AspectEmo: Multi-Domain Corpus of Consumer Reviews for Aspect-Based Sentiment Analysis

Jan Kocón∗, Jarema Radom†, Ewa Kaczmarsz-Wawryk†, Kamil Wabnic†, Ada Zajaczkowska∗, Monika Zaśko-Zielinska†

∗Department of Artificial Intelligence, Wrocław University of Science and Technology, Poland
†Institute of Polish Studies, University of Wrocław, Poland
∗{jan.kocon,jarema.radom,ada.zajaczkowska}@pwr.edu.pl
†{ewa.kaczmarsz,kamil.wabnic,monika.zasko-zielinska}@uwr.edu.pl

Abstract—Aspect-based sentiment analysis (ABSA) is a text analysis method that categorizes data by aspects and identifies the sentiment assigned to each aspect. Aspect-based sentiment analysis can be used to analyze customer opinions by associating specific sentiments with different aspects of a product or service. Most of the work in this topic is thoroughly performed for English, but many low-resource languages still lack adequate annotated data to create automatic methods for the ABSA task. In this work, we present annotation guidelines for the ABSA task for Polish and preliminary annotation results in the form of the AspectEmo corpus, containing over 1.5k consumer reviews annotated with over 63k annotations. We present an agreement analysis on the resulting annotated corpus and preliminary results in the form of transformer-based models trained on AspectEmo.

Index Terms—AspectEmo, aspect-based sentiment analysis, ABSA, corpus, Polish

I. INTRODUCTION

Sentiment analysis has been one of the most active fields of research in Natural Language Processing (NLP) [1]–[5]. It has been researched as a response to the growing availability of opinion-rich resources, such as personal blogs and online review sites [6]. Rapid development can also be attributed to high business demand for user feedback. One of such use cases is processing user reviews, for instance, in e-commerce systems. In most cases, sentiment analysis focuses on inferring the sentiment of the entire review or sentence. However, it is not always the optimal choice since one review or sentence might contain information on both positive and negative aspects of a product or service. A user might like the quality of a product or service but loathe the price. The Aspect Based Sentiment Analysis (ABSA) is proposed as a method to extract user attitudes towards specific aspects of a product or service. This approach allows inferring more in-depth sentiment, which could be invaluable for business as it could, e.g., allow a company to focus on refining negative aspects of their products or services while maintaining aspects valued by their users. ABSA focuses on the relationship between three clearly defined elements – a subject, an aspect of the subject, and a sentiment expressed towards subject and aspect.

Aspects are sub-themes to the main theme. They are also called local topics in opposition to the global topic. The issue of thematic organization of discourse includes a higher-level topic of the text, called the discourse theme, which is related to the overall object of speech or the global purpose of the discourse, and lower-level topics called sentence themes. In the discourse analysis, the global theme is combined with the concept of the macrostructure of the text. Macrostructure is the global semantic structure of a discourse, which is usually revealed in the title, heading, or summary sentences. A global theme is indicated for the text under analysis, which is determined by the source of the text. This data takes the form of titles or headings. They are initially used by a viewer to find the information they are looking for or to focus their attention on a specific topic. The conceptual basis of the topic, i.e., what the section is about, can be used to determine aspects, which we treat as sub-topics. The distinction of a particular aspect is thus determined by the conceptual layer of the lexical unit, not by its label. Therefore aspects must be distinguished from their textual determiners, of which there may be many more than distinguished aspects in specific utterances.

Subjects (aspects) in the discourse analysis are most often delimited within sentence boundaries, which cannot always be applied to the analyzed material. For example, Internet texts do not always have sentence boundaries clearly delineated by the creator of the text, as consistent punctuation is omitted. Rigid delimitation of aspect boundaries, that is, understood as limiting the aspect to the sentence boundaries, is therefore irrelevant. If we treat the sub-topic as the center of attention, then with respect to the word-based approach, its center will be the lexical unit, and the other elements will be the context of the word, which influences the evaluation. We consider that the boundary of an aspect is determined by the appearance of a new determinant that is being evaluated. Ultimately, we should conclude that the following elements occur within aspect boundaries: (1) the aspect determiner, which is the subject; (2) the aspect evaluation, or comment (rHEME), which describes what the aspect is. In the close vicinity of the comment, there may be intensifiers, de-intensifiers, or other
entities that modify the introduced evaluation.

