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

Jan Kocoń*, Jarema Radom*, Ewa Kaczmarz-Wawryk†, Kamil Wabnic†,
Ada Zajączkowska*, Monika Zaśko-Zielińska†
*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.kaczmarz,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 using 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

This work was financed by (1) the National Science Centre, Poland, project no. 2019/33/B/HS2/02814; (2) the Polish Ministry of Education and Science, CLARIN-PL; (3) the European Regional Development Fund as a part of the 2014-2020 Smart Growth Operational Programme, CLARIN – Common Language Resources and Technology Infrastructure, project no. POIR.04.02.00-00C002/19; (4) the statutory funds of the Department of Artificial Intelligence, Wrocław University of Science and Technology.

2375-9259/21/$31.00 ©2021 IEEE
DOI 10.1109/ICDMW53433.2021.00027

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); when 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.

TABLE I
THE NUMBER OF ANNOTATED TEXTS IN EACH **DOMAIN** SUBCORPUS, THEIR VOLUME (EXPRESSED IN TOKENS), THE NUMBER OF ANNOTATIONS AND SUPER-ANNOTATIONS.

Domain	Texts	Tokens	Annotations	Super-ann.
school	498	48326	8775	4580
medicine	386	49935	7206	3667
hotels	498	72284	19784	10305
products	200	26911	6274	2718
all	1582	197456	42039	21270

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 plWordNet. 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-subcorpus.

TABLE II
PSA VALUES [%] FOR ASPECT BOUNDARIES ONLY (PSA_A) AND FOR ASPECTS AND THEIR SENTIMENT TAGS ($PSA_{A\&S}$).

Subcorpus	PSA_A	$PSA_{A\&S}$
school	77.84	56.39
medicine	76.49	63.50
hotels	76.69	65.14
products	68.14	57.09
all	75.62	61.83

TABLE III
$PSA_{ASPECT\&SENTIMENT}$ VALUES FOR DIFFERENT SENTIMENT CATEGORIES IN SUBCORPORA.

Subcorpus	SN	WN	0	WP	SP	AMB	ALL
school	60.44	33.08	57.95	39.24	66.03	28.01	56.39
medicine	65.31	28.47	57.37	22.45	74.69	20.00	63.50
hotels	65.95	42.14	54.64	37.23	75.60	21.88	65.14
products	58.50	35.85	50.80	26.17	66.42	16.67	57.09
all	64.23	36.10	55.10	35.70	72.23	22.62	61.83

Table III presents the PSA values per tag broken down into subcorpora. 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 IV
NUBER OF ANNOTATIONS FOR DIFFERENT SENTIMENT CATEGORIES IN SUBCORPORA.

Subcorpus	SN	WN	0	WP	SP	AMB	ALL
school	1456	526	2854	841	2741	357	8775
medicine	2658	274	1356	196	2552	170	7206
hotels	7248	617	3631	419	7421	448	19784
products	1723	212	1571	107	2469	192	6274
all	13085	1629	9412	1563	15183	1167	42039

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 Miłkowski 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 (ELMo[1]). 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 – PyTorch[2] 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 1×10^{-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 CUDA[3] 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

[1] https://clarin-pl.eu/dspace/handle/11321/690
[2] https://pytorch.org/docs
[3] https://developer.nvidia.com/cuda-zone

