Multi-Level Sentiment Analysis of PolEmo 2.0: Extended Corpus of Multi-Domain Consumer Reviews

Jan Kocoń
Wrocław University of Science and Technology
Wrocław, Poland
jan.kocon@pwr.edu.pl

Piotr Miłkowski
Wrocław University of Science and Technology
Wrocław, Poland
piotr.milkowski@pwr.edu.pl

Monika Zaśko-Zielińska
University of Wrocław
Institute of Polish Studies
Wrocław, Poland
monika.zasko-zielinska@uwr.edu.pl

Abstract

In this article we present an extended version of PolEmo – a corpus of consumer reviews from 4 domains: medicine, hotels, products and school. Current version (PolEmo 2.0) contains 8,216 reviews having 57,466 sentences. Each text and sentence was manually annotated with sentiment in 2+1 scheme, which gives a total of 197,046 annotations. We obtained a high value of Positive Specific Agreement, which is 0.91 for texts and 0.88 for sentences. PolEmo 2.0 is publicly available under a Creative Commons copyright license. We explored recent deep learning approaches for the recognition of sentiment, such as Bidirectional Long Short-Term Memory (BiLSTM) and Bidirectional Encoder Representations from Transformers (BERT).

1 Introduction

In recent years, we have observed a growing interest in methods of effective sentiment analysis, especially in subjective, opinion-forming online texts. This trend is perfectly illustrated by Figure 1, which compares the popularity of two terms: *customer feedback* and *sentiment analysis*. A very dynamic growth has been observed since 2010, which correlates with the increase in the number of scientific research in this area. Many studies focus on the perception of emotion and sentiment in text messages and, for example, their impact on election results (Ramteke et al., 2016), prediction of future events (Zhang and Skiena, 2010) and security issues around the world (Subramaniyaswamy et al., 2017; Al-Rowaily et al., 2015). Automatic sentiment analysis systems have proven to be effective in analyzing many different types of text data such as emails, blogs, news, tweets and books (Medhat et al., 2014). The introduction of advanced computational techniques (machine learning, deep learning) in natural language processing has resulted in a significant increase in sentiment analysis techniques (Zhang et al., 2018). This increase for some languages is effectively limited by the lack of good quality resources for this task, especially in the form of manually annotated corpora (Balahur and Turchi, 2012; Dashtipour et al., 2016).

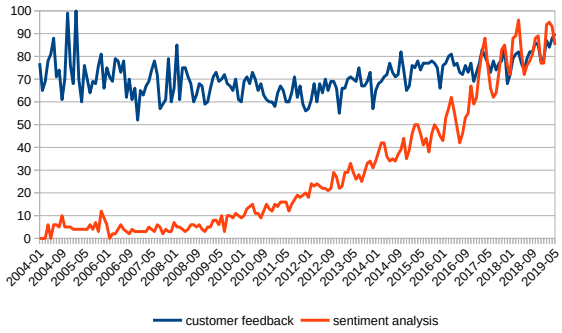

Figure 1: Google Trends (https://trends.google.com) data showing interest in time for search terms "customer feedback" and "sentiment analysis". On the vertical axis 100 means biggest search term popularity.

Analysis of the existing language resources in the area of sentiment analysis shows that they largely concern the English language (Dashtipour et al., 2016). However, there is a clear growing interest in other languages, often much more complex than English (e.g. Slavic languages in the area of loose syntax and rich inflection) and new resources become available for them, e.g., Slovene (Bučar et al., 2018), Czech (Habernal and Brychcín, 2013) or Russian (Rogers et al., 2018). Due to a small number of available corpora manually annotated with sentiment for the Polish language, we decided that the construction of the PolEmo resource will be a valuable contribution to the collection of publicly available resources for sentiment analysis and may in the future provide a basis for the creation of shared tasks, in which the recognition of sentiment for the Polish language

Proceedings of the 23rd Conference on Computational Natural Language Learning, pages 980–991
Hong Kong, China, November 3-4, 2019.

will also be included. Both for the construction of the corpus and for further research, we used the experience from the work on the manual annotation of the Polish WordNet – plWordNet 4.0 Emo (Janz et al., 2017; Kocoń et al., 2018a,b) – as a result of which the sentiment metadata of more than 55,000 lexical units were described.

The main objectives of the article are to present:

- The current state of resources related to the analysis of sentiment for the Polish language;
- The method of selecting data for the PolEmo 2.0 corpus, the annotation method, the annotation results and the analysis of annotation errors;
- The results of research related to the automatic analysis of sentiment, with particular emphasis on the importance of the *text domain* in this topic.

The key contribution of these studies includes:

- Detailed description of the procedure of building PolEmo 2.0: manually annotated corpus of consumer reviews from 4 domains (medicine, school, hotels, products) at 2 levels of sentiment granularity (document, sentence);
- Detailed analysis of manual annotation with regard to frequently occurring errors;
- Development of methods based on deep learning (BiLSTM, BERT), adapted to PolEmo 2.0 corpus, also using sentiment lexicon generated from plWordNet 4.0 Emo;
- Performing tests on sets prepared for the analysis of the quality of methods (1) evaluated on texts within a given domain, (2) evaluated on texts from various domains (3) trained on texts that do not include a given domain and tested on a given domain;
- Comparison of deep learning methods with classic methods (Logistic Regression), especially in the context of the ability to generalize the problem of recognizing sentiment and providing semantic representation, which is as independent of the domain as possible;
- Making PolEmo 2.0 corpus available under an open license.

2 Related Work

There are several well-known resources annotated with sentiment for English, e.g.: MPQA 3.0 (Deng and Wiebe, 2015), the Stanford Sentiment Treebank (Socher et al., 2013), Amazon Product Data (He and McAuley, 2016), Pros And Cons Dataset (Ganapathibhotla and Liu, 2008), corpora developed within the Semantic Evaluation workshops (Nakov et al., 2016; Pontiki et al., 2016), SentiWordNet (Baccianella et al., 2010) or Opinion Lexicon (Hu and Liu, 2004). There are also different approches and tools used for multilingual sentiment analysis (Lo et al., 2017) which are based on transformations on the existing resources. In this section we are focusing on the resources prepared directly for Polish.

