TwitterEmo: Annotating Emotions and Sentiment in Polish Twitter

Stanisław Bogdanowicz[1] [0009-0008-3975-464X], Hanna Cwynar[2] [0009-0004-3011-1446], Aleksandra Zwierzchowska[3] [0000-0002-7322-7535], Cezary Klamra[4][0000-0003-4321-8862], Witold Kieraś[5] [0000-0002-8062-5881] and Łukasz Kobyliński[6] [0000-0003-2462-0020].

Institute of Computer Science, Polish Academy of Sciences, Jana Kazimierza 5, 01-248 Warszawa, Poland

[1]stan.bogdanowicz97@gmail.com, [2]hanna.cwynar@gmail.com, [3]aazwierzchowska@gmail.com, [4]c.klamra@ipipan.waw.pl, [5]wkieras@ipipan.waw.pl, [6]lkobylinski@ipipan.waw.pl

Abstract. This article presents TwitterEmo, a new dataset for emotion and sentiment analysis in Polish. TwitterEmo provides a non-domain-specific and colloquial language dataset, which includes Plutchik's eight basic emotions and sentiment annotations for 36,280 tweets collected over a one-year period. Additionally, a sarcasm category is included, making this dataset unique in Polish computational linguistics. Each entry was annotated by at least four annotators. We present the results of the evaluation using several language models, including HerBERT and TrelBERT. The TwitterEmo dataset is a valuable resource for developing and training machine learning models, broadening possible applications of emotion recognition methods in Polish, and contributing to social studies and research on media bias.

Keywords: Natural Language Processing, Emotion Recognition, Sentiment Analysis.

1 Introduction

The field of computational linguistics and natural language processing (NLP) increasingly uses social media platforms like Twitter to gather real-world data on user opinions and emotions. Potential restrictions or cancellation of Twitter's unrestricted academic access make the value of such datasets even more significant. This paper presents a new dataset of Polish tweets annotated with emotions, sentiment, and sarcasm, which is a valuable resource for training machine learning models. Unlike existing domain-specific emotion-annotated datasets, TwitterEmo fills a significant gap in the field by providing a non-domain-specific dataset for investigating emotions and sentiment in Polish texts. The dataset covers various topics, making it useful for social studies and research on media bias. Additionally, the dataset includes sarcasm annotations, which is a supplementary characteristic that may contribute to the study of emotions in language. This paper describes the creation and annotation process of the dataset and presents data analysis, including correlations between emotions and the distribution of sentiment over time. Preliminary tests were conducted to evaluate the data in terms of machine learning.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_20

2 Related Works

A large number of sentiment analysis datasets developed to date have been specifically designed for widely known languages, such as English [4], yet few similar resources exist for lesser-known languages such as Polish.

With regard to sentiment annotation, one notable mention of a Polish resource is the dataset prepared for the sentiment recognition shared task during the PolEval2017 campaign [16]. The dataset consists of 1,550 sentences obtained from consumer reviews from three specific domains. The sentences are annotated at the level of phrases determined by the dependency parser, with each node of the dependency tree receiving one of the three sentiment classes: negative, neutral, and positive. Other instances of sentiment annotated domain-specific corpora for Polish are PolEmo 2.0 [7] and the Wroclaw Corpus of Consumer Reviews [8].

In emotion annotation, it is common to employ Plutchik's [13] Wheel of Emotions, which is exemplified by the collection of texts under the Sentimenti project [9], specifically developed for Polish. Additionally, a set of guidelines [15] has been developed to assist with emotion annotation in Polish consumer reviews.

Furthermore, as preliminary evidence suggests that the occurrence of sarcasm may cause up to a 50% decrease in accuracy in the automatic detection of sentiment in text and recognition of sarcasm based on the occurrence of hashtags may provide inaccurate results [14], manual annotation of sarcasm appears to be a fairly relevant consideration for machine learning applications.

As opposed to aforementioned datasets for Polish, text samples in our dataset come from unspecified and diverse domains. They exhibit a variety of stylistic features and tend to be highly opinionated due to their informal nature, thus rendering them a valuable source for sentiment analysis. To the best of our knowledge, the TwitterEmo dataset is the first for Polish that includes sentiment and emotion annotation along with experimental sarcasm annotation.

