Polish-ASTE: Aspect-Sentiment Triplet Extraction Datasets for Polish

Marta Lango*, **Borys Naglik**[2], **Mateusz Lango**[1,3], **Iwo Naglik**[1,4]
[1] Poznan University of Technology, Faculty of Computing and Telecommunications
[2] Adam Mickiewicz University, Faculty of Mathematics and Computer Science
[3]Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
[4] deepsense.ai, (research@deepsense.ai) * independent researcher

Abstract

Aspect-Sentiment Triplet Extraction (ASTE) is one of the most challenging and complex tasks in sentiment analysis. It concerns the construction of triplets that contain an aspect, its associated sentiment polarity, and an opinion phrase that serves as a rationale for the assigned polarity. Despite the growing popularity of the task and the many machine learning methods being proposed to address it, the number of datasets for ASTE is very limited. In particular, no dataset is available for any of the Slavic languages. In this paper, we present two new datasets for ASTE containing customer opinions about hotels and purchased products expressed in Polish. We also perform experiments with two ASTE techniques combined with two large language models for Polish to investigate their performance and the difficulty of the assembled datasets. The new datasets are available under a permissive licence and have the same file format as the English datasets, facilitating their use in future research.

Keywords: aspect-sentiment triplet extraction, sentiment analysis, language resources, Polish

1. Introduction

Aspect Sentiment Triplet Extraction (ASTE) is a recently proposed sentiment analysis (SA) task (Peng et al., 2020) that involves the extraction of triplets comprising:

- aspect phrase - a text span that represents a particular feature or attribute of the item, for which an opinion is being expressed
- sentiment polarity - often categorised as positive, negative or neutral and refers to the emotional tone being expressed regarding the given aspect.
- opinion phrase - a text span that explicitly conveys the sentiment towards the aspect.

Two examples of sentences with extracted ASTE triplets are given below and in Figure 1.

> Cena jest rozsądna, ale obsługa słaba.
> *The price is reasonable, but the service is poor.*
> Triplets: (Cena$_{\text{price}}$, rozsądna$_{\text{reasonable}}$, Positive), (obsługa$_{\text{service}}$, słaba$_{\text{poor}}$, Negative)
>
> Pokój był czysty i bardzo przytulny.
> *The room was clean and very cosy.*
> Triplets: (Pokój$_{\text{room}}$, czysty$_{\text{clean}}$, Positive), (Pokój$_{\text{room}}$, bardzo przytulny$_{\text{very cosy}}$, Positive)

Note that multi-word phrases and one-to-many links are possible.

Since the ASTE triples provide the comprehensive ”What, How and Why” information regarding the sentiment (Peng et al., 2020), the task quickly gained significant research attention. Many machine learning techniques have been proposed, including GTS (Wu et al., 2020), JET (Xu et al., 2020), Span-based approach (Xu et al., 2021), PBF (Li et al., 2021), GAS (Zhang et al., 2021), a two-stage approach (Huang et al., 2021), BMRC (Liu et al., 2022), SBC (Chen et al., 2022) and EPISA (Naglik and Lango, 2023). However, the experimental evaluation of these models is limited to English only, since datasets for other languages are not available.

In this paper, we introduce two novel datasets for ASTE that contain customer opinions on hotels and purchased products, all expressed in Polish. Furthermore, we conduct experiments using two ASTE techniques in conjunction with two state-of-the-art large language models designed for Polish to assess the difficulty posed by the newly curated datasets to already existing ASTE methods.

2. Related works

There are several datasets constructed for sentiment analysis in Polish. PolEmo 2.0 (Kocoń et al., 2019a) contains over 57k sentences with annotated sentiment polarity, later extended to AspectEmo (Kocoń et al., 2021) with sentiment assigned to each token of the sentence. Even more fine-grained annotation can be found in the Polish Sentiment Treebank[1], used in the PolEval-2017 challenge (Wawer and Ogrodniczuk, 2017), which is a dependency treebank with sentiment annotations for each subphrase of the sentence.