2.1 Polish Sentiment Corpora

There are corpora for the Polish language that can be used for automatic sentiment analysis. One of them is a corpus prepared for the sentiment recognition shared task within *PolEval2017*[1] workshop (Wawer and Ogrodniczuk, 2017). The corpus contains 1550 sentences annotated at the level of phrases determined by the dependency parser. The sentences came from consumer reviews and covered 3 domains: *perfume*, *clothing* and *other*. Each node of the dependency tree received one of the three sentiment annotations: -1 (negative), 0 (neutral), 1 (positive). Most of the systems participating in the PolEval2017 competition used Tree LSTM adapted to dependency trees, including the best system, which reached an accuracy of 79% on this data.

Another resource is *HateSpeech*[2] corpus containing 2,000 posts crawled from public Polish web. These texts were annotated for hate speech. The annotator team reached an agreement score of Krippendorff's $\alpha = 0.6$ (Krippendorff, 2018). The SVM model trained on a subset of 1500 texts (containing equal amounts of hate speech and non-hate speech) obtained the precision of 0.8 (Troszyński and Wawer, 2017).

Other interesting resource is the *Polish Corpus of Suicide Notes* (PCSN) (Zaśko-Zielińska, 2013). The PCSN is one of very few such resources in the world. It includes 1,244 genuine SNs that have been scanned and manually transcribed. Each SN

[1] http://2017.poleval.pl/index.php/tasks/

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

was linguistically annotated on several levels, including selected semantic and pragmatic phenomena (Zaśko-Zielińska, 2013). The annotation is stored in a TEI-based format (Marcińczuk et al., 2011) with corrected version in a separate layer. PCSN includes also a subcorpus of 334 counterfeited SNs (elicited). They were created by volunteers who were asked to imitate a real SN for imaginary person whose characteristic had been provided at the beginning of the experiment. Most volunteers were told that the notes written by them would be used *to deceive* the computer program. Due to the sensitive nature of the texts and legal obligations of the author, the corpus is not publicly available. In the experiment described in article (Piasecki et al., 2017) we have collected 3,200 texts from the Internet as examples of non-letters. Using SVM with a rich set of features we obtained 90,06% (F1-score) in the task of distinguishing between genuine SNs, counterfeited SNs and non-letters.

2.2 Polish Sentiment Lexicons

One of the largest Polish sentiment lexical resources in terms of number of annotations is *plWordNet 4.0 Emo*[3] (Janz et al., 2017; Kocoń et al., 2018a). This dataset is available under the WordNet 3.0 license. It was built within CLARIN-PL[4] project (Piasecki, 2014). The manual annotation is done at the level of lexical units (Zaśko-Zielińska et al., 2015). Available values for polarity are: *strong negative, weak negative, neutral, weak positive, strong positive, ambiguous*. One annotator could assign only one of these values for a single lexical unit. There are more than 83,000 annotations covering more than 54,000 lexical units and 41,000 synsets (Kocoń et al., 2018b). About 22,000 of the polarity annotations are different than neutral and these annotations cover 13,000 lexical units and 9,000 synsets (22% of all synsets containing annotated units). plWordNet 4.0 Emo is used in the research presented in this article as a knowledge base for the sentiment recognition task.

Another lexicon is the *Nencki Affective Word List (NAWL)*[5] (Wierzba et al., 2015; Riegel et al., 2015). It is a database of Polish words suitable for studying various aspects of language and emotions. 2902 Polish words from the NAWL were presented to 265 subjects, who were instructed to rate them according to the intensity of each of the five basic emotions: happiness, anger, sadness, fear and disgust. The total number of ratings was 385,575.

The next resource is called the *Polish Sentiment Dictionary*[6] (Wawer, 2012; Wawer and Rogozinska, 2012). It contains 3,704 words with sentiment scores computed using supervised methods presented in (Wawer and Rogozinska, 2012).

Recently, a new resource has appeared in the Sentimenti project, containing a large database of annotated lexical units and annotated texts. Details are described in Section 2.3.

2.3 Sentimenti Project

This year, the first results of the Sentimenti[7] project (Kocoń et al., 2019a) were published, which were aimed at creating methods of analyzing texts written on the Internet in terms of emotions aroused by the recipients of the analysed content. A large database has been created, in which 30,000 lexical units from plWordNet database (Piasecki et al., 2014) and 7,000 texts were annotated. Most of the texts were consumer reviews from the domain of hotels and medicine. The elements were annotated by 20,000 unique Polish respondents in the Computer Assisted Web Interview survey and more than 50 marks were obtained for each element. Within each mark, polarisation of the element, stimulation and basic emotions aroused by the recipients are determined. The total number of manual annotations is 3,742,611 for texts and 19,141,041 for lexical units The first results concerning the automatic recognition of polarity and emotions for this set are presented in (Kocoń et al., 2019a) and propagation of this annotation with the use of Heterogeneous Structured Synset Embeddings is presented in (Kocoń et al., 2019b). Due to the commercial nature of the Sentimenti project, it is planned to publish only 20% of the project data available soon. The data will be published at the main project's site[7].

The Sentimenti project has interested both the scientific community and business. Within the CLARIN-PL project, we decided that in addition to a large annotated plWordNet lexicon, there

[3] http://plwordnet.pwr.edu.pl
[4] https://clarin-pl.eu
[5] https://exp.lobi.nencki.gov.pl/nawl-analysis
[6] http://zil.ipipan.waw.pl/SlownikWydzwieku
[7] https://sentimenti.com/

should also be a large corpus annotated with sentiment, available under an open license. In the next part we present the works related to the preparation of PolEmo.

3 PolEmo Sentiment Corpus

3.1 Motivation

Linguistic research on sentiment recognition involves two approaches: (1) *bottom-up* from the perspective of analysing the occurrence of emotional words and (2) *top-down* from the perspective of the entire document. The first attempt is usually a consequence of the creation of the sentiment lexicon, e.g. manual annotation of the WordNet (Baccianella et al., 2010). The second results from the analysis of the specific text content in which we see that the sentiment of a word or phrase changes under the influence of the surrounding context (Taboada et al., 2008). This change may vary depending on the domain of the text.

A discourse perspective in sentiment analysis is an attempt to address limitations of *bottom-up* methods (e.g. problems with negation, focusing on adjectives). It used findings of Rhetorical Structure Theory (Mann and Thompson, 1988). The attempt bears in mind local and global orientation in the text, discourse structure or topicality (Taboada et al., 2008). It allows the researcher to extract the most important sentences from the text in the perspective of the entire discourse context: nucleus satellite method (Wang et al., 2012). The relevance of the sentences is evaluated in relation to the main topic and the analysis omits some less important parts of the text.