Aspect-based sentiment analysis has been researched in multiple domains such as tourism [7], product reviews [8], [9], social reviews [10] and recommender systems [11]. Multiple works attempted to solve ABSA using various methods such as multichannel deep convolutional neural networks [12], lexicon-based approach [13], Spearman’s rank correlation coefficient [14], ontology-based approaches [15], factorized Latent Dirichlet Allocation (LDA) [16], two-layered Long Short-Term Memory model (LSTM) [17], and Bidirectional Encoder Representations from Transformers (BERT) [18]. There have been works focused on sentiment analysis in the Polish language [19]. However, there has not been a single one dedicated to ABSA. ABSA was studied across other low-resource languages, e.g., Roman Urdu [20], Turkish [21] and Vietnamese [22]. Wang [23] performs ABSA task on the Chinese dataset containing information from multiple different domains (cosmetic, book, hotel, movie, cellphone, restaurant) and using multiple models, based on Term Frequency, Conditional Random Fields (CRF), Support Vector Machine, Naive Bayes, Hidden Markov Model, Recurrent Neural Network, and K-Nearest Neighbours. This approach is similar to what this work tries to accomplish. However, a different set of methods is used in our work, which notably outperforms each of the previously enumerated models. Cruz [24] attempts to address the case of implicit aspect extraction with the use of CRF. In this work, we will deal with both implicit and explicit aspects. Afmuus [25] explores the fuzzy rule-based method. Rules were extracted from nouns and noun phrases using the Fuzzy Unordered Rule Induction Algorithm. For the purpose of aspect identification, in this work, we use a predefined set of labels, which were annotated by experts following a 2+1 scheme.

To our best knowledge, nobody has attempted to annotate a corpus and perform ABSA on a low-resource language such as Polish and on a dataset containing reviews from multiple different domains (hotels, products, school, medicine) with the usage of a wide range of BERT-based pre-trained models or LSTM. The main contributions of the paper are (1) development of ABSA annotation guidelines for a multi-domain consumer review corpus, (2) performing large-scale annotation (1.6k texts and 63k annotations in total) by domain experts (linguists) in a 2+1 mode along with a disagreement verification procedure, (3) evaluation of an automated approach to the ABSA task using transformer-based SOTA models.

II. Related Work

Lexical resources for sentiment analysis are primarily WordNets containing emotion annotation, e.g., SentiWordNet [26], WordNet-Affect [27], or pLWordNet 4.0 Emo [28], [29], whose design facilitates the propagation of results using automated methods. It is noted, however, that such resources are often not fully covered by emotion annotations, are often only available for certain languages, prompting the construction of further multilingual corpora [30], and sometimes have a limited range of annotations that includes only positive, negative, or neutral affect [31]. Therefore, the emotionally annotated wordnets and dictionaries are complemented by project-specific corpus resources containing collections of texts in specific languages, organized by genre or domain. As far as research presented in this article is concerned, the key resources are these in the Polish language version. So far, Polish language corpora have been developed, which are the subject of research projects on sentiment analysis. For the Sentimenti project, in addition to annotating selected lexical units from the Polish wordnet (Słowosieć), emotional annotation was also applied to consumer reviews [32]. This corpus (PolEmo) is continuously used to develop methods for machine extraction of sentiment from texts [32], [33]. The corpus established within the framework of the PolEval 2017 project [34], which contains the opinions of users of various types of products (such as perfume, clothes) with added negative, positive, and neutral annotations, is definitely noteworthy. Another interesting corpus related to mining emotions from the text is the HateSpeech corpus consisting of 2000 posts published on Polish websites and annotated for the presence of hate speech. Semantic and pragmatic annotations can be found in the Polish Corpus of Suicide Notes [35] which consists of authentic and simulated suicide letters. The Polish linguistic corpora also include a corpus combining a treebank, Twitter posts, and online film reviews, which is used to predict sentiment in short pieces of text [36].