[1]http://zil.ipipan.waw.pl/TreebankWydzwieku

LREC-COLING 2024, pages 12821–12828

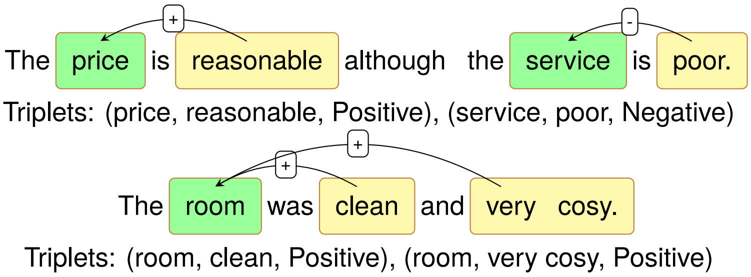

Figure 1: ASTE triplets extracted from two example sentences. Opinion and aspect phrases are highlighted in yellow and green, respectively. The +/- sign denote positive/negative sentiment.

Short informal texts are at the focus of Twitter-Emo corpus (Bogdanowicz et al., 2023) which provides annotation for detecting sarcasm, sentiment and basic emotions. Ptaszynski et al. (2019) collected dataset for cyberbullying detection. Similarly, HateSpeech corpus[2](Troszyński and Wawer, 2017) comprises online posts with offensive content.

Allegro Reviews dataset (Rybak et al., 2020) contains customer reviews with 1-5 ratings from a popular Polish e-commerce marketplace. Finally, Wroclaw Corpus of Consumer Reviews Sentiment (WCCRS, Kocoń et al. 2019b) is a multi-domain dataset of Polish textual reviews, available on a permissive license that allows its adaptation and transformation.

None of the aforementioned Polish datasets is suitable for training machine learning models for ASTE, but naturally such datasets do exist for English. The datasets originate from SemEval shared tasks (Pontiki et al., 2014, 2015, 2016), for which the following corpora were constructed: 14lab (opinions about laptops), 14res, 15res and 16res (restaurant reviews). Fan et al. (2019) extended these corpora by adding opinion term annotation and Peng et al. (2020) finally compiled them for ASTE.

It is important to mention that there are other related tasks like structured SA or fine-grained SA, which have datasets constructed for different languages e.g. for Basque and Catalan (Barnes et al., 2018), German (Klinger and Cimiano, 2014), Czech (Steinberger et al., 2014) or Hindi (Akhtar et al., 2016). Some of these tasks even involve the extraction of some triplets from texts, but they contain opinion holders, aspect categories, or others.

3. New ASTE Datasets for Polish

3.1. Construction of datasets

The customer reviews for the new datasets were taken from the training part of Wroclaw Corpus of Consumer Reviews Sentiment (WCCRS) (Kocoń et al., 2019), which contains online customer reviews from four domains: medicine, hotels, products and students' opinions about lectures. The sentiment polarity in WCCRS is provided at the sentence and document level, but not at the aspect level. Therefore, the sentiment annotation from WCCRS was not used during the construction of our ASTE datasets.

We selected two WCCRS domains: hotels (reviews of hotels, originally taken from TripAdvisor) and products (buyers' opinions on products from Ceneo.pl). Following previous work for English (Fan et al., 2019; Peng et al., 2020), we annotated (aspect phrase, opinion phrase, sentiment polarity) triples in each sentence, where aspect/opinion phrases are text spans and the considered sentiment polarities are: positive, negative and neutral.

The datasets were annotated by Polish native speakers using Doccano annotation platform (Nakayama et al., 2018). Every annotator studied the annotation guidelines prepared by an NLP expert with experience in sentiment analysis. All doubts and questions of the annotators were discussed during the Q&A sessions. Quality control was done by reviewing the annotators' output by NLP experts who were allowed to correct them if needed (less than 5% of examples). To measure the inter-annotator agreement, we asked one of our annotators to re-annotate 50 examples from the hotel datasets that had previously been annotated by another annotator. The annotator provided the same annotation for 78% of the sentences.

A basic summary of the annotation guidelines is provided below.

- Aspects should be properties of the whole item/service being described in a review. For instance, in a hotel review the aspects might be "room" or "parking lot" but not "kettle".

- Opinion phrases are the shortest possible phrases that provide a sentiment polarity for an aspect, without changing the intensity of the sentiment polarity. For example, in

[2] http://zil.ipipan.waw.pl/HateSpeech

the text “very good room” the phrase “very good” is an opinion phrase because a shorter phrase “good” would lose information about the sentiment intensity.