3 Data Gathering

3.1 Data Source and Preprocessing

The data used for annotation comes from the period from 01.09.2021 to 31.08.2022. A hundred tweets per day were scraped giving 36,500 entries of up to 280 characters, some of which were excluded in the process of annotation concluding in 36,280 tweets total. A sample of 156 tweets annotated as irrelevant is included in the final dataset. Our dataset comprises tweets collected from an extended version of the set of accounts considered in the Ogrodniczuk and Kopeć [12] study (75%) as well as from the category Twitter Trends (25%). Samples were collected from 36,080 unique accounts. All emoji symbols were removed in order to avoid them from suggesting the emotions. User tags were left for the references to other users' tweets to be detectable. In the published version, user tags were replaced with a unified token '*@anonymized_account*.'

3.2 Annotation Methodology and Guidelines

We followed the methodology used in the creation of similar datasets for Polish to obtain compatible and comparable results. We performed text-level annotations on each tweet separately, with four annotators (five for the first 8,000 tweets) providing their annotations. The annotations were then consolidated, and any discrepancies were discussed during group meetings to achieve a consistent and unambiguous result. In cases where uniformity was impossible, the sentiment was rendered ambiguous. We used Google Spreadsheets to carry out the annotation process, with each annotator allocated an individual sheet. A separate spreadsheet was used to collect annotations from each annotator's sheet, allowing us to identify discrepancies, calculate inter-annotator agreement, and facilitate group annotation.

We devised a set of guidelines for annotating emotions, sentiment, and sarcasm, which serves as an expansion to the previously formulated CLARIN-PL instruction developed for annotating product and service reviews with emotions [15].

Emotions. We adopted the Plutchik’s [13] model, which delineates a discrete set of eight basic emotions (‘joy’, ‘sadness’, ‘trust’, ‘disgust’, ‘fear’, ‘anger’, ‘surprise’, ‘anticipation’). For each tweet in the dataset, annotators were asked to assign emotions consistent with the emotions expressed by the author of the text. It was emphasized that contrasting emotions should not be assigned to a single tweet. Moreover, the annotators were instructed to annotate predetermined dyads (e.g., ‘love’, ‘pessimism’, ‘aggression’). Finally, in the event of a conflict between emotions the annotators were asked to select a predominant emotion or, in the most difficult cases, leave the tweet for group annotation.

Sentiment. The sentiment could take one of the following four values: ‘negative’, ‘positive’, ‘neutral’, or ‘ambivalent’ (indicated as ‘positive’ + ‘negative’). For each tweet, after annotating emotions, the annotators were instructed to annotate its sentiment, which ideally would follow from the previously annotated emotions (‘joy’/’trust’ corresponds to ‘positive’; ‘anticipation’/’surprise’ corresponds to ‘neutral’; ‘fear’/’sadness’/’disgust’/’anger’ corresponds to ‘negative’). Nevertheless, the annotation of emotions and sentiment remained independent. In ambiguous instances, if a tweet contained both positive and negative sentiment, and neither was predominant, the annotators were allowed to mark both.

Sarcasm. To manually annotate sarcasm, we followed the theoretical approach mentioned in previous studies [5] and identified ‘sarcasm’ as an inconsistency between the literal content conveyed in the text (positive) and the sentiment intended by the author (negative). In cases where such a contrast could not be detected, annotators could mark the statement as a snide remark/irony (annotated as ‘disgust’ in the spreadsheet). For every sample in the dataset, annotators marked sarcasm in a binary way, indicating its presence (1) or absence (0).

3.3 Positive Specific Agreement

We used Positive Specific Agreement (PSA) to evaluate the results of manual annotation. The instructions were discussed and adjusted during the annotation process to increase the PSA values. However, due to the linguistic and thematic diversity of the data, it was difficult to provide unambiguous instructions for annotation, which resulted in low agreement between annotators. The level of agreement was higher for sentiment annotations (76.20%) than for emotions (55.29%) or sarcasm (25.27%). The overall PSA for the dataset was 66.31%.