There are interesting articles focused at domain-oriented sentiment analysis (Kanayama and Nasukawa, 2006), where a system is trained on labeled reviews from one source domain but is meant to be deployed on another (Glorot et al., 2011). The latter article describes the research carried out on the Amazon Product Data (He and McAuley, 2016). The ratings were assigned to reviews by authors of the reviews. Moreover, the ratings were applied to the entire text. Our idea was to obtain such a set of reviews that would be rated by the recipients and not by the authors of the content. The annotation should take into account not only the level of the entire review, but also the level of the individual sentences of the review. Additionally, this dataset was supposed to be a multi-domain one, to evaluate potential knowledge transfer across domains.

ID	Name	Source	Author	Subject
H	hotels	tripadvisor.com	visitor	hotel
M	medicine	znanylekarz.pl	patient	doctor
S	school	polwro.pl	student	teacher
P	products	ceneo.pl	buyer	product

Table 1: Each review is described in its domain **ID** and domain **Name** with the given **Source** of the review, **Author**'s type and the general **Subject** of the review.

3.2 Dataset

In the initial part of the work, presented in article (Kocoń et al., 2019), we have chosen online customer reviews from four domains, presented in Table 1. At the beginning of our work we had only 1000 texts for each of the following domains: *school, products, medicine*. In the case of *product* reviews, we also had metadata from the reviewer, how many stars he assigned to a specific review (from 1 to 5, where 5 means the most positive review). We used this information to select the reviews for the corpus, where 200 reviews from each star category were added.

On the basis of a preliminary analysis of several dozen examples of opinions, we have come to the conclusion that neutral examples are very difficult to find in the case of reviews. In the meantime, the corpus was extended by 8000 texts from the category Medicine and 17000 texts from the category Hotels, also with a uniform distribution in relation to the star categories available in the source data (also 1 to 5). In order to capture the phenomenon of neutral text, we decided to add 2000 new texts to each of the last two fields (medicine, hotels). These texts were fragments of articles from information portals on hotel industry[8] and health[9].

In Section 3.3 we present how the genre structure of a customer review affects the text sentiment polarity. It is an enhancement of the discourse perspective in sentiment analysis.

3.3 Pilot Annotation

Our CLARIN-PL pilot study on sentiment analysis of customer reviews was conducted in 2018. The initial part of the analysis included 3,000 reviews. Each text was manually annotated by two annotators: a psychologist and a linguist,

[8] http://ehotelarstwo.com
[9] http://naukawpolsce.pap.pl/zdrowie

who worked according to the general guidelines. The annotation tool used for this task was Inforex[10] (Marcińczuk et al., 2012; Marcińczuk and Oleksy, 2019) – a web-based system for text corpora management, annotation and analysis, available as an open source project. In the pilot project, we decided to deal with the sentiment annotation of the entire text. There was also an attempt to manually extract descriptions of particular aspects of the review. In both annotation cases we used the same tag system that is used in plWordNet Emo for lexical units: [+m] (strong positive), [+s] (weak positive), [-m] (strong negative), [-s] (weak negative), [amb] (ambiguous). We assumed that reviews are always characterised by a certain polarity, which is why we did not use the [0] (neutral) tag in the pilot annotation.

In the process of annotation we focused mainly on the strategic places of the text. In the consumer review these are opening and closing sentences, i.e. a text frame. The opening sentences consist of the general opinion of the author about the subject of the review, and the closing sentences contain the author's recommendation for the review recipients. The annotators have developed their first overall rating based on these two segments. In the text, review authors changed their opinions only subtly. Regardless of the modification of the main opinion in the text, we did not use the [amb] tag when the frame of the text was clearly positive or negative. Polarity of the text frame was influenced not only by the lexical content, but also by non-verbal elements: emoticons or multiplication of punctuation marks, e.g. exclamation marks.

The annotators were also recommended to distinguish those parts of the text that are placed in one sentence, but relate to different aspects (e.g. the teacher's appearance or teaching skills). This task turned out to be very difficult, specially in specifying, even with the help of guidelines, how to mark precisely in the text the boundaries of a given aspect. The Positive Specific Agreement (Hripcsak and Rothschild, 2005) between the annotators in the task of annotating the boundaries of aspects was below 0.15. The concept of annotation was radically changed and presented in Section 3.4.

[10] https://github.com/CLARIN-PL/Inforex

3.4 PolEmo Annotation Guidelines

In the main stage of the project we decided to annotate the sentiment for the whole text (a *meta* level) and the *sentence* level. We assumed that this strategy allows to establish the acceptable value of PSA, because the division of the text into sentences was determined by the MACA[11] tool (Radziszewski and Śniatowski, 2011), so the task was limited only to annotating the sentiment of the sentence. We followed the rule that the *meta* annotation results partially from sentence annotations, however the frame polarity is the main factor for the final meta annotation. We have prepared the following annotation tags, regardless of whether the entire text or sentence is annotated:

- SP – entirely positive;
- WP – generally positive, but there are some negative aspects within the review;
- 0 – neutral;
- WN – generally negative, but there are some positive aspects within the review;
- SN – entirely negative;
- AMB – there are both positive and negative aspects in the text that are balanced in terms of relevance.

This time we used [0] tag (neutral) because in the main stage of the project we extended the corpus with neutral texts presented in Section 3.2. Also reviews that are not neutral often contain neutral sentences.

We tested the new guidelines on a subset of 50 documents, achieving a PSA of 80% for the meta level and 78% for the sentence level. In the second iteration of the annotation guidelines improvement, the values were 87% (meta) and 85% (sentence). In the last iteration of the improvement of the guidelines, the annotators reached a PSA of 90% (meta) and 87% (sentence).

3.5 PolEmo 2.0 Annotation Analysis

We decided to publish the first results of the research on the PolEmo 1.0 corpus when the number of annotated reviews reached 8462 and the number of annotated sentences was 35724 (Kocoń et al.,

[11] Morphological Analysis Converter and Aggregator: http://nlp.pwr.edu.pl/redmine/projects/libpltagger/wiki

2019). Due to the fact that in PolEmo 2.0 there are only those annotated elements that received 2 annotations from linguists and were agreed by the super-annotator, this time we provide 8216 reviews and 57466 sentences. In Section 5 we present Table 7 with the final distribution of annotations and Table 6 with the number of elements in each domain (evaluation data splits). In this section we focus on annotation agreement and annotation errors.