- If the review contains an opinion phrase but no aspect, it should not be annotated.
- If the opinion can be assigned to the phrase that denotes the whole described item, it should be treated as an aspect. However, if it is possible to associate the opinion phrase with a more precise aspect, the name of the whole item should be ignored. For instance: ”Everything is good” will not be annotated as there is no aspect phrase, ”The hotel is good” should yield the (hotel, good, positive) triplet, but ”The hotel has good rooms” should be annotated as (rooms, good, positive).
- The aspect phrases should not contain prepositions, even if they determine the noun declension e.g. ”W pokoju wszystko było okej” (In the room everything was good) will be annotated as (pokoju/room, okej/good, positive)
- While assigning sentiment polarity, the context and the intention of the opinion giver should be taken into account.

While annotating WCCRS corpus, we noticed that some of the texts did not appear to be customer reviews, but rather press releases discussing, for example, hotel construction or the tourism industry. Such sentences were flagged by our annotators and excluded from further analysis in this paper. However, we release them along with our annotations for possible future experiments with out-of-domain texts.

Our datasets are available under open source licence CC BY-SA 4.0 Deed[3] on our GitHub repository[4] with the file format being compatible with English datasets constructed for ASTE (Wu et al., 2020). This will allow easier integration with already existing code bases and hopefully will incentivise comparison of ASTE models on the newly proposed benchmark for Polish.

3.2. Characteristics of new datasets

The qualitative analysis of the constructed datasets can be found in Table 1. For comparison, the same statistics were computed for four English datasets from SemEval competitions. Furthermore, to facilitate comparisons, we computed a weighted average of each statistic value over all datasets for each language, using the size of the corpora (measured by the number of sentences) as the average weight.

The average number of triples in a sentence is slightly higher in the Polish datasets, mostly due to more frequent opinion phrases forming triples with the same aspect. An important feature of the proposed datasets that makes them more challenging are multi-word opinion phrases, which are almost six times more frequent than in the English datasets. The average length of an opinion phrase is 2.97 words in the hotels dataset and 2.22 words in the products' dataset, while this value does not exceed 1.25 in any of the English corpora. Similarly, the number of triples containing solely single word spans in both aspect and opinion phrases is about 1.75 times lower in the Polish datasets. Such triples are much easier to construct by machine learning methods, making the Polish datasets more challenging for them.

The distribution of sentiment polarity in the Polish datasets is significantly more balanced between positive and negative classes than in the English counterparts. However, in datasets for both languages, the triplets with Neutral sentiment are quite rare. Note, that the issue of class imbalance is common in sentiment classification (Lango, 2019).

4. Experimental evaluation

4.1. Experimental setup

The aim of experimental evaluation is to assess the difficulty of the newly constructed datasets for Polish. We selected two deep learning approaches for the comparison:

- Grid Tagging Scheme (GTS, Wu et al., 2020) is a classical approach for ASTE which builds a classifier predicting a particular word-by-word matrix. Due to the special coding scheme used, the matrix can be converted to aspect-sentiment triplets by a dedicated decoding algorithm.
- Exploiting Phrase Interrelations Span-level Approach (EPISA, Naglik and Lango, 2023) is a recent technique that first generates all suitable phrases from a given text, and then constructs final triples with a 2-dimensional CRF model that exploits interrelations between the phrases. To the best of our knowledge, this is the state-of-the-art approach for ASTE for English.

To construct text representation, the aforementioned models leverage masked language models (MLM) like BERT (Devlin et al., 2019) or DeBERTa (He et al., 2020) which are pre-trained on

[3] https://creativecommons.org/licenses/by-sa/4.0/
[4] https://anonymous.4open.science/r/Polish-ASTE-Datasets-anonymous/