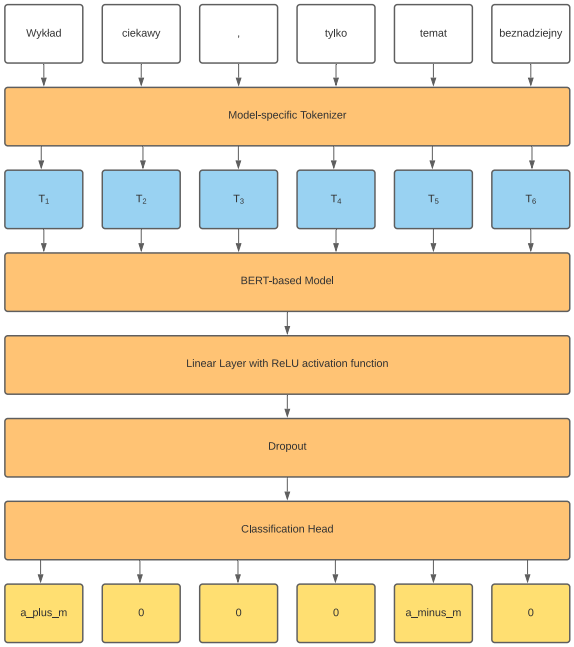

Fig. 1. Overview of all BERT-based models

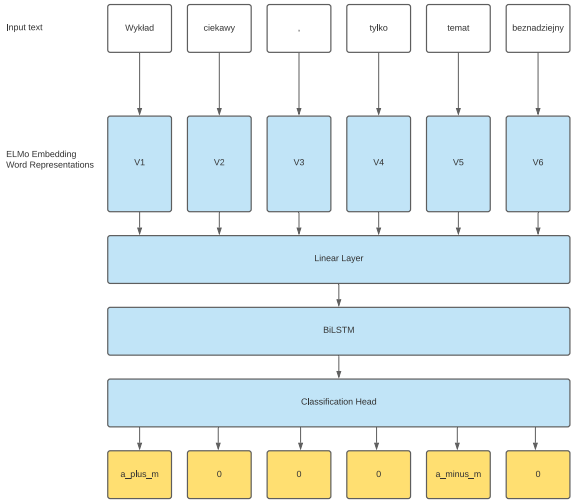

Fig. 2. Overview of implemented LSTM-based model

TABLE V
F_1 AND ACCURACY SCORES ACHIEVED BY EVALUATED MODELS.

Model	F_1	Accuracy	Recall
LSTM	42.63 ± 0.60	41.50 ± 2.10	44.03 ± 2.03
mBERT	43.65 ± 0.72	45.76 ± 1.55	41.79 ± 1.17
Polish RoBERTa	63.73 ± 0.55	66.61 ± 1.53	61.14 ± 1.30
XLM-Roberta	65.13 ± 0.43	66.99 ± 0.99	**63.38 ± 0.75**
HerBERT	**65.22 ± 0.29**	**67.56 ± 1.01**	63.05 ± 0.86

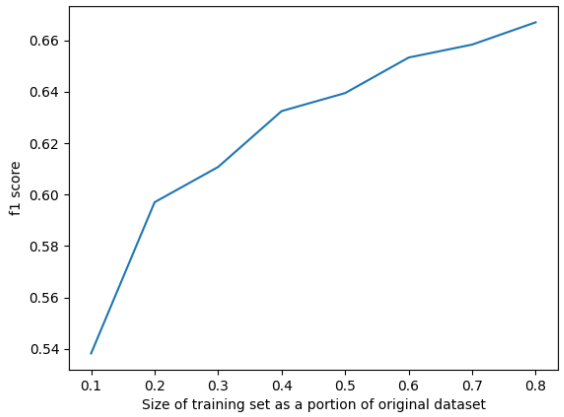

Fig. 3. The learning curve of XLM-R on portions of the original dataset used as a training set.

TABLE VI
F_1 SCORE AND ACCURACY PER LABEL, EVALUATED ON HERBERT.

Label	F_1	Accuracy	Support
SP	77.42	83.47	1015
SN	64.87	69.42	1039
0	56.85	55.04	501
WN	28.85	32.61	67
WP	28.24	24.00	41
AMB	15.65	23.08	91

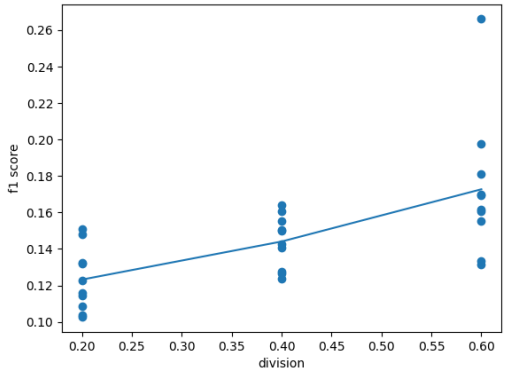

Fig. 4. The learning curve of XLM-R for AMB label, parts of the original dataset used as a training set.

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_1 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_1 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[4].