DOI: 10.1007/978-3-031-36021-3_20

3.4 Rendering the Final Annotation

The rendition of the final dataset was composed of two steps. Firstly, the annotations were automatically summed up given the conditions presented in Table 1. The conflicting pairs of emotions were considered. Hence, a threshold of minimum two consistent annotations of an emotion was set, along with the restriction that the number of annotations of a given emotion needs to be higher than the number of annotations of the opposed emotion. While it was intended for the sentiment to match and follow directly from the emotions it was not always the case. Thus, sentiment was rendered independently. The conditions never rendered a 'None' sentiment.

Table 1. Conditions for automated totaling of annotations of opposed pairs of emotions and sentiment

Result	Conditions	Example
Emotion A	AND(A>B;A>1)	joy=3; sadness=1
Emotion B	AND(A<B;B>1)	joy=1; sadness=2
Conflict	OR(AND(A=1;B=1);AND(A=2;B=2))	joy=2; sadness=2
None	else	joy=1; sadness=0
Positive (A)	AND(A>B;A>C;A>1)	positive=4; negative=0; neutral=1
Negative (B)	AND(A<B;B>C; B>1)	positive=0; negative=3; neutral=1
Neutral (C)	AND(B<C;A<C;C>1)	positive=1; negative=1; neutral=2
None	AND(A<2;B<2;C<2)	positive=1; negative=1; neutral=1
Ambivalent	else	positive=2; negative=2; neutral=0

The second step required a group discussion of specific cases. After singular annotations were added up according to the conditions, each case of conflict between emotions was discussed in group as well as each tweet marked with 'sarcasm' and/or 'to be discussed'. Totally a number of 5,207 tweets were dealt with by group annotation, the results of which were written over the results of automatic rendition.

4 Data Analysis

4.1 Overall Results

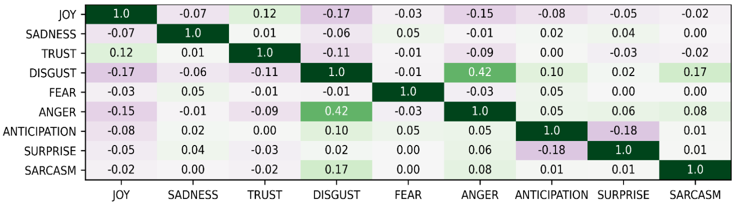
Table 2 presents the overall results of the annotation process. The most frequent (34,85%) emotion is 'anticipation'. Which may be a result of a very liberal instruction for annotating this emotion – most tweets referring to the future with various attitudes were counted as anticipation. Surprisingly, the least frequent emotion in the dataset is 'fear' with only 0,90% frequency of occurrence.

Table 2. Overall counts and frequencies for each category

Category	Count	Frequency	Category	Count	Frequency
Joy	4168	11,49%	Surprise	2352	6,48%
Sadness	1680	4,63%	Positive	3986	10,99%
Trust	1620	4,47%	Negative	10729	29,57%
Disgust	8361	23,05%	Neutral	**18385**	**50,68%**
Fear	**328**	**0,90%**	Ambivalent	3039	8,38%
Anger	6364	17,54%	Sarcasm	751	2,07%
Anticipation	**12645**	**34,85%**	Irrelevant	156	0,43%

4.2 Co-occurrence of Emotions

We have computed correlations of every pair of emotions (see Figure 1). The most strongly positively correlated pairs were ‘disgust’ and ‘anger’ (r=0.42), ‘disgust’ and ‘sarcasm’ (r=0.17), ‘joy’ and ‘trust’ (r=0.12), and ‘disgust’ and ‘anticipation’ (r=0.11). Combination of ‘disgust’ and ‘anger’ corresponds to the predefined ‘contempt’ dyad, while combination of ‘joy’ and ‘trust’ corresponds to the ‘love’ dyad. The most strongly negatively correlated pairs were ‘anticipation’ and ‘surprise’ (r=-0.18), ‘joy’ and ‘disgust’ (r=-0.17), ‘joy’ and ‘anger’ (r=-0.15) and ‘trust’ and ‘disgust’ (r=-0.11). These results were fully expected and justified by the instructions given to annotators not to assign conflicting emotions to one text.