	English				Polish		Norm. avg.	
	14lap	14res	15res	16res	hot.	prod.	Eng.	Pol.
Number of sentences	1453	2068	1075	1393	590	511	n/a	n/a
Number of triplets	2349	3909	1747	2247	1197	851	1.69	1.85
incl. with negative sentiment	774	754	401	483	541	376	0.40	0.83
incl. with neutral sentiment	225	286	61	90	58	54	0.10	0.10
incl. with positive sentiment	1350	2869	1285	1674	598	421	1.18	0.92
Number of aspect phrases	2030	3392	1507	1946	798	693	1.46	1.35
incl. single word aspect phrases	1292	2545	1102	1427	681	526	1.04	1.09
incl. multi-word aspect phrases	738	847	405	519	117	167	0.42	0.26
Number of opinion phrases	2030	3409	1620	2078	1156	827	1.51	1.79
incl. single word opinion phrases	1705	3037	1421	1829	412	343	1.32	0.68
incl. multi-word opinion phrases	325	372	199	249	744	484	0.19	1.10
Number of one-to-many relation	535	812	307	388	323	150	0.33	0.42
incl. one aspect-to-many opinions	281	443	208	263	289	128	0.20	0.37
incl. one opinion-to-many aspects	254	369	99	125	34	22	0.13	0.05
N. of triplets w/ single words spans	1305	2631	1140	1478	385	287	1.07	0.61
N. of triplets w/ a multi-word phrases	1044	1278	607	769	812	564	0.61	1.24
incl. with multi-word opinion and single word aspect	207	302	149	188	649	377	0.14	0.92
incl. with multi-word aspect and single-word opinion	684	875	403	513	46	70	0.41	0.11
Mean sentence length (words)	18.4	16.9	15.0	14.9	16.4	21.0	16.43	18.50
Mean length of aspect phrases	1.47	1.40	1.45	1.44	1.26	1.40	1.44	1.32
Mean length of opinion phrases	1.25	1.16	1.19	1.19	2.97	2.22	1.20	2.62

Table 1: Quantitative characteristics of the proposed corpora for Polish (hotel, products) and, for comparison. of already existing corpora for English. Additionally, an average normalized by the size of the corpus (number of sentences) is computed over all corpora for a given language (Norm. avg.).

		GTS			EPISA		
Language	Dataset	Precision	Recall	F1	Precision	Recall	F1
English	14lap	58.54	50.65	54.30	66.98	60.55	63.56
	14res	68.71	67.67	68.17	75.29	72.56	73.89
	15res	60.69	60.54	60.61	66.44	64.74	65.54
	16res	67.39	66.73	67.06	71.12	72.45	71.77
Polish	products (TrelBERT)	45.15	40.17	41.74	50.01	43.36	46.07
	products (HerBERT)	40.65	33.79	35.65	48.18	44.09	45.89
	hotels (TrelBERT)	42.07	37.82	39.08	49.07	41.66	44.72
	hotels (HerBERT)	37.35	28.28	30.02	48.79	42.76	45.47

Table 2: The experimental results of ASTE task, measured with Precision, Recall and F1 for two methods: GTS and EPISA. For Polish datasets two variants of these methods are explored, with different backbone masked language models: TrelBERT and HerBERT.

English texts. Therefore, in our study we reimplemented them replacing the original MLMs with two established architectures for Polish, namely HerBERT (Mroczkowski et al., 2021) and TrelBERT (Szmyd et al., 2023).

Following related works, the models are evaluated using three criteria: precision, recall, and F1-score. An aspect/opinion phrase is deemed accurate only if it perfectly aligns with the gold standard (exact match). A triplet is considered a true positive only if all its parts are correct. All metric values presented were calculated on the respective test set and averaged across ten separate training runs. 60% of dataset examples were used for training set and 20% for validation and test set, respectively. Examples without any annotated triplets were omitted during training.

4.2. Results

The results are presented in Table 2. As expected, the more recent EPISA method produced better results in terms of F1 score for all datasets, with an average improvement of 7.5 percentage points over GTS. For both Polish datasets, the combination of GTS with TrelBERT yielded better results than its HerBERT counterpart. Similarly, the TrelBERT variant achieved the highest F1-score value for EPISA on products corpora. On the other hand, HerBERT performed slightly better on hotels corpora when paired with EPISA.

The proposed Polish datasets proved to be much more challenging for existing methods than the existing English datasets. The F1 score averaged over all English datasets is 63.5% for GTS and 68.7% for EPISA, while the average performance of the generally better performing TrelBERT variants is 40.5% and 45.4% for GTS and EPISA respectively. The difference in triplet extraction performance between Polish and English is approx. 23 percentage points for both methods, which calls for future research in ASTE for under-resourced languages.

5. Summary

In this paper we have presented two datasets for the Aspect Sentiment Triplet Extraction task for the Polish language. The new datasets have the same format as commonly used English datasets, which facilitates the comparison and development of machine learning ASTE models for both languages. The ASTE datasets for Polish are characterised by a higher average number of triplets and a higher frequency of one-to-many relations than their English counterparts. Multi-word opinion phrases are also significantly more common.