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

[1] C. Yang, H. Zhang, B. Jiang, and K. Li, "Aspect-based sentiment analysis with alternating coattention networks," *Information Processing & Management*, vol. 56, no. 3, pp. 463 – 478, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0306457318306344

[2] S. Akhtar, D. Ghosal, A. Ekbal, P. Bhattacharyya, and S. Kurohashi, "All-in-one: Emotion, sentiment and intensity prediction using a multi-task ensemble framework," *IEEE transactions on affective computing*, 2019.

[3] M. S. Akhtar, A. Ekbal, and E. Cambria, "How intense are you? predicting intensities of emotions and sentiments using stacked ensemble [application notes]," *IEEE Computational Intelligence Magazine*, vol. 15, no. 1, pp. 64–75, 2020.

[4] M. E. Basiri, S. Nemati, M. Abdar, E. Cambria, and U. R. Acharya, "Abcdm: An attention-based bidirectional cnn-rnn deep model for sentiment analysis," *Future Generation Computer Systems*, vol. 115, pp. 279–294, 2021.

[5] E. Cambria, Y. Li, F. Z. Xing, S. Poria, and K. Kwok, "Senticnet 6: Ensemble application of symbolic and subsymbolic ai for sentiment analysis," in *Proceedings of the 29th ACM international conference on information & knowledge management*, 2020, pp. 105–114.

[6] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Comput. Linguist*, vol. 35, no. 2, pp. 311–312, 2009.

[7] E. Marrese-Taylor, J. D. VelÃˇsquez, and F. Bravo-Marquez, "A novel deterministic approach for aspect-based opinion mining in tourism products reviews," *Expert Systems with Applications*, vol. 41, no. 17, pp. 7764–7775, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417414003315

[8] S. Moghaddam and M. Ester, "On the design of lda models for aspect-based opinion mining," in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, ser. CIKM '12. New York, NY, USA: Association for Computing Machinery, 2012, p. 803–812. [Online]. Available: https://doi.org/10.1145/2396761.2396863

[9] S. Moghaddam and M. Esterr, "Aspect-based opinion mining from product reviews," in *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, 2012, pp. 1184–1184.

[10] J. Mir and M. Usman, "An effective model for aspect based opinion mining for social reviews," in *2015 Tenth International Conference on Digital Information Management (ICDIM)*, 2015, pp. 49–56.

[11] A. Da'u, N. Salim, I. Rabiu, and A. Osman, "Weighted aspect-based opinion mining using deep learning for recommender system," *Expert Systems with Applications*, vol. 140, p. 112871, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417419305810

[12] A. Dau, N. Salim, I. Rabiu, and A. Osman, "Recommendation system exploiting aspect-based opinion mining with deep learning method," *Information Sciences*, vol. 512, pp. 1279–1292, 2020.

[13] F. Wogenstein, J. Drescher, D. Reinel, S. Rill, and J. Scheidt, "Evaluation of an algorithm for aspect-based opinion mining using a lexicon-based approach," in *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, ser. WISDOM '13. New York, NY, USA: Association for Computing Machinery, 2013. [Online]. Available: https://doi.org/10.1145/2502069.2502074

[14] A. K. J and A. S, "Aspect-based opinion ranking framework for product reviews using a spearman's rank correlation coefficient method," *Information Sciences*, vol. 460-461, pp. 23–41, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S002002551830358X

[15] J. Wątróbski, "An ontology-based approach to opinion mining tools selection," *Studies & Proceedings of Polish Association for Knowledge Management*, no. 89, 2018.

[16] S. Moghaddam and M. Ester, "The flda model for aspect-based opinion mining: Addressing the cold start problem," in *Proceedings of the 22nd International Conference on World Wide Web*, ser. WWW '13. New York, NY, USA: Association for Computing Machinery, 2013, p. 909–918. [Online]. Available: https://doi.org/10.1145/2488388.2488467

[17] I. Sindhu, S. Muhammad Daudpota, K. Badar, M. Bakhtyar, J. Baber, and M. Nurunnabi, "Aspect-based opinion mining on student's feedback for faculty teaching performance evaluation," *IEEE Access*, vol. 7, pp. 108 729–108 741, 2019.