	JOY	SADNESS	TRUST	DISGUST	FEAR	ANGER	ANTICIPATION	SURPRISE	SARCASM
JOY	1.0	-0.07	0.12	-0.17	-0.03	-0.15	-0.08	-0.05	-0.02
SADNESS	-0.07	1.0	0.01	-0.06	0.05	-0.01	0.02	0.04	0.00
TRUST	0.12	0.01	1.0	-0.11	-0.01	-0.09	0.00	-0.03	-0.02
DISGUST	-0.17	-0.06	-0.11	1.0	-0.01	0.42	0.10	0.02	0.17
FEAR	-0.03	0.05	-0.01	-0.01	1.0	-0.03	0.05	0.00	0.00
ANGER	-0.15	-0.01	-0.09	0.42	-0.03	1.0	0.05	0.06	0.08
ANTICIPATION	-0.08	0.02	0.00	0.10	0.05	0.05	1.0	-0.18	0.01
SURPRISE	-0.05	0.04	-0.03	0.02	0.00	0.06	-0.18	1.0	0.01
SARCASM	-0.02	0.00	-0.02	0.17	0.00	0.08	0.01	0.01	1.0

Fig. 1. Correlations of pairs of emotions

4.3 Distribution of Sentiment Over Time

Figure 2 depicts the distribution of annotated sentiment over time. Ambivalent sentiment was assigned to the tweets published after mid-November more frequently due to the fact that the first 8,000 tweets the number of annotators was reduced from 5 to 4. We suspected that the beginning of the Russian invasion of Ukraine might affect the distribution of the sentiment of tweets in subsequent months. However, no such changes can be identified in the graph. However, in tweets posted around New Year’s Eve, a significant increase of positive sentiment can be observed.

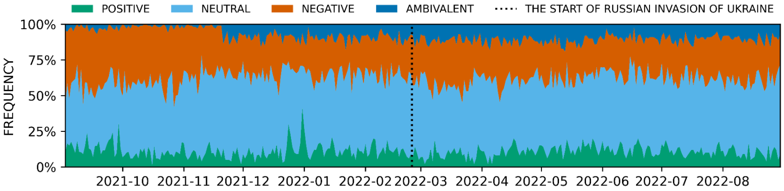


Fig. 2. Distribution of sentiment over time.

4.4 Co-occurrence of Emotions and Sentiment

Because of independent annotation of emotions and sentiment we have examined their co-occurrence (see Figure 3). Unsurprisingly, positive emotions are positively correlated with ‘positive’ sentiment (‘joy’, r=0.79; ‘trust’, r=0.39), and negative emotions, with ‘negative’ sentiment (‘anger’, r=0.62; ‘disgust’, r=0.71; ‘fear’, r=0.11; ‘sadness’, r=0.22). Sarcasm was most strongly correlated with ‘negative’ sentiment (r=0.18). Sentiment of tweets containing ‘surprise’ or ‘anticipation’ was most frequently annotated as ‘negative’ or ‘neutral’ rather than ‘positive’ or ‘ambivalent’.

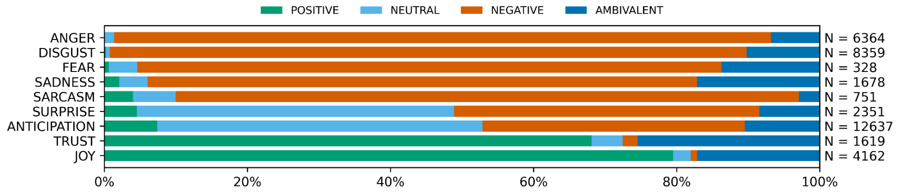


Fig. 3. Distribution of sentiment for every emotion.

5 Experiments

We evaluated three models on our dataset: one for emotion recognition, one for sentiment classification, and a multi-task model that performs both tasks jointly. The F1-score metric was used to measure the predictive performance of emotions and sentiment classifications, with the macro average used to address class imbalance. The models were trained with a fixed learning rate, batch size, and number of epochs. We excluded irrelevant and sarcastic tweets. The dataset was split into train(80%), dev(10%), and test(10%) sets using iterative stratified sampling.