III. AspectEmo Corpus

The approach we propose to analyze sentiment in consumer reviews takes into account the manual annotation phase, which is relatively time-consuming and cost-intensive. Thus its application is often limited to a smaller group of texts [37] with some attempt to replace it with crowdsourcing or automated methods that should be selected based on a comparison with the results of manual annotation [38]. So far, comparisons of manual annotations for sentiment analysis focused on overtone evaluation rather than taking into account the compatibility of boundaries of the context necessary for aspect annotation. The developed annotation procedures need to cover both the aspect identification stage and the boundary determination stage in which the given element is evaluated by the text author. We assume that the assessed aspect may be represented in the text by different text determinants. It is, therefore, not possible to automatically extract aspect determinants beforehand that could be used by people performing crowd-coding in the semi-manual extraction of aspects [39]. Similarly, it is very difficult to assume in advance what context boundaries should be taken into account in aspect sentiment. Each time, they are determined based on the aspect of a (de)intensifier related to the determinant.

A. Dataset

AspectEmo Corpus is an extended version of a publicly available PolEmo 2.0 corpus [40] that was used in many projects on the use of different methods in sentiment analysis.
The AspectEmo corpus consists of four subcorpora, each containing online customer reviews from the following domains: school, medicine, hotels, and products:

- The *school* subcorpus consists of 1,000 documents (94642 tokens, approximately 95 tokens per text). The opinions were sourced from the polwro.pl discussion forum and are opinions on the courses and lecturers at the Wrocław University of Technology. Students are the authors of the opinions.
- The *medicine* subcorpus consists of 3510 documents (478505 tokens, approximately 136 tokens per text). The texts were sourced from the Znanylekarz.pl portal, which is intended to help patients find a doctor and make appointments. These are opinions on doctors authored by patients.
- The *hotels* subcorpus consists of 4200 documents (578259 tokens, approximately 137 tokens per text). The subcorpus consists of texts from the English version of tripadvisor.com. These are opinions on hotels authored by guests.
- The *products* subcorpus consists of 1000 documents (135217 tokens, approximately 135 tokens per text). The texts come from ceneo.pl. This is a price comparison engine that also allows for adding feedback on products. Thus, text authors should be the users of the assessed products.

IV. Annotation Guidelines

The annotation process was carried out using the Inforex tool [41]. One of the features of Inforex, in which the multi-genre corpus was annotated, is a module for annotation reconciliation called the Agreement module. In this module, two annotators make their annotations independently, and they assign them to a given token. During validation, consistency between the annotators’ decisions taken for the same text fragment is checked together with consistency between the changing members of the annotators’ pair or the decision of the super-annotator. The development of guidelines for annotating sentiment in consumer reviews for the AspectEmo corpus included the development of rules for detecting aspect indicators in the text and the establishment of rules for assigning a characteristic to individual aspects.

A. Aspect identification

For the correctness of the annotation process, it was important to clarify how aspects can be expressed and what are the principles for marking them. We have determined that we annotate every occurrence of a marker in the text to take account of the individual ratings rather than generalizing them into sequences ambiguous in terms of attributes. We marked the markers regardless of the assessment that appears next to them. If the suffix in the text did not carry any value, it obtained a neutral annotation. The above findings allowed for comparing the compatibility between the annotators. According to our guidelines, the aspect mark can be expressed with a:

- **noun**, for example: *guy, nurse, bathroom, access, etc.*, even if it is a predicative in a nominal predicate as we consider the noun to be the first in the hierarchy, for example: *he is a scientist.*
- **pronoun**: personal (*he, she*); possessive (*his, her*); generalising (*all*).
- **verb**, if this verb is in third person, which means it corresponds to a noun, *lecturer, doctor, plate, etc.*, meaning: *asks, examines, acts*, or takes the form of a gerund (*conducting classes, examining a patient*); it is a conjunction of the nominal predicate, *e.g.*, *is*; if the indicator is a compound predicate, for example: *likes to experiment*, then we annotate the first element.
- **adjective**, when this adjective takes a nominal function (*demanding*), *i.e.*, it is not accompanied by another noun (*a demanding woman*), nor is it part of a nominal predicate (*he is how? – he is demanding*); or when it is part of a nominal predicate that lacks conjunction (*is*), and there is nothing to be annotated.
- **adverb**, if this adverb is part of a nominal predicate, and the conjunction is missing from the sentence, *e.g.: close* to the centre – for distance.

If several determinants are identified in the sentence for the aspect under analysis, their hierarchy is as follows: noun, pronoun, verb, adjective. For example, if a noun and pronoun are in a given local context, then we select a noun, and if it is, for example, a verb, pronoun, and adjective, then we select a pronoun.