L	D	SN	WN	0	WP	SP	AMB	A
T	H	91.91	36.29	99.41	39.38	91.61	40.11	79.73
	M	94.09	26.42	99.05	22.37	96.28	42.46	89.52
	P	94.06	23.33	100.0	47.62	85.95	33.68	78.76
	S	87.50	20.00	00.00	36.07	92.52	54.19	77.03
	A	92.87	32.20	99.18	37.10	93.48	41.86	83.41
S	H	93.78	00.00	88.40	00.20	93.05	33.94	85.39
	M	90.43	28.75	91.84	26.58	93.43	39.04	88.83
	P	91.27	01.20	48.42	06.90	90.84	30.50	76.82
	S	79.21	00.00	26.56	02.76	81.39	33.73	60.78
	A	91.93	11.94	87.21	07.24	92.12	33.86	84.56

Table 2: Positive Specific Agreement for annotations obtained at the level (L) of text (T) and sentence (S) for each domain (D): hotels (H), medicine (M), products (P), school (S) and all (A).

Table 2 presents PSA values obtained at the level of text and sentence for all domains. The overall PSA value for texts is 83.41% and for sentences is 84.56%. It is worth noting that for the domains to which we have not added neutral texts (products, school), there are practically no neutral annotations at the text level (see Table 7). The highest values are obtained for the most obvious categories (SP, SN and 0), regardless of the level of text description. For the remaining categories PSA value is lower than 40.00% in most cases.

D	A/ WP	SN/ WN	SP/ WP	A/ WN	A/ SN	A/ SP	R R	A/WP/ WN
H	28.55	22.07	18.33	17.08	07.86	03.12	02.99	47.63
M	18.66	26.24	14.29	17.49	12.24	04.37	06.71	37.32
P	28.16	24.27	13.59	19.42	10.68	02.91	00.97	48.54
S	36.21	07.76	28.45	10.34	06.03	08.62	02.59	49.14
A	26.69	22.07	17.82	16.79	09.02	03.89	03.74	45.23

Table 3: Distribution (%) of disagreements between annotators at the text level. A – AMB tag, A/WP – one annotator assigned [AMB], other – [WP]. R is the rest of rare occurring combinations. A/WP/WN is the sum of A/WP, A/WN and WN/WP.

Table 3 presents the distribution of disagreements between annotators at the text level. The most common disagreement is within the pair of tags [AMB/WP]. Nearly half of the disagreements are related to any pair of AMB, WP and WN tags. This suggests that annotators, despite the guidelines, have difficulty in judging the relevance of aspects regardless of the domain, or it is a very subjective task.

D	SN/ 0	A/ SN	A/ 0	A/ WP	SP/ 0	A/ SP	SP/ WP	R R	A/WP/ WN
H	10.52	14.29	05.65	19.80	09.42	07.88	09.31	04.30	30.40
M	34.66	08.10	05.02	04.98	15.93	03.32	06.68	04.35	11.62
P	07.84	21.08	33.57	06.21	05.57	09.00	05.17	02.15	09.93
S	04.63	13.90	26.59	08.66	06.45	20.44	12.19	02.01	12.49
A	16.22	13.80	13.23	11.69	10.20	08.20	08.08	18.58	19.07

Table 4: Distribution (%) of disagreements between annotators at the sentence level. A – AMB tag, A/WP – one annotator assigned [AMB], other – [WP]. R is the rest of rare occurring combinations. A/WP/WN is the sum of A/WP, A/WN and WN/WP.

Table 4 presents the distribution of disagreements between annotators at the sentence level. The most common disagreement is within the pair of tags [SN/0]. This time the cases of disagreements between A/WP/WN tags are less than 20%. Most of the errors are related to the neutral sentence marking. The analysis of specific cases and a discussion with linguists showed that in the task of annotating sentences it is difficult to isolate a sentence from the context and sometimes the annotation of the next sentence was a consequence of the sentiment of the previous sentence.

We have found that it is difficult to decide on the relevance of the aspects and without creating a hierarchy of relevance of aspects for a given domain it will be hard to achieve better agreement for WP/WN/AMB tags. Due to the fact that mistakes are often within these tags, we have combined them into one AMB tag. PolEmo 2.0 will also be available for the original tags, but research (Kocoń et al., 2019) has shown that machine learning methods achieve F-score for WP/WN/AMB classes no higher than PSA. The evaluation data in this research has WP/WN/AMB tags merged into one AMB tag. Table 5 presents PSA values after the merging step. The total PSA increased from 83% to 91% for texts and from 85% to 88% for sentences.

4 Multi-Level Sentiment Recognition

Recently deep neural networks show relatively good performance among all available methods of processing such information (Glorot et al., 2011). Possibility of retrieving data from different sources like social networks (Pak and Paroubek, 2010), publicly available discussion boards or

L	D	SN	0	AMB	SP	A
T	H	91.92	99.42	78.50	91.62	89.39
	M	94.09	99.05	70.25	96.28	93.43
	P	94.06	100.0	77.82	85.95	89.07
	S	87.50	00.00	80.78	92.52	88.32
	A	92.87	99.18	76.87	93.48	90.91
S	H	93.78	88.40	65.64	93.05	89.83
	M	90.43	91.84	59.40	93.43	90.13
	P	91.27	48.42	41.22	90.84	79.12
	S	79.21	26.56	45.48	81.39	65.68
	A	91.92	87.21	56.82	92.12	87.50

Table 5: Positive Specific Agreement for annotations with WP/WN/AMB merged into one AMB tag, obtained at the level (L) of text (T) and sentence (S) for each domain (D): hotels (H), medicine (M), products (P), school (S) and all (A).

marketing platforms connected with proper annotations on training data set can provide not only simple positive, negative or neutral classification but lead to accurate fine-grained sentiment prediction (Guzman and Maalej, 2014).

We selected the same classifiers for the recognition tasks as in (Kocoń et al., 2019): (1) Logistic Regression as a fastText recognition model (Joulin et al., 2017) with KGR10 word embeddings (Kocoń and Gawor, 2018) providing a baseline for text classification; (2) BiLSTM (Zhou et al., 2016) in two variants: KGR10 embeddings as features only and KGR10 embeddings extended with general polarity information from sentiment dictionary described in (Kocoń et al., 2019); (3) BERT (Devlin et al., 2018) with additional sequence classification layer.