[18] M. Hoang, O. A. Bihorac, and J. Rouces, "Aspect-based sentiment analysis using BERT," in *Proceedings of the 22nd Nordic Conference on Computational Linguistics*. Turku, Finland: Linköping University Electronic Press, Sep.–Oct. 2019, pp. 187–196. [Online]. Available: https://www.aclweb.org/anthology/W19-6120

[19] P. Lula and K. Wójcik, "Sentiment analysis of consumer opinions written in Polish," *Economics and Management*, vol. 16, no. 1, pp. 1286–1291, 2011.

[4]https://clarin-pl.eu/dspace/handle/11321/849

[20] R. Zahid, M. O. Idrees, H. Mujtaba, and M. O. Beg, "Roman urdu reviews dataset for aspect based opinion mining," in *2020 35th IEEE/ACM International Conference on Automated Software Engineering Workshops (ASEW)*, 2020, pp. 138–143.

[21] E. Akbaş, "Aspect based opinion mining on turkish tweets," Ph.D. dissertation, bilkent university, 2012.

[22] N. T. T. Thuy, N. X. Bach, and T. M. Phuong, "Leveraging foreign language labeled data for aspect-based opinion mining," in *2020 RIVF International Conference on Computing and Communication Technologies (RIVF)*, 2020, pp. 1–6.

[23] W. Wang, G. Tan, and H. Wang, "Cross-domain comparison of algorithm performance in extracting aspect-based opinions from chinese online reviews," *International Journal of Machine Learning and Cybernetics*, vol. 8, no. 3, pp. 1053–1070, 2017.

[24] I. Cruz, A. F. Gelbukh, and G. Sidorov, "Implicit aspect indicator extraction for aspect based opinion mining." *Int. J. Comput. Linguistics Appl.*, vol. 5, no. 2, pp. 135–152, 2014.

[25] M. Afzaal, M. Usman, A. Fong, S. Fong, and Y. Zhuang, "Fuzzy aspect based opinion classification system for mining tourist reviews," *Advances in Fuzzy Systems*, vol. 2016, pp. 1–14, 10 2016.

[26] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining." in *Lrec*, vol. 10, no. 2010, 2010, pp. 2200–2204.

[27] C. Strapparava, A. Valitutti *et al.*, "Wordnet affect: an affective extension of wordnet." in *Lrec*, vol. 4, no. 1083-1086. Citeseer, 2004, p. 40.

[28] A. Janz, J. Kocon, M. Piasecki, and M. Zasko-Zielinska, "plWordNet as a basis for large emotive lexicons of Polish," *Proceedings of Human Language Technologies as a Challenge for Computer Science and Linguistics Poznan: Fundacja Uniwersytetu im. Adama Mickiewicza w Poznaniu*, pp. 189–193, 2017.

[29] J. Kocoń, A. Janz, and M. Piasecki, "Context-sensitive sentiment propagation in wordnet," in *Proceedings of the 9th Global Wordnet Conference*, 2018, pp. 329–334.

[30] V. Bobicev, V. Maxim, T. Prodan, N. Burciu, and V. Angheluş, "Emotions in words: Developing a multilingual wordnet-affect," in *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, 2010, pp. 375–384.

[31] Y. Rao, X. Quan, L. Wenyin, Q. Li, and M. Chen, "Building word-emotion mapping dictionary for online news," in *SDAD 2012 The 1st International Workshop on Sentiment Discovery from Affective Data*, 2012, p. 28.

[32] J. Kocoń, A. Janz, P. Miłkowski, M. Riegel, M. Wierzba, A. Marchewka, A. Czoska, D. Grimling, B. Konat, K. Juszczyk, K. Klessa, and M. Piasecki, "Recognition of emotions, valence and arousal in large-scale multi-domain text reviews," in *Human Language Technologies as a Challenge for Computer Science and Linguistics*, Z. Vetulani and P. Paroubek, Eds. Poznań, Poland: Wydawnictwo Nauka i Innowacje, 2019, pp. 274–280.

[33] J. Kocoń, P. Miłkowski, and K. Kanclerz, "Multiemo: Multilingual, multilevel, multidomain sentiment analysis corpus of consumer reviews," in *International Conference on Computational Science*. Springer, 2021, pp. 297–312.