The conducted experiments with two ASTE deep learning approaches adopted to Polish by coupling them with two different Polish language models showed that the constructed datasets are challenging for modern machine learning techniques.

6. Frequently Asked Questions

1. The data used for annotation is a subset of an existing data set - why is it being presented as two datasets?

   The research on ASTE is almost exclusively carried out on the English datasets discussed in our paper. These datasets contain texts from only one domain, e.g. opinions about restaurants or laptops. The performance achieved by the current SOTA methods on English is still not fully satisfactory and production-ready to move to a more challenging scenario of multi-domain datasets. As we expected that training ASTE for Polish may be even more challenging, we considered it necessary to produce single-domain datasets as well. However, just as all four English datasets can be merged into one larger dataset, the same is true for the Polish datasets. One can also merge all the datasets in our work to test the multilingual multi-domain scenario.

2. The lower metric values obtained on Polish datasets are interpreted as the datasets being more challenging. Is it possible that the model replication/implementation for Polish has issues?

   Both the GTS and EPISA algorithms have been run by us for Polish and English using original implementation provided by the authors. For EPISA we were able to reproduce the exact results of the original work on English, as the original implementation of EPISA was available with fixed random seed. GTS results differed slightly from the original work on English, but none of the differences were statistically significant. The Polish datasets have the same input format as the English datasets, and the only change with respect to the original methods is to replace an English MLM with a Polish MLM, so it is unlikely that there are any implementation issues for Polish.

3. Is it possible that the methods used were just optimized for English data, and this causes the lower performance on the Polish data?

   Although current SOTA methods based on deep learning do not appear to explicitly model language-dependent features that would favour English over Polish, it is possible that they are inadvertently optimised for English data. For example, published methods tend to perform better on English benchmarks, leading to the promotion of such methods during development, despite the fact that other design choices might work better for other languages. Developing datasets for languages other than English is the first step in investigating this issue.

7. Acknowledgements

The research has been partially supported by 0311/SBAD/0743 PUT University grants. M. Lango was supported by National Science Centre, Poland (Grant No. 2022/47/D/ST6/01770). This work used resources of the LINDAT/CLARIAH-CZ Research Infrastructure (Czech Ministry of Education, Youth, and Sports project No. LM2018101).

8. Bibliographical References

Md Shad Akhtar, Asif Ekbal, and Pushpak Bhattacharyya. 2016. Aspect based sentiment analysis in Hindi: Resource creation and evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2703–2709, Portorož, Slovenia. European Language Resources Association (ELRA).

Jeremy Barnes, Toni Badia, and Patrik Lambert. 2018. MultiBooked: A corpus of Basque and Catalan hotel reviews annotated for aspect-level sentiment classification. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Stanisław Bogdanowicz, Hanna Cwynar, Aleksandra Zwierzchowska, Cezary Klamra, Witold Kieraś, and Łukasz Kobyliński. 2023. Twitter-emo: Annotating emotions and sentiment in polish twitter. In *Computational Science – ICCS 2023*, pages 212–220, Cham. Springer Nature Switzerland.

Yuqi Chen, Keming Chen, Xian Sun, and Zequn Zhang. 2022. Span-level bidirectional cross-attention framework for aspect sentiment triplet extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhifang Fan, Zhen Wu, Xin-Yu Dai, Shujian Huang, and Jiajun Chen. 2019. Target-oriented opinion words extraction with target-fused neural sequence labeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2509–2518, Minneapolis, Minnesota. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Lianzhe Huang, Peiyi Wang, Sujian Li, Tianyu Liu, Xiaodong Zhang, Zhicong Cheng, Dawei Yin, and Houfeng Wang. 2021. First target and opinion then polarity: Enhancing target-opinion correlation for aspect sentiment triplet extraction.

Roman Klinger and Philipp Cimiano. 2014. The USAGE review corpus for fine grained multi lingual opinion analysis. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2211–2218, Reykjavik, Iceland. European Language Resources Association (ELRA).

Jan Kocoń, Piotr Miłkowski, and Monika Zaśko-Zielińska. 2019a. Multi-level sentiment analysis of PolEmo 2.0: Extended corpus of multi-domain consumer reviews. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 980–991, Hong Kong, China. Association for Computational Linguistics.