B. Sentiment in aspect

The guidelines for the sentiment annotation in the current project indicate that evaluation should be formulated based on the context of your local indicator. The aspect annotation includes the following values: strong negative – SN, weak negative – WN, neutral – 0, weak positive – WP, strong positive – SP, ambiguous – AMB. The annotated value was always determined based on information contained in the text and not on the possible default judgment of the recipient. As the basic purpose of the consumer opinion is predominantly assessment, the aspect indicators are mostly polarised (positive or negative). Only exceptionally, in the fragments devoted to advise or hints, which are informative, do they remain neutral.

In the annotation process, we took into account:

- sentiment behind aspect determinant, *e.g.*, *quack doctor* instead of *physician*; *ruin* instead of *building*;
- sentiment-loaded words occurring locally together with accompanying modifiers, intensifiers *s* (*e.g. very – extremely, tragic – tragically bitter*), de-intensifiers (*slightly, sometimes, usually*), or conjunctions affecting sentiment, *e.g.: if, but*.

Thus, if local context includes information recommended, we determine sentiment as *strongly positive* because there is no de-intensifier here, and the intensity-indicating expression itself (*recommended*) has a positive sentiment.
V. Annotation Analysis

The manual annotation of texts from the AspectEmo corpus followed the 2+1 scheme, which means that there were two annotators and one super-annotator. In addition, each subcorpus annotation series was accompanied by meetings devoted to post-super-annotation. They took place twice after having annotated 50 texts, twice another 100 texts, and then after 200 texts. This allowed for making some corrections to annotation guidelines, thus clarifying the annotations and adding specific examples.

At this stage, 1700 texts were annotated: 500 from the school, medicine, hotels subcorpora, and 200 from the products subcorpus. Currently, the number of documents annotated by two annotators is 1582, with the total number of annotations of 42104 and the number of super-annotations of 21270. Table I provides information on the number and volume of annotated texts in subsequent subcorpora and the overall distribution of the annotation. We subsequently annotated resources from the school, medicine, hotels, products subcorpora, which corresponded to an increased annotation complexity level as well as an increase in the number of people evaluated, as more places, services and functions got evaluated. In the products corpus, both material objects, cultural texts, and electronics were reviewed.

The analysis of annotation agreement covered the agreement of the annotation boundaries, i.e., the local context is taken into account by the annotator when assessing a specific sentiment determinant and sentiment agreement (strong negative, weak negative, neutral, weak positive, strong positive, ambiguous), which corresponds to the description of sentiments in lexical units in pWordNet. We observed that improving the agreement of the annotation boundaries has always led to improvements in the sentiment agreement. Table II shows the annotation agreement for specific subcorpora. It has been calculated using the Positive Specific Agreement method (PSA) [42]. The PSA values are listed to indicate annotation boundaries, limits, and values. PSA for the annotation boundaries indicates spots in which annotators marked aspects subject to valuation in the text. PSA for boundaries and tags includes both the annotator’s indication of the aspect and the assessment of its sentiment. The average PSA in the AspectEmo corpus for annotation boundary is 75.62%, and for the boundaries and tags combined, it is 61.83%. The lowest PSA values were found in the products sub-corpus, 68.14% and 57.09%, respectively. This is likely to be due to the wide variety of aspects related to the occurrence of opinions on extremely varied goods in the sub-corpus.

Table III presents the PSA values per tag broken down into subcorpus. The highest PSA values were found for extremes SP (72.23%) and SN (64.23%), followed by neutral ones (55.0%). The lower PSA annotations values have unambiguous annotations: WN - 36.10%, WP - 35.70%, and AMB - 22.62%. This may be linked to the difficulty annotators face when assessing these aspects, which do not have an unambiguous sentiment. The tags that have a lower PSA value are also generally less frequently picked by annotators, as shown in Table IV. Similar lower PSA results for ambiguous tags were obtained in similar studies on the analysis of sentiment at text and sentence levels [40].