As a main model we adopt HerBERT [10], a Transformer-based model trained specifically for Polish. Because of the specific tweets style, we also utilized HerBERT that was trained using almost 100 million messages extracted from Polish Twitter TrelBERT [1]. This model is available only in the base architecture version, so we decided to train our own large model (HerBERT-large-T), for which we used tweets from the TwitterEmo and PolEval2019 task [11]. We use this set of tweets to fine-tune the HerBERT-large model using the intermediate masked language model task as the training objective with the probability of 15% to randomly mask tokens in the input. Additionally, the multilingual XLM-R model [3] was also taken into consideration, as well as its version pretrained on 198 million multilingual tweets [2].

5.1 Sentiment Analysis

Table 3 reports sentiment classification results. As well as in the following case of emotion recognition, the best model turned out to be HerBERT-large (62.70% average F1-macro).

Table 3. Sentiment classification F1 results on TwitterEmo test set for single-task models

Sentiment	HerBERT base	HerBERT large	TrelBERT	HerBERT large-T	XLM-R base	XLM-R large	XLM-Twitter
Positive	68.89	72.11	**73.87**	72.90	67.19	73.58	69.37
Negative	76.24	**79.27**	78.44	78.54	73.17	78.07	73.29
Neutral	83.86	**84.60**	84.46	83.99	81.33	84.11	82.18
Ambivalent	4.97	**14.83**	9.60	12.56	1.91	10.58	3.56
Macro avg	58.49	**62.70**	61.59	62.00	55.90	61.59	57.10

5.2 Emotion Recognition

Table 4 displays the results of emotion recognition evaluation. The highest average macro F1-score was achieved by the HerBERT-large model, reaching 57.17%. The two emotions with the lowest recognition scores were 'trust' and 'fear,' which is consistent with their low frequency in the dataset, with 4.47% and 0.90% frequencies, respectively (see Table 2). Among the Polish base models, TrelBERT achieved over 3pp higher macro F1-score compared to the model without pretraining on tweets. However, the HerBERT-large-T results were below expectations, possibly due to the small training set. It only marginally improved performance for 'trust,' 'anger,' and 'surprise'. The results of the multilingual XLM-R model were good, but they were still lower than the results of the native models.

Table 4. Emotion recognition F1 results on TwitterEmo test set for single-task models

Emotion	HerBERT	HerBERT	TrelBERT	HerBERT	XLM-R	XLM-R	XLM-
	base	large		large-T	base	large	Twitter
Joy	68.26	70.99	72.80	70.85	68.24	72.83	66.33
Sadness	41.88	50.17	43.42	49.03	39.23	47.95	37.35
Trust	11.46	32.52	22.97	33.33	3.59	30.04	8.89
Disgust	67.09	73.21	70.89	72.70	63.43	71.82	64.21
Fear	10.81	40.00	5.71	38.10	0.00	33.33	0.00
Anger	61.51	66.05	63.91	66.38	58.53	66.61	60.42
Anticipation	69.90	71.16	71.57	70.84	69.55	71.60	69.61
Surprise	46.68	53.30	51.96	53.52	41.64	49.28	41.73
Macro avg	47.20	**57.17**	50.40	56.84	43.03	55.43	43.57

6 Conclusions

This paper presents the results of creating a baseline dataset for emotion recognition and sentiment analysis of non-domain-specific and colloquial language tweets in Polish. The dataset covers various topics and spans a full year, allowing for an analysis of emotions and sentiment over time. TwitterEmo fills the gap in the field of emotion and sentiment datasets in Polish by overcoming the limitations of the domain-centered approach and broadening applications of emotion recognition methods in social studies. Annotation of sarcasm introduces novelty to computational linguistics in Polish. The dataset will be made available via an online repository.[1]

Results of preliminary model training on the dataset were presented, showing the HerBERT large model achieving the highest average F1-score. Further research will determine the utility of the dataset in detecting emotions related to specific events and topics in Polish public debate. The dataset can also be used for media bias detection in social media and to detect trends in public opinion based on social media entries. Finally, the absence of emoji symbols in the annotation can be confronted with emotions ascribed to them, which may present interesting results on the use of emoji in Twitter and their potential in emotion recognition research.