Jan Kocoń, Jarema Radom, Ewa Kaczmarz-Wawryk, Kamil Wabnic, Ada Zajaczkowska, and Monika Zaśko-Zielińska. 2021. AspectEmo 1.0: Multi-domain corpus of consumer reviews for aspect-based sentiment analysis. CLARIN-PL digital repository.

Jan Kocoń, Monika Zaśko-Zielińska, and Piotr Miłkowski. 2019b. Multi-level analysis and recognition of the text sentiment on the example of consumer opinions. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2019*.

Mateusz Lango. 2019. Tackling the problem of class imbalance in multi-class sentiment classification: An experimental study. *Foundations of Computing and Decision Sciences*, 44(2):151–178.

Yuncong Li, Fang Wang, Wenjun Zhang, Sheng hua Zhong, Cunxiang Yin, and Yancheng He. 2021. A more fine-grained aspect-sentiment-opinion triplet extraction task.

Shu Liu, Kaiwen Li, and Zuhe Li. 2022. A robustly optimized BMRC for aspect sentiment triplet extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 272–278, Seattle, United States. Association for Computational Linguistics.

Iwo Naglik and Mateusz Lango. 2023. Exploiting phrase interrelations in span-level neural approaches for aspect sentiment triplet extraction. In *Advances in Knowledge Discovery and Data Mining*, pages 222–233, Cham. Springer Nature Switzerland.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

Haiyun Peng, Lu Xu, Lidong Bing, Fei Huang, Wei Lu, and Luo Si. 2020. Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8600–8607.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495, Denver, Colorado. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 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.

Michal Ptaszynski, Agata Pieciukiewicz, and Paweł Dybała. 2019. Results of the poleval 2019 shared task 6: First dataset and open shared task for automatic cyberbullying detection in polish twitter. *Proceedings of the PolEval 2019 Workshop*, page 89.

Piotr Rybak, Robert Mroczkowski, Janusz Tracz, and Ireneusz Gawlik. 2020. KLEJ: Comprehensive benchmark for polish language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1191–1201, Online. Association for Computational Linguistics.

Josef Steinberger, Tomáš Brychcín, and Michal Konkol. 2014. Aspect-level sentiment analysis in Czech. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 24–30, Baltimore, Maryland. Association for Computational Linguistics.

Marek Troszyński and Aleksander Wawer. 2017. Czy komputer rozpozna hejtera? wykorzystanie uczenia maszynowego (ml) w jakościowej analizie danych. *Przegląd Socjologii Jakościowej*, 13(2):62–80.

Aleksander Wawer and Maciej Ogrodniczuk. 2017. Results of the poleval 2017 competition: Sentiment analysis shared task. In *8th Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*.

Zhen Wu, Chengcan Ying, Fei Zhao, Zhifang Fan, Xinyu Dai, and Rui Xia. 2020. Grid tagging scheme for aspect-oriented fine-grained opinion extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2576–2585, Online. Association for Computational Linguistics.

Lu Xu, Yew Ken Chia, and Lidong Bing. 2021. Learning span-level interactions for aspect sentiment triplet extraction. In *Proceedings of the 59th Annual Meeting of the ACL and the 11th IJCNLP (Volume 1: Long Papers)*, pages 4755–4766. Association for Computational Linguistics.

Lu Xu, Hao Li, Wei Lu, and Lidong Bing. 2020. Position-aware tagging for aspect sentiment triplet extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2339–2349. Association for Computational Linguistics.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021. Towards generative aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the ACL and the 11th IJCNLP (Volume 2: Short Papers)*, pages 504–510. Association for Computational Linguistics.

9. Language Resource References

Kocoń, Jan and Zaśko-Zielińska, Monika and Miłkowski, Piotr and Janz, Arkadiusz and Piasecki, Maciej. 2019. *Wroclaw Corpus of Consumer Reviews Sentiment (WCCRS)*. [link].

Mroczkowski, Robert and Rybak, Piotr and Wr 'oblewska, Alina and Gawlik, Ireneusz. 2021. *HerBERT: Efficiently Pretrained Transformer-based Language Model for Polish*. Association for Computational Linguistics. [link].

Szmyd, Wojciech and Kotyla, Alicja and Zobniów, Michał and Falkiewicz, Piotr and Bartczuk, Jakub and Zygadło, Artur. 2023. *TrelBERT: A pre-trained encoder for Polish Twitter*. Association for Computational Linguistics. [link].