We changed the architecture of BiLSTM and BERT architecture. In case of BiLSTM, instead of fixed input length we changed the model to work with text of any length. The input tensor shape is (None, 300) for *embedding-only* variant (BiLSTM) and (None, 306) for *embedding+dictionary* variant (BiLSTMd). We changed the shape of the second gaussian noise layer to (None, 300)/(None, 306), respectively. Next layers remain the same, i.e. (1) BiLSTM layer with 1024 hidden units, (2) dropout layer (ratio 0.2). Last dense layer changed due to the reduction of sentiment labels from 6 to 4 by label merging process described in Section 3.5. For BERT we used the same architecture as in (Kocoń et al., 2019) for the whole texts, but we changed it for sentences. We reduced the maximum sequence length from 512 to 64 (covers more than 99% of sentences) and we increased batch size from 32 to 128.

5 Evaluation

As in article (Kocoń et al., 2019a; Kocoń et al., 2019), we prepared three variants of evaluation of the sentiment classification methods:

- *SD – Single Domain* – evaluation sets created using elements from the same domain;
- *DO – Domain Out* – train/dev sets created using elements from 3 domains, test set from the remaining domain. This variant allows to evaluate the ability of the classification method to capture the domain-independent sentiment features;
- *MD – Mixed Domains* – SD train/dev/test sets joined respectively. This variant allows to examine the ability of the classifier to generalise the task of sentiment analysis in all available domains.

We use *SDT*, *DOT*, and *MDT* abbreviations for *text* evaluation types and *SDS*, *DOS*, and *MDS* for *sentence* evaluation types. We use also prefixes of domains (*Hotels, Medicine, School, Products*) as suffixes for *SD** and *DO** variants, e.g. SDS-H is a Single Domain evaluation type performed on Sentences within *Hotels* domain, whereas DOT-M is a Domain-Out evaluation type performed on Texts trained on texts outside *Medicine* domain and tested on texts from that domain.

Table 6 shows the number of texts and sentences annotated by linguists for all evaluation types, with division into the number of elements within training, validation and test sets. The distribution of labels for each domain (both texts and sentences) is presented in Table 7.

6 Results

Table 8 presents the values of F1-score for each label, global F1-score, micro-AUC and macro-AUC for all evaluation types related to the texts. In case of evaluation for a single domain for each label, fastText (using Logistic Regression) outperformed other classifiers in 16 out of 28 distinguishable cases. The worst results are obtained for *ambiguous* cases, but in 9 out of 13 cases F1-score is higher than 0.5 and this result is much better, than obtained for intermediate labels (*weak* positive and *weak* negative) presented in work (Kocoń

Type	Domain	Train	Dev	Test	SUM
SDT	Hotels	3165	396	395	3956
	Medicine	2618	327	327	3272
	Products	387	49	48	484
	School	403	50	51	504
DOT	!Hotels	3408	427	-	3835
	!Medicine	3955	496	-	4451
	!Products	6186	774	-	6960
	!School	6170	772	-	6942
MDT	All	6573	823	820	8216
SDS	Hotels	19881	2485	2485	24851
	Medicine	18126	2265	2266	22657
	Products	5942	743	742	7427
	School	2025	253	253	2531
DOS	!Hotels	26093	3262	-	29355
	!Medicine	27848	3481	-	31329
	!Products	40032	5004	-	45036
	!School	43949	5494	-	49443
MDS	All	45974	5745	5747	57466

Table 6: The number of texts/sentences for each evaluation type in train/dev/test sets.

Type	Domain	SP	AMB	0	SN
SDT	Hotels	25.61	24.29	10.77	39.33
	Medicine	29.37	09.57	24.11	36.95
	Products	11.16	27.48	00.41	60.95
	School	51.39	38.29	00.00	10.32
	All	27.84	19.47	14.81	37.88
SDS	Hotels	29.55	12.26	17.05	41.15
	Medicine	23.18	06.26	39.48	31.08
	Products	24.61	19.86	09.36	46.17
	School	35.56	37.38	08.89	18.17
	All	26.67	11.98	24.54	36.81

Table 7: The distribution (%) of annotations in a given domain for the following sets: SDT – single domain texts (100%=8216), SDS – single domain sentences (100%=57466).

et al., 2019). BERT classifier performs much better (14 out of 28 cases) in domain-out knowledge transfer evaluation (DOT). For this evaluation type only 4 times fastText was better. These observations are consistent with the results of article (Kocoń et al., 2019a) for *valence* dimensions.

7 Conclusions

BERT's performance is below the expectations of this advanced method in case of the classification of the whole texts. Looking at both tables (8 and 9), BERT's results are the best in 64 out of 182 label-specific cases. BiLSTM outperformed other methods in 48 cases. Adding an external sentiment dictionary helped in 40 label-specific cases. Overall BiLSTM performance is better in 88 out of 182 cases. BERT dominance (when distinguishing between BiLSTM and BiLSTMd) is observed in DOT and all sentence cases. MDT case is the most promising in terms of the further use of the recognition method in applications such as brand monitoring or early crisis detection. The values of the general F1, micro AUC and macro AUC are the highest for BiLSTM variants (see Table 6).

We published PolEmo 2.0 in CLARIN-PL DSpace repository[12] under the Creative Commons 4.0 License. We also intend to test the contextualized embedding that we are currently building using the ELMo deep word representations method (Peters et al., 2018), with the use of the large KGR10 corpus presented in work (Kocoń et al., 2019a). We also want to train the basic BERT model with the use of KGR10 to investigate whether it will improve the quality of sentiment recognition. It is also very interesting to use the propagation of sentiment annotation in WordNet (Kocoń et al., 2018a,b), to increase the coverage of the sentiment dictionary and to potentially improve the recognition quality as well. This objective can be achieved by other complex methods such as OpenAI GPT-2 (Radford et al., 2019) and domain dictionaries construction methods utilising WordNet (Kocoń and Marcińczuk, 2016).

[12] http://hdl.handle.net/11321/710

References

Khalid Al-Rowaily, Muhammad Abulaish, Nur Al-Hasan Haldar, and Majed Al-Rubaian. 2015. Bisal–a bilingual sentiment analysis lexicon to analyze dark web forums for cyber security. *Digital Investigation*, 14:53–62.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204.