Acknowledgements
The project is financed under the 2014-2020 Smart Development Operational Programme, Priority IV: Increasing the scientific and research potential, Measure 4.2: Development of modern research infrastructure of the science sector, No. POIR.04.02.00-00C002/19, "CLARIN - Common Language Resources and Technology Infrastructure" and by the project co-financed by the Minister of Education and Science under the agreement 2022/WK/09.

[1] The repository available at https://huggingface.co/datasets/clarin-pl/twitteremo.

References

1. Bartczuk, J., Dziedzic, K., Falkiewicz, K., Kotyla, A., Szmyd, W., Zobniów, M., Zygadło, A.: TrelBERT, https://huggingface.co/deepsense-ai/trelbert, last accessed 2023/03/01.
2. Barbieri, F., Espinosa, A., Camacho-Collados, J.: XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference, pp. 258-266. European Language Resources Association, Marseille, France, 2022.
3. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. CoRR (2019).
4. Dashtipour, K., Poria, S., Hussain, A., Cambria, E., Hawalah, A. Y. A., Gelbukh, A., Zhou, Q.: Multilingual sentiment analysis: state of the art and independent comparison of techniques. Cogn Comput 8, pp. 757–771 (2016). doi.org/10.1007/s12559-016-9415-7.
5. Davidov, D., Tsur, Or., Rappoport, A.: Semi-supervised recognition of sarcastic sentences in Twitter and Amazon. Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL '10). Association for Computational Linguistics, pp. 107–116 (2010).
6. Ganesan, K. A., Zhai, C. X.: Opinion-based entity ranking. Information Retrieval, pp. 116-150 (2012).
7. Kocoń, J., Miłkowski, P., Zaśko-Zielińska, M.: Multi-level sentiment analysis of Pol-Emo 2.0: Extended corpus of multi-domain consumer reviews. In: Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), pp. 980-991. Association for Computational Linguistics, Hong Kong, China (2019a).
8. Kocoń, J., Zaśko-Zielińska, M., Miłkowski, P., Janz, A., Piasecki, M.: Wrocław Corpus of Consumer Reviews Sentiment. CLARIN-PL (2019c). hdl.handle.net/11321/700.
9. Kocoń, J., Janz, A., Miłkowski, P., Riegel, M., Wierzba, M., Marchewka, A., Czoska, A., Grimling, D., Konat, B., Juszczyk, K., Klessa, K., Piasecki, M.: Recognition of emotions, polarity and arousal in large-scale multi-domain text reviews. In: Vetulani, Z., Paroubek, P. (eds.) Human Language Technologies as a Challenge for Computer Science and Linguistics, pp. 274-280. Wydawnictwo Nauka i Innowacje, Poznań, Poland (2019d).
10. Mroczkowski, M., Rybak, P., Wróblewska, A., Gawlik, I.: HerBERT: Efficiently Pretrained Transformer-based Language Model for Polish. In: Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing. pp. 1–10. Association for Computational Linguistics, Kiyv, Ukraine (Apr 2021).
11. Ogrodniczuk, M., Kobyliński, Ł.: Proceedings of the PolEval 2019 Workshop. Warszawa, Poland (2019).
12. Ogrodniczuk, M., Kopeć, M.: Lexical Correction of Polish Twitter Political Data. In: Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, pp. 115–125. Association for Computational Linguistics, Vancouver, Canada (2017).
13. Plutchik, R.: The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. American Scientist 89(4), pp. 344-350 (2001). http://www.jstor.org/stable/27857503.
14. Sykora, M., Elayan, S., Jackson, T. W.: A qualitative analysis of sarcasm, irony and related #hashtags on Twitter. Big Data & Society (2020). doi.org/10.1177/2053951720972735.
15. Wabnic, K., Zaśko-Zielińska, M., Kaczmarz, E., Matyka, E., Zajączkowska, A., Kocoń, J.: Guidelines for annotating consumer reviews with basic emotions. CLARIN-PL (2022). hdl.handle.net/11321/909.
16. Wawer, A., Ogrodniczuk, M.: 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).

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_20