TeamMachadoPardo	22.1050
TeamUFPR	17.1908
TeamOwl	2.6265

6.2. Task 2 - Sentiment Orientation Extraction

We evaluate the training strategies in 4.2, according to subsets 5.2, in terms of the following metrics: accuracy (acc.), f1-macro (f1), and f1 on each class; positive - f1(pos), neutral - f1(neu), and negative - f1(neg).

Table 5 illustrates 3 best results for each training subset described under subsection 5.2, and also has zero-shot results for GPT-3 [15], which has not been fine-tuned on any of our data. The best results for each specific metric are highlighted in bold.

6.2.1. GPT-3

It is important to mention that we did not translate the reviews into English before turning them into prompts for GPT-3 through the OpenAI's text completion API endpoint [1]. The prompts were designed as described on subsection 4.2. Furthermore, due to time constraints, we did not perform predictions for the entire test set, but only for a subset of 391 reviews.

6.2.2. Review preprocessing

For all the best models tested, we achieved the best results by taking the entire review and the aspect term as an input, instead of trying to extract from the review only the sentences which are relevant to the aspect term.

[1] https://beta.openai.com/docs/api-reference/completions

Table 5
Best results on training subsets. We also include zero-shot results obtained with GPT-3 [15] . The symbol {MNLI, XNLI}* refers to mDeBERTa previously trained on MNLI and XNLI, explained in 4.2.4.

subset	model	external data	acc.	f1	f1(pos)	f1(neu)	f1(neg)
subset 1	PTT5 large	-	**86.8**	**78.8**	**92.6**	62	**81.7**
	PTT5 large	target swap	86.3	78.3	92.4	63	79.6
	PTT5 base	target swap	86.3	78.2	92.6	**63.1**	78.8
	GPT-3	-	80.0	66.0	90.0	36.0	73.0
subset 2	mDeBERTa base	{MNLI, XNLI}*	**85.1**	75.9	**92.4**	**57.1**	78.2
	mDeBERTa base	MAMs, Evalita, Semeval	84.6	**76.6**	91.6	55.2	**83**
	mDeBERTa base	-	83.3	73.5	91.2	52.1	77.1
subset 3	PTT5 large	-	**77.4**	**75.6**	**82.3**	60.6	**83.8**
	mDeBERTa base	MAMs, Evalita, Semeval	76.9	**75.6**	82	**62**	82.8
	mDeBERTa base	{MNLI, XNLI}*	74.2	72.1	81.6	58.1	76.7

6.2.3. PTT5 Large

On subset 1 and subset 3, PTT5 large without external data provides the best results, and mDeBERTa base surpasses PTT5 large only on subset 2. Therefore, in our final submission, we opted for a conditional text generation approach where we build a voting ensemble of PTT5 large models, as described under subsection 5.2.

Table 6 shows the final results of the competition for the SOE task and the metrics of our final submission in comparison with other teams.

Table 6
Competition final results for the Task 2 (SOE).

team_name	bacc	f1	precison	recall
TeamDeepLearningBrasil	**82.3756**	**81.7988**	**81.3144**	**82.3756**
Teampiln	78.8619	77.4794	76.5911	78.8619
TeamUFSCAR	62.8992	61.2248	65.5697	62.8992
TeamPeAm	62.8992	61.2248	65.5697	62.8992
TeamUFPR	62.8992	61.2248	65.5697	62.8992
TeamOwl	53.5995	57.2803	68.9396	53.5996

7. Conclusion

We submitted the best performing system on ABSAPT 2022 at IberLEF 2022, achieving new state-of-the-art results on both ATE and SOE tasks. For ATE, we used an ensemble of RoBERTa and mDeBERTa's models trained in Portuguese and multilingual datasets, respectively. For SOE, we employed a voting ensemble of PTT5 large without external data.

In future work, we plan to experiment with a coreference resolution step to improve aspect term extraction by removing ambiguity from the reviews [16]. It may also be interesting to consider fine-tuning Transformer models for both subtasks simultaneously through a multi-task learning framework [17].