VI. Automatic Aspect-Based Sentiment Analysis

The most closely related work in terms of use case proposes a solution for a dataset of faculty teaching performance evaluation is [17]. The proposed solution includes two LSTM networks, first focused on aspect extraction, second on opinion mining. The output of the first LSTM network is being fed directly to the second one. The authors emphasize the problem with sentences where more than one aspect has been addressed without using any connector. Another limitation is using a

<table>
<thead>
<tr>
<th>Domain</th>
<th>Texts</th>
<th>Tokens</th>
<th>Annotations</th>
<th>Super-ann.</th>
</tr>
</thead>
<tbody>
<tr>
<td>school</td>
<td>498</td>
<td>48326</td>
<td>8775</td>
<td>4380</td>
</tr>
<tr>
<td>medicine</td>
<td>386</td>
<td>49935</td>
<td>7206</td>
<td>3667</td>
</tr>
<tr>
<td>hotels</td>
<td>498</td>
<td>72284</td>
<td>19784</td>
<td>10305</td>
</tr>
<tr>
<td>products</td>
<td>200</td>
<td>26911</td>
<td>6274</td>
<td>2718</td>
</tr>
<tr>
<td>all</td>
<td>1582</td>
<td>197456</td>
<td>42039</td>
<td>21270</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Subcorpus</th>
<th>SN</th>
<th>WN</th>
<th>WP</th>
<th>SP</th>
<th>AMB</th>
<th>ALL</th>
</tr>
</thead>
<tbody>
<tr>
<td>school</td>
<td>60.44</td>
<td>33.08</td>
<td>57.95</td>
<td>39.24</td>
<td>66.03</td>
<td>56.39</td>
</tr>
<tr>
<td>medicine</td>
<td>65.31</td>
<td>28.47</td>
<td>57.37</td>
<td>22.45</td>
<td>74.69</td>
<td>63.50</td>
</tr>
<tr>
<td>hotels</td>
<td>65.95</td>
<td>42.14</td>
<td>54.64</td>
<td>37.23</td>
<td>75.60</td>
<td>65.14</td>
</tr>
<tr>
<td>products</td>
<td>58.50</td>
<td>35.85</td>
<td>50.80</td>
<td>26.17</td>
<td>66.42</td>
<td>57.09</td>
</tr>
<tr>
<td>all</td>
<td>64.23</td>
<td>36.10</td>
<td>55.10</td>
<td>33.70</td>
<td>72.23</td>
<td>61.83</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Subcorpus</th>
<th>SN</th>
<th>WN</th>
<th>WP</th>
<th>SP</th>
<th>AMB</th>
<th>ALL</th>
</tr>
</thead>
<tbody>
<tr>
<td>school</td>
<td>2658</td>
<td>274</td>
<td>1356</td>
<td>196</td>
<td>2552</td>
<td>170</td>
</tr>
<tr>
<td>medicine</td>
<td>2658</td>
<td>274</td>
<td>1356</td>
<td>196</td>
<td>2552</td>
<td>170</td>
</tr>
<tr>
<td>hotels</td>
<td>2658</td>
<td>274</td>
<td>1356</td>
<td>196</td>
<td>2552</td>
<td>170</td>
</tr>
<tr>
<td>products</td>
<td>2658</td>
<td>274</td>
<td>1356</td>
<td>196</td>
<td>2552</td>
<td>170</td>
</tr>
<tr>
<td>all</td>
<td>2658</td>
<td>274</td>
<td>1356</td>
<td>196</td>
<td>2552</td>
<td>170</td>
</tr>
</tbody>
</table>

Table I: The Number of Annotated Texts in Each Domain Subcorpus, Their Volume (Expressed in Tokens), the Number of Annotations and Super-Annotations.

Table II: PSA Values (%) for Aspect Boundaries Only (PSAA) and for Aspects and Their Sentiment Tags (PSA&S).

Table III: PSA Aspect-Sentiment Values for Different Sentiment Categories in Subcorpora.

Table IV: Number of Annotations for Different Sentiment Categories in Subcorpora.
predefined list of aspects rather than extracting them from sentences.

One approach to the duality of the ABSA task, which also applies to Aspect-based Opinion Mining (ABOM), is an alternating coattention network [1]. A coattention network consists of two LSTM networks that can learn both attention representations for target and context. The output of both LSTM networks is processed in coattention encoder, which proceeds to the classifier. Such an approach can outperform most of the existing LSTM-attention methods yet still struggles with negation modifiers and implicit sentiment.

Another recent work [43] tries to address problems relevant to deep learning networks, namely tendency of attention mechanism to over-fit and force network to focus on the insignificant part of sentence resulting in loss of key information and lack of flexibility in case of low availability of training data. Authors propose a model merging lexicon information with attention LSTM to emphasize on strengths of both approaches. Results confirm the effectiveness of the approach, stating that fine-grained, domain and aspect-specific lexicon could significantly improve performance.