Alexandra Balahur and Marco Turchi. 2012. Multilingual sentiment analysis using machine translation? In *Proceedings of the 3rd workshop in computational approaches to subjectivity and sentiment analysis*, pages 52–60. Association for Computational Linguistics.

Jože Bučar, Martin Žnidaršič, and Janez Povh. 2018. Annotated news corpora and a lexicon for sentiment analysis in slovene. *Language Resources and Evaluation*, 52(3):895–919.

T	C	SP	AMB	0	SN	F1	micro	macro
SDT-H	1	83.58	55.56	**98.80**	85.47	80.25	94.28	73.96
	2	**87.31**	64.24	97.56	88.44	**84.05**	95.44	75.87
	3	84.69	**67.39**	96.30	**89.97**	**84.05**	**96.71**	**76.77**
	4	83.50	59.88	93.83	86.90	81.01	95.62	74.94
SDT-M	1	**82.83**	**36.84**	**98.65**	81.48	83.18	95.62	73.35
	2	78.35	18.18	96.60	78.29	77.37	92.99	70.83
	3	75.13	15.87	94.67	76.19	74.31	91.92	70.13
	4	80.75	00.00	97.30	**85.61**	**83.79**	**96.37**	**74.29**
SDT-P	1	**40.00**	**54.55**	00.00	**85.29**	**75.00**	**93.09**	**63.65**
	2	00.00	00.00	00.00	82.93	70.83	87.49	35.50
	3	00.00	08.70	00.00	67.65	50.00	77.82	44.43
	4	00.00	00.00	00.00	84.34	72.92	89.21	39.81
SDT-S	1	**81.36**	**66.67**	00.00	**50.00**	**74.00**	**84.27**	**59.85**
	2	65.31	60.47	00.00	25.00	60.00	76.92	56.23
	3	72.73	57.89	00.00	28.57	64.00	76.12	53.97
	4	71.79	00.00	00.00	00.00	56.00	79.48	51.02
DOT-H	1	77.63	41.77	**90.48**	80.85	73.16	90.39	71.30
	2	74.37	25.00	85.71	73.28	66.08	85.96	67.75
	3	82.52	**52.69**	86.42	82.14	**76.46**	92.77	73.17
	4	**83.84**	47.27	85.71	**83.43**	76.20	**94.15**	**73.46**
DOT-M	1	76.40	20.00	81.89	78.26	**74.01**	89.54	66.99
	2	73.81	20.62	88.89	76.38	70.03	88.34	68.92
	3	73.14	23.08	88.41	78.33	72.48	**91.71**	70.94
	4	**78.11**	**23.30**	**92.20**	**78.84**	72.78	90.81	**71.01**
DOT-P	1	50.00	57.14	00.00	78.69	68.75	90.27	72.90
	2	**66.67**	55.17	00.00	75.86	66.67	88.90	**74.73**
	3	50.00	**64.29**	00.00	**85.25**	**75.00**	**93.76**	72.04
	4	**40.00**	52.17	40.00	82.54	70.83	90.65	72.06
DOT-S	1	72.73	**59.26**	00.00	**33.33**	60.00	76.97	**60.24**
	2	73.47	56.25	00.00	26.67	58.00	82.03	59.79
	3	78.43	23.08	00.00	26.67	50.00	76.92	58.62
	4	**80.00**	52.94	00.00	28.57	**62.00**	**83.71**	58.89
MDT-A	1	82.20	53.64	95.73	84.06	80.37	93.69	73.61
	2	**87.22**	**61.92**	95.20	**88.17**	**84.39**	96.41	**76.44**
	3	84.33	55.63	94.37	86.61	81.71	95.19	75.36
	4	85.40	56.75	**96.07**	85.97	82.07	**96.72**	76.43
MDT-H	1	84.42	54.44	**98.80**	84.37	79.49	93.46	73.62
	2	**86.73**	**65.14**	95.00	**89.09**	**83.80**	96.06	76.33
	3	85.00	58.33	96.30	86.80	81.27	95.24	75.44
	4	85.86	63.58	95.00	87.91	82.78	**96.82**	**76.52**
MDT-M	1	81.82	30.00	96.60	83.27	82.57	95.21	73.30
	2	**88.32**	**36.36**	95.95	87.55	**86.24**	97.16	75.92
	3	84.38	32.14	96.55	**88.12**	84.10	95.77	74.95
	4	86.01	32.65	**97.96**	86.79	85.02	**97.37**	**76.12**
MDT-P	1	50.00	**72.73**	00.00	91.18	**83.33**	93.54	74.56
	2	**66.67**	66.67	00.00	**92.31**	**83.33**	**94.86**	**76.25**
	3	33.33	53.85	00.00	87.10	72.92	92.23	73.35
	4	50.00	42.86	00.00	77.42	64.58	93.26	68.60
MDT-S	1	77.78	66.67	00.00	**57.14**	70.00	85.85	62.86
	2	**87.27**	73.68	00.00	28.57	**78.00**	**94.63**	66.48
	3	**87.27**	**82.35**	00.00	25.00	**78.00**	93.53	**66.70**
	4	84.21	66.67	00.00	00.00	74.00	93.55	66.52

Table 8: F1-scores for text-oriented evaluation. Training sets for evaluation types (T) are the same as in Table 6 rows 1-9. Classifiers: (1) logistic regression (fastText), (2) BiLSTM on word embeddings only (3) BiLSTMd – word embeddings extended using polarity dictionary (4) BERT. Evaluation types are explained in Section 5.