[34] A. Wawer and M. Ogrodniczuk, "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*, 2017.

[35] M. Piasecki, K. Młynarczyk, and J. Kocoń, "Recognition of genuine Polish suicide notes," in *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, 2017, pp. 583–591.

[36] A. Wawer and J. Sobiczewska, "Predicting sentiment of Polish language short texts," in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, 2019, pp. 1321–1327.

[37] A. Al-Laith, M. Shahbaz, H. F. Alaskar, and A. Rehmat, "Arasencorpus: A semi-supervised approach for sentiment annotation of a large arabic text corpus," *Applied Sciences*, vol. 11, no. 5, p. 2434, 2021.

[38] M. Boukes, B. van de Velde, T. Araujo, and R. Vliegenthart, "What's the tone? easy doesn't do it: Analyzing performance and agreement between off-the-shelf sentiment analysis tools," *Communication Methods and Measures*, vol. 14, no. 2, pp. 83–104, 2020.

[39] M. Farhadloo, R. A. Patterson, and E. Rolland, "Modeling customer satisfaction from unstructured data using a bayesian approach," *Decision Support Systems*, vol. 90, pp. 1–11, 2016.

[40] J. Kocoń, P. Miłkowski, and M. Zaśko-Zielińska, "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)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 980–991. [Online]. Available: https://www.aclweb.org/anthology/K19-1092

[41] M. Marcińczuk, J. Kocoń, and B. Broda, "Inforex–a web-based tool for text corpus management and semantic annotation," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 2012, pp. 224–230.

[42] G. Hripcsak and A. S. Rothschild, "Technical Brief: Agreement, the F-Measure, and Reliability in Information Retrieval." *JAMIA*, vol. 12, no. 3, pp. 296–298, 2005. [Online]. Available: http://dblp.uni-trier.de/db/journals/jamia/jamia12.html#HripcsakR05

[43] L. Bao, P. Lambert, and T. Badia, "Attention and lexicon regularized LSTM for aspect-based sentiment analysis," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 253–259. [Online]. Available: https://www.aclweb.org/anthology/P19-2035

[44] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[45] H. Xu, L. Shu, P. S. Yu, and B. Liu, "Understanding pre-trained bert for aspect-based sentiment analysis," *arXiv preprint arXiv:2011.00169*, 2020.

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

[47] X. Li, L. Bing, W. Zhang, and W. Lam, "Exploiting bert for end-to-end aspect-based sentiment analysis," in *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, 2019, pp. 34–41.

[48] X. Li, X. Fu, G. Xu, Y. Yang, J. Wang, L. Jin, Q. Liu, and T. Xiang, "Enhancing bert representation with context-aware embedding for aspect-based sentiment analysis," *IEEE Access*, vol. 8, pp. 46 868–46 876, 2020.

[49] M. Marcińczuk, "KPWr n82 NER model (on Polish RoBERTa base)," 2020, CLARIN-PL digital repository. [Online]. Available: http://hdl.handle.net/11321/743

[50] X. Ou and H. Li, "Ynu@ dravidian-codemix-fire2020: Xlm-roberta for multi-language sentiment analysis," *FIRE (Working Notes)*, 2020.

[51] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[52] J. Kocoń, A. Figas, M. Gruza, D. Puchalska, T. Kajdanowicz, and P. Kazienko, "Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach," *Information Processing & Management*, vol. 58, no. 5, p. 102643, 2021.

[53] K. Kanclerz, A. Figas, M. Gruza, T. Kajdanowicz, J. Kocon, D. Puchalska, and P. Kazienko, "Controversy and conformity: from generalized to personalized aggressiveness detection," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 5915–5926. [Online]. Available: https://aclanthology.org/2021.acl-long.460

[54] P. Milkowski, M. Gruza, K. Kanclerz, P. Kazienko, D. Grimling, and J. Kocon, "Personal bias in prediction of emotions elicited by textual opinions," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*. Online: Association for Computational Linguistics, Aug. 2021, pp. 248–259. [Online]. Available: https://aclanthology.org/2021.acl-srw.26