Bidirectional encoder representations from transformers (BERT) [44] are designed to pre-train deep bidirectional representations from the unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks. BERT is versatile, and it is specific behavior must be interpreted within the scope of the task. Thus [45] have investigated BERT’s behavior on ABSA datasets. Work concludes that BERT uses very few self-attention heads to encode context words and opinion words for an aspect, most features in the representation of an aspect are dedicated to fine-grained semantics of the domain and the aspect itself instead of carrying summarized opinions from its context.

Fine-tuning BERT for numerous natural language processing (NLP) tasks, including ABOM-like tasks, has been widely researched recently. Hoang et al. [18] proposed a model based on fine-tuned BERT, which managed to outperform previous state-of-the-art results for ABSA. [46] introduces fine-tuned BERT and achieves new state-of-the-art results on SentiHood and SemEval2014 Task 4 datasets. State-of-the-art results were achieved the same year with the model introduced in work [47]. The model consisted of a Self-attention network built on top of fine-tuned BERT. Li et al. [48] proposed a new method named GBCN. The fine-grained sentiment analysis task needs auxiliary information to distinguish each aspect. BERT by itself is not capable of providing extra contextual information. To address this problem gating mechanism is introduced. The input text is fed to BERT as well as to the context-aware embedding layer to generate both BERT representation and refined context-aware embeddings separately. Then a gating mechanism is being used to control the propagation of sentiment features from BERT with context-aware embeddings. GBCN achieves new state-of-the-art results on SentiHood and SemEval2014. Janz and Milkowski et al. [49] employed Polish RoBERTa to solve the named entity recognition (NER) task and achieved state-of-the-art results on KPWr corpus. Ou et al. [50] utilized XLM-RoBERTa for Multi-language Sentiment Analysis. It achieved the best F-Score in Malayalam-English and ranked third in Tamil-English.

VII. Experiments

There are two problems at focus. The first aims to achieve state-of-the-art performance on AspectEmo corpus using previously discussed architectures. To properly confront results achieved with each architecture, performances will be measured using Accuracy, Recall, and F1 score. Overall results will be discussed in terms of potential improvement. The second problem focuses on the time and space complexity of evaluated models.

A. Setup

In each case, the dataset was divided into batches of tokens and corresponding labels. Those in the case of LSTM were vectorized with Embeddings from Language Models (ELMo1). In the case of BERT-based approaches, each one has its own pre-trained tokenizer used on word level to prepare inputs for models. In all approaches, each batch of embedded or tokenized text is fed to a model, which processes sequence capturing sentence-wide context.

In the case of BERT-based approaches – Figure 1, there are three additional layers, linear layer on the top of pre-trained model, dropout layer, and linear classification layer. In the case of LSTM – Figure 2, a batch is fed into a linear layer, then to a two-layer bidirectional LSTM cell, and finally to another linear layer used for classification. The same loss function was used in all approaches – PyTorch2 implementation of CrossEntropyLoss, which combines LogSoftmax and negative log-likelihood loss. The same optimizer was used across all approaches, namely Pytorch’s implementation of Adam [51] with a learning rate of 5e-5, adam epsilon of $1 \times 10^{-8}$, and weight decay of 0.01. The dimension of hidden states is set to 300 in the case of LSTM, 768 in the case of mBERT, and 1024 in the rest of the cases. In the first experiment, max_sequence_length is always set to 128. In all cases, dropout is set to 0.3, and batch_size is equal to 32. The evaluation was conducted with CUDA3 on GPU: RTX 3090 24GB RAM and Processor: AMD Ryzen Threadripper 3970X 32-Core Processor.

B. Results

Table V contains $F_1$, Recall, and Accuracy results for all chosen approaches. HerBERT outperforms other models, but XLM-Roberta comes as a close second. mBERT significantly underperforms compared to other BERT-based approaches, but as mentioned previously, all BERT-based models use large versions aside from mBERT, which only offers the base model. Noteworthy, LSTM performs only slightly worse

1https://clarin.pl.eu/dspace/handle/11321/690
2https://pytorch.org/docs
3https://developer.nvidia.com/cuda-zone
Fig. 1. Overview of all BERT-based models

Fig. 2. Overview of implemented LSTM-based model

TABLE V
F1 AND ACCURACY SCORES ACHIEVED BY EVALUATED MODELS.