T	C	SP	AMB	0	SN	F1	micro	macro
SDS-H	1	71.98	40.00	64.49	75.90	68.21	83.48	64.44
	2	**82.51**	53.93	72.23	**84.29**	78.31	**93.78**	**73.40**
	3	81.69	51.41	71.21	84.21	77.99	93.43	73.03
	4	82.46	**56.65**	**75.33**	84.21	**78.99**	92.97	72.98
SDS-M	1	67.58	25.90	73.33	64.06	66.18	82.41	61.67
	2	72.36	**31.75**	78.20	71.17	71.96	90.67	70.09
	3	74.49	29.13	79.62	72.58	73.33	**91.18**	**70.39**
	4	**75.69**	27.24	**81.33**	**73.77**	**74.53**	90.76	69.72
SDS-P	1	62.22	**35.34**	33.93	73.19	60.78	80.13	59.96
	2	62.21	28.34	**40.65**	**74.48**	60.78	81.82	61.34
	3	**66.67**	31.46	36.36	73.94	61.32	**83.05**	**62.51**
	4	**66.67**	16.77	36.04	74.07	**62.80**	82.63	60.82
SDS-S	1	**59.34**	58.37	34.29	**42.50**	**54.55**	**77.34**	**59.64**
	2	47.06	47.85	34.29	28.26	43.08	68.40	53.11
	3	45.16	51.61	**35.56**	26.97	43.87	73.38	56.71
	4	51.31	**63.24**	18.18	00.00	51.78	76.17	52.96
DOS-H	1	61.49	26.94	46.98	62.32	54.53	74.29	57.88
	2	72.57	34.60	58.97	74.56	66.56	87.02	67.76
	3	**72.76**	**42.29**	60.50	74.80	**67.81**	**87.89**	**68.21**
	4	70.42	42.12	**60.89**	**74.81**	66.96	85.71	68.07
DOS-M	1	48.58	21.18	56.83	55.56	50.33	71.50	55.83
	2	**61.87**	26.37	62.44	**64.55**	59.47	80.72	63.67
	3	58.68	24.77	63.00	63.00	58.41	80.83	63.51
	4	**61.87**	**27.21**	**66.58**	64.25	**60.75**	**81.80**	**65.08**
DOS-P	1	54.21	23.77	28.92	58.81	47.04	69.03	53.20
	2	66.28	33.33	**35.34**	72.20	**59.30**	**81.78**	**63.82**
	3	**66.47**	30.61	31.50	72.05	58.36	81.15	62.98
	4	64.26	**35.82**	30.95	**72.78**	58.76	78.58	62.11
DOS-S	1	38.52	42.05	**34.92**	30.30	37.15	59.92	52.56
	2	53.25	43.90	19.35	**46.03**	44.27	71.52	58.91
	3	**58.82**	47.50	23.73	41.79	46.64	71.10	**61.07**
	4	55.13	**51.89**	29.79	44.07	**49.01**	**73.09**	59.20
MDS-A	1	66.17	32.36	63.05	66.73	61.27	79.33	61.45
	2	**77.43**	47.21	74.09	**79.40**	74.13	91.48	71.70
	3	77.10	45.88	74.30	78.73	73.70	**91.52**	**71.83**
	4	76.65	**47.76**	**76.70**	79.27	**74.36**	91.19	71.80
MDS-H	1	72.09	33.13	61.42	72.88	65.43	81.43	62.66
	2	**82.82**	51.63	73.18	84.23	78.51	**93.64**	73.19
	3	81.73	54.51	72.68	**84.77**	78.59	93.80	**73.53**
	4	**82.82**	**55.41**	**74.76**	84.52	**78.91**	93.04	73.12
MDS-M	1	63.02	23.12	68.42	61.87	61.37	79.79	60.19
	2	**76.10**	34.88	79.19	75.27	74.44	91.55	70.72
	3	75.27	35.29	79.60	72.51	73.42	91.21	70.72
	4	75.12	**40.00**	**81.83**	**75.50**	**75.67**	**91.71**	**71.52**
MDS-P	1	56.89	31.85	31.75	63.39	52.16	73.92	56.03
	2	67.75	**36.44**	35.93	76.90	63.88	86.03	65.86
	3	**70.65**	35.34	40.00	**77.89**	**65.23**	**87.23**	**67.14**
	4	65.19	33.33	**42.60**	75.53	62.26	84.60	65.06
MDS-S	1	52.17	48.68	26.67	41.44	46.25	69.03	54.72
	2	59.17	**64.42**	34.15	**54.55**	**58.50**	79.16	**62.17**
	3	**61.71**	50.81	30.43	52.00	52.96	78.05	62.10
	4	58.62	53.47	**34.29**	50.53	53.36	**81.38**	61.85

Table 9: F1-scores for sentence-oriented evaluation. Training sets for evaluation types (T) are the same as in Table 6 rows 1-9. Classifiers: (1) logistic regression (fastText), (2) BiLSTM on word embeddings only (3) BiLSTMd – word embeddings extended using polarity dictionary (4) BERT. Evaluation types are explained in Section 5.

Kia Dashtipour, Soujanya Poria, Amir Hussain, Erik Cambria, Ahmad YA Hawalah, Alexander Gelbukh, and Qiang Zhou. 2016. Multilingual sentiment analysis: state of the art and independent comparison of techniques. *Cognitive computation*, 8(4):757–771.

Lingjia Deng and Janyce Wiebe. 2015. Mpqa 3.0: An entity/event-level sentiment corpus. In *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1323–1328.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Murthy Ganapathibhotla and Bing Liu. 2008. Mining opinions in comparative sentences. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 241–248. Association for Computational Linguistics.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 513–520.

Emitza Guzman and Walid Maalej. 2014. How do users like this feature? a fine grained sentiment analysis of app reviews. In *2014 IEEE 22nd international requirements engineering conference (RE)*, pages 153–162. IEEE.

Ivan Habernal and Tomáš Brychcín. 2013. Unsupervised improving of sentiment analysis using global target context. In *Proceedings of RANLP 2013*. Association for Computational Linguistics.

Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*, pages 507–517. International World Wide Web Conferences Steering Committee.

George Hripcsak and Adam S. Rothschild. 2005. Technical Brief: Agreement, the F-Measure, and Reliability in Information Retrieval. *JAMIA*, 12(3):296–298.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

Arkadiusz Janz, Jan Kocoń, Maciej Piasecki, and Monika Zaśko-Zielińska. 2017. plWordNet as a Basis for Large Emotive Lexicons of Polish. In *LTC'17 8th Language and Technology Conference*, Poznań, Poland. Fundacja Uniwersytetu im. Adama Mickiewicza w Poznaniu.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics.

Hiroshi Kanayama and Tetsuya Nasukawa. 2006. Fully automatic lexicon expansion for domain-oriented sentiment analysis. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 355–363. Association for Computational Linguistics.

Jan Kocoń and Michał Gawor. 2018. Evaluating KGR10 Polish word embeddings in the recognition of temporal expressions using BiLSTM-CRF. *Schedae Informaticae*, 27.

Jan Kocoń, Arkadiusz Janz, and Maciej Piasecki. 2018a. Classifier-based Polarity Propagation in a Wordnet. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC'18)*.

Jan Kocoń, Arkadiusz Janz, and Maciej Piasecki. 2018b. Context-sensitive Sentiment Propagation in WordNet. In *Proceedings of the 9th International Global Wordnet Conference (GWC'18)*.