Acknowledgments

This work has been supported by the AI Center of Excellence (Centro de Excelência em Inteligência Artificial – CEIA) of the Institute of Informatics at the Federal University of Goiás (INF-UFG).

References

[1] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, SemEval-2014 task 4: Aspect based sentiment analysis, in: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Association for Computational Linguistics, Dublin, Ireland, 2014, pp. 27–35. URL: https://aclanthology.org/S14-2004. doi:10.3115/v1/S14-2004.

[2] Z. Toh, W. Wang, Dlirec: Aspect term extraction and term polarity classification system., in: SemEval@ COLING, 2014, pp. 235–240.

[3] A. Nazir, Y. Rao, L. Wu, L. Sun, Issues and challenges of aspect-based sentiment analysis: A comprehensive survey, IEEE Transactions on Affective Computing 13 (2022) 845–863. doi:10.1109/TAFFC.2020.2970399.

[4] L. De Mattei, G. De Martino, A. Iovine, A. Miaschi, M. Polignano, G. Rambelli, Ate absita@ evalita2020: Overview of the aspect term extraction and aspect-based sentiment analysis task, 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020), (2020).

[5] V. Basile, M. Di Maro, D. Croce, L. Passaro, Evalita 2020: Overview of the 7th evaluation campaign of natural language processing and speech tools for italian, in: 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop, EVALITA 2020, volume 2765, CEUR-ws, 2020.

[6] E. Di Rosa, A. Durante, App2check@ ate absita 2020: Aspect term extraction and aspect-based sentiment analysis, Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020), Online. CEUR. org (2020).

[7] M. Bennici, ghostwriter19@ ate_absita: Zero-shot and onnx to speed up bert on sentiment analysis tasks at evalita 2020, EVALITA Evaluation of NLP and Speech Tools for Italian-December 17th, 2020 (2020) 80.

[8] W. Zhang, X. Li, Y. Deng, L. Bing, W. Lam, Towards generative aspect-based sentiment analysis, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), 2021, pp. 504–510.

[9] S. U. S. Chebolu, F. Dernoncourt, N. Lipka, T. Solorio, Exploring conditional text generation for aspect-based sentiment analysis, arXiv preprint arXiv:2110.02334 (2021).

[10] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional

transformers for language understanding, 2018. URL: https://arxiv.org/abs/1810.04805. doi:10.48550/ARXIV.1810.04805.

[11] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Brew, Huggingface's transformers: State-of-the-art natural language processing, CoRR abs/1910.03771 (2019). URL: http://arxiv.org/abs/1910.03771. arXiv:1910.03771.

[12] E. Loper, S. Bird, Nltk: The natural language toolkit, arXiv preprint cs/0205028 (2002).

[13] D. Carmo, M. Piau, I. Campiotti, R. Nogueira, R. Lotufo, Ptt5: Pretraining and validating the t5 model on brazilian portuguese data, 2020. URL: https://arxiv.org/abs/2008.09144. doi:10.48550/ARXIV.2008.09144.

[14] T. Liesting, F. Frasincar, M. M. Trusca, Data augmentation in a hybrid approach for aspect-based sentiment analysis, 2021. URL: https://arxiv.org/abs/2103.15912. doi:10.48550/ARXIV.2103.15912.

[15] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, 2020. URL: https://arxiv.org/abs/2005.14165. doi:10.48550/ARXIV.2005.14165.

[16] G. S. Chauhan, Y. K. Meena, D. Gopalani, R. Nahta, A mixed unsupervised method for aspect extraction using bert, Multimedia Tools and Applications (2022) 1–26.

[17] M. Schmitt, S. Steinheber, K. Schreiber, B. Roth, Joint aspect and polarity classification for aspect-based sentiment analysis with end-to-end neural networks, 2018. URL: https://arxiv.org/abs/1808.09238. doi:10.48550/ARXIV.1808.09238.