<table>
<thead>
<tr>
<th>Model</th>
<th>F1</th>
<th>Accuracy</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM</td>
<td>42.63</td>
<td>± 0.60</td>
<td>41.50 ± 2.10</td>
</tr>
<tr>
<td>mBERT</td>
<td>43.65</td>
<td>± 0.72</td>
<td>45.76 ± 1.55</td>
</tr>
<tr>
<td>Polish RoBERTa</td>
<td>63.73</td>
<td>± 0.55</td>
<td>66.61 ± 1.53</td>
</tr>
<tr>
<td>XLM-Roberta</td>
<td>65.13</td>
<td>± 0.43</td>
<td>66.99 ± 0.99</td>
</tr>
<tr>
<td>HerBERT</td>
<td>65.22</td>
<td>± 0.29</td>
<td>67.56 ± 1.01</td>
</tr>
</tbody>
</table>

Fig. 3. The learning curve of XLM-R on portions of the original dataset used as a training set.

Fig. 4. The learning curve of XLM-R for AMB label, parts of the original dataset used as a training set.

TABLE VI
F1 SCORE AND ACCURACY PER LABEL, EVALUATED ON HerBERT.

<table>
<thead>
<tr>
<th>Label</th>
<th>F1</th>
<th>Accuracy</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>SP</td>
<td>77.42</td>
<td>83.47</td>
<td>1015</td>
</tr>
<tr>
<td>SN</td>
<td>64.87</td>
<td>69.42</td>
<td>1039</td>
</tr>
<tr>
<td>0</td>
<td>56.85</td>
<td>55.04</td>
<td>501</td>
</tr>
<tr>
<td>WN</td>
<td>28.85</td>
<td>32.61</td>
<td>67</td>
</tr>
<tr>
<td>WP</td>
<td>28.24</td>
<td>24.00</td>
<td>41</td>
</tr>
<tr>
<td>AMB</td>
<td>15.65</td>
<td>23.08</td>
<td>91</td>
</tr>
</tbody>
</table>
than mBERT-base. Performance of all models is less than satisfactory, to find the reason XLM-R was used to draw a learning curve using F score and gradual increase of the size of training dataset - Figure 3. Learning curve is not converging on the currently available dataset. With the increase of data F scores above 0.7 seem to be within the range of current solutions. Not all labels are supported equally. Table VI contains a breakdown of the best model’s performance (HerBERT) divided into labels. It seems that the performance per label is tied to support of the label within the training set. It reinforces the idea that broadening the dataset with a focus on underrepresented labels might lead to better model performance. Figure 4 shows a learning curve for AMB label. It follows the problem noticed in similar work on PolEmo 2.0 [40] – the ambiguous label is not only underrepresented but also hard to assign for human experts as well as for deep learning models.

VIII. CONCLUSIONS

Although the annotation agreement persisted at a relatively satisfactory level, there were some difficulties. In the case of the noun + apposition (e.g., spa area), where one annotator marks the aspect as area and the other as spa, it proved to be problematic. It was also not an easy task to determine the aspect of construction without a directly expressed object of opinion. However, these constructions were unambiguously characterized and therefore had to be annotated. As always, it was difficult to annotate chaotic, disorganized texts with distorted coherence, especially those in which there were many comparisons and references to other objects (e.g., other models of electronic equipment) than the evaluated product or service.

In this work, we have carried out a successful attempt of solving aspect-based sentiment analysis problems on AspectEmo dataset with numerous approaches, including LSTM and BERT-based solutions. HerBERT turned out to be the best performing model, with XLM-RoBERTa being a close second. Results leave room for improvement. Thus, we have researched and discussed potential ways to improve the results in future works, such as increasing the dataset, focusing on certain labels, and applying our recent personalized (user-based) approach [52]–[54] to the sequence classification task. AspectEmo dataset is publicly available in the CLARIN-PL DSpace repository.

From the perspective of annotating the whole text, it will be important to establish a hierarchy of aspects connected to the ontology developed on the basis of the AspectEmo corpus data. This will make it possible to indicate which aspects have a decisive influence on the evaluation of a product or service and which, according to the reviewers, are of marginal importance.

REFERENCES