Jan Kocoń, Monika Zaśko-Zielińska, and Piotr Miłkowski. 2019. 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*.

Jan Kocoń, Arkadiusz Janz, Miłkowski Piotr, Monika Riegel, Małgorzata Wierzba, Artur Marchewka, Agnieszka Czoska, Damian Grimling, Barbara Konat, Konrad Juszczyk, Katarzyna Klessa, and Maciej Piasecki. 2019a. Recognition of emotions, polarity and arousal in large-scale multi-domain text reviews. In Zygmunt Vetulani and Patrick Paroubek, editors, *Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 274–280. Wydawnictwo Nauka i Innowacje, Poznań, Poland.

Jan Kocoń, Arkadiusz Janz, Monika Riegel, Małgorzata Wierzba, Artur Marchewka, Agnieszka Czoska, Damian Grimling, Barbara Konat, Konrad Juszczyk, Katarzyna Klessa, and Maciej Piasecki. 2019b. Propagation of emotions, arousal and polarity in WordNet using Heterogeneous Structured Synset Embeddings. In *Proceedings of the 10th International Global Wordnet Conference (GWC'19)*, Wrocław, Poland.

Jan Kocoń and Michał Marcińczuk. 2016. Generating of Events Dictionaries from Polish WordNet for the Recognition of Events in Polish Documents. In *Text, Speech and Dialogue, Proceedings of the 19th International Conference TSD 2016*, volume 9924 of *Lecture Notes in Artificial Intelligence*, Brno, Czech Republic. Springer.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Siaw Ling Lo, Erik Cambria, Raymond Chiong, and David Cornforth. 2017. Multilingual sentiment analysis: from formal to informal and scarce resource languages. *Artificial Intelligence Review*, 48(4):499–527.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Michal Marcińczuk, Monika Zaśko-Zielińska, and Maciej Piasecki. 2011. Structure annotation in the Polish corpus of suicide notes. In *Text, Speech and Dialogue - 14th International Conference, TSD 2011, Pilsen, Czech Republic, September 1-5, 2011. Proceedings*, volume 6836 of *Lecture Notes in Computer Science*, pages 419–426. Springer.

Michał Marcińczuk, Jan Kocoń, and Bartosz Broda. 2012. Inforex – a web-based tool for text corpus management and semantic annotation. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Michał Marcińczuk and Marcin Oleksy. 2019. Inforex —a Collaborative System for Text Corpora Annotation and Analysis Goes Open. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2019*.

Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2016. Semeval-2016 task 4: Sentiment analysis in twitter. In *Proceedings of the 10th international workshop on semantic evaluation (semeval-2016)*, pages 1–18.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, pages 1320–1326.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Maciej Piasecki. 2014. User-driven language technology infrastructure—the case of clarin-pl. In *Proceedings of the Ninth Language Technologies Conference. Ljubljana, Slovenia*.

Maciej Piasecki, Marek Maziarz, Stanisław Szpakowicz, and Ewa Rudnicka. 2014. PLWordNet as the Cornerstone of a Toolkit of Lexico-semantic Resources. In *Proc. 7th International Global Wordnet Conference*, pages 304–312.

Maciej Piasecki, Ksenia Młynarczyk, and Jan Kocoń. 2017. Recognition of genuine Polish suicide notes. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 583–591.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, AL-Smadi Mohammad, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 19–30.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, page 8.

Adam Radziszewski and Tomasz Śniatowski. 2011. Maca — a configurable tool to integrate Polish morphological data. In *Proceedings of the Second International Workshop on Free/Open-Source Rule-Based Machine Translation*.

Jyoti Ramteke, Samarth Shah, Darshan Godhia, and Aadil Shaikh. 2016. Election result prediction using twitter sentiment analysis. In *2016 international conference on inventive computation technologies (ICICT)*, volume 1, pages 1–5. IEEE.

Monika Riegel, Małgorzata Wierzba, Marek Wypych, Łukasz Żurawski, Katarzyna Jednoróg, Anna Grabowska, and Artur Marchewka. 2015. Nencki affective word list (nawl): The cultural adaptation of the berlin affective word list–reloaded (bawl-r) for polish. *Behavior Research Methods*, 47(4):1222–1236.

Anna Rogers, Alexey Romanov, Anna Rumshisky, Svitlana Volkova, Mikhail Gronas, and Alex Gribov. 2018. Rusentiment: An enriched sentiment analysis dataset for social media in russian. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 755–763.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

V Subramaniyaswamy, R Logesh, M Abejith, Sunil Umasankar, and A Umamakeswari. 2017. Sentiment analysis of tweets for estimating criticality and security of events. *Journal of Organizational and End User Computing (JOEUC)*, 29(4):51–71.

Maite Taboada, Kimberly Voll, and Julian Brooke. 2008. Extracting sentiment as a function of discourse structure and topicality. *Simon Fraser Univeristy School of Computing Science Technical Report*.

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

Fei Wang, Yunfang Wu, and Likun Qiu. 2012. Exploiting discourse relations for sentiment analysis. *Proceedings of COLING 2012: Posters*, pages 1311–1320.

Aleksander Wawer. 2012. Mining co-occurrence matrices for so-pmi paradigm word candidates. In *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 74–80. Association for Computational Linguistics.

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*.

Aleksander Wawer and Dominika Rogozinska. 2012. How much supervision? corpus-based lexeme sentiment estimation. In *2012 IEEE 12th International Conference on Data Mining Workshops*, pages 724–730. IEEE.

Małgorzata Wierzba, Monika Riegel, Marek Wypych, Katarzyna Jednoróg, Paweł Turnau, Anna Grabowska, and Artur Marchewka. 2015. Basic emotions in the nencki affective word list (nawl be): New method of classifying emotional stimuli. *PLoS One*, 10(7):e0132305.

Monika Zaśko-Zielińska. 2013. *Listy pożegnalne: w poszukiwaniu lingwistycznych wyznaczników autentyczności tekstu*. Wydawnictwo Quaestio, Wrocław.

Monika Zaśko-Zielińska, Maciej Piasecki, and Stan Szpakowicz. 2015. A large wordnet-based sentiment lexicon for polish. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 721–730.

Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253.

Wenbin Zhang and Steven Skiena. 2010. Trading strategies to exploit blog and news sentiment. In *Fourth international aAAI conference on weblogs and social media*.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 207–212.