Personal Bias in Prediction of Emotions Elicited by Textual Opinions

Piotr Miłkowski*, **Marcin Gruza***, **Kamil Kanclerz***,
Przemysław Kazienko*, **Damian Grimling**†, **Jan Kocoń***
*Wrocław University of Science and Technology, Wrocław, Poland
†Sentimenti Sp. z o.o., Poznań, Poland
{piotr.milkowski,marcin.gruza,kamil.kanclerz,
przemyslaw.kazienko,jan.kocon}@pwr.edu.pl
damian@sentimenti.com

Abstract

Analysis of emotions elicited by opinions, comments, or articles commonly exploits annotated corpora, in which the labels assigned to documents average the views of all annotators, or represent a majority decision. The models trained on such data are effective at identifying the general views of the population. However, their usefulness for predicting the emotions evoked by the textual content in a particular individual is limited. In this paper, we present a study performed on a dataset containing 7,000 opinions, each annotated by about 50 people with two dimensions: valence, arousal, and with intensity of eight emotions from Plutchik's model. Our study showed that individual responses often significantly differed from the mean. Therefore, we proposed a novel measure to estimate this effect – Personal Emotional Bias (PEB). We also developed a new BERT-based transformer architecture to predict emotions from an individual human perspective. We found PEB a major factor for improving the quality of personalized reasoning. Both the method and measure may boost the quality of content recommendation systems and personalized solutions that protect users from hate speech or unwanted content, which are highly subjective in nature.

1 Introduction

Emotions are a very important component of natural human communication. Collectively, we tend to react quite similarly emotionally to phenomena around us, but at the level of the individual, some differences can be discerned in the intensity of the emotions experienced. Various emotional models have been used in different studies. In Russell and Mehrabian (1977), emotional states are located in a multidimensional space, with valence (negative/positive), arousal (low/high) and dominance explaining most of the observed variance. Another approach distinguishes different number of basic, discrete emotions, e.g. six by Ekman and Friesen (1976) and eight by Plutchik (1982).

We can observe continuous interest in sentiment analysis and emotion recognition within the filed of natural language processing (Kocoń and Maziarz, 2021; Alswaidan and Menai, 2020; Kanclerz et al., 2020). Recently, they commonly rely on deep machine learning methods applied to large amounts of textual data (Yadav and Vishwakarma, 2020; Kocoń et al., 2019b; Kocoń et al., 2019). Nevertheless, emotion recognition remains a challenging task. One of the reasons is the lack of high quality annotated data, where annotators are a representative sample of the whole population. Commonly, a small number (usually 2 to 5) of trained annotators are involved. Due to differences between individual opinions, reinforced by multiple choice possibilities (6 or 8 emotions), this often leads to low inter-annotator agreement (Hripcsak and Rothschild, 2005). Averaging the annotations collected in such a way can still be a good input for effective systems recognizing the most likely emotional responses shared by most people. This, however, is not suitable to make accurate inferences about emotions to be evoked in specific individuals.

In this work, we developed a method to predict text-related emotions that most closely reflect the reactions of a given reader. In addition to the classical approach of providing only texts to the model input, we extended it with our new feature – Personal Emotional Bias (PEB). It reflects how an individual perceived the texts they evaluated in the past. In this way, we switched from averaging labels for annotated texts to individual text annotations. We tested the impact of PEB on individual recognition quality of emotion dimensions, also in a setup including multilingual transformer-based architecture for the following languages: Dutch, En-

Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop, pages 248–259
August 5–6, 2021. ©2021 Association for Computational Linguistics

glish, Polish, French, German, Italian, Portuguese, Russian and Spanish. Our experimental evaluation revealed that emotional annotation of just a few texts is appears to be enough to calculate the approximate value of Personal Emotional Bias for a given user. This, in turn, enables us to significantly improve personalized reasoning. Since texts are independently annotated with ten emotional states, each with its own level, we trained and tested both multi-task classifiers and multivariate regressors.

This work is inspired by our initial idea of human-centred processing presented in (Kocoń et al., 2021). In addition, in paper (Kanclerz et al., 2021), we have shown that mixing user conformity measures with document controversy is efficient in personalized recognition of aggressiveness in texts.

2 Related work

The studies have shown that the recognition of emotions should take into account the subjective assessments of individual annotators (Neviarouskaya et al., 2009; Chou and Lee, 2019; Kocoń et al., 2019a). A personal bias related to the individual beliefs may have its origins in the demographic background and many factors such as the first language, age, education (Wich et al., 2020a; Al Kuwatly et al., 2020), country of origin (Salminen et al., 2018), gender (Bolukbasi et al., 2016; Binns et al., 2017; Tatman, 2017; Wojatzki et al., 2018), and race (Blodgett and O'Connor, 2017; Sap et al., 2019; Davidson et al., 2019; Xia et al., 2020). The uniqueness of person's annotations may also be derived from their political orientations and not respecting them can significantly reduce the effectiveness of the classifier (Wich et al., 2020b).

The most common approach to mitigate the impact of personal bias on method performance is to utilize only annotations provided by the experts (Waseem, 2016). However, we should be aware that selecting a small group of experts poses a risk of involving too few annotators for too many documents (Wiegand et al., 2019) or creating unfair models, that will discriminate minorities (Dixon et al., 2018). Besides, it may be difficult to find the sufficient number of experts. To resolve this, non-expert annotators can be involved. An average of annotations from non-expert is enough to achieve expert-level labeling quality (Snow et al., 2008). Personal bias also affects the model evaluation process. Therefore, annotations from a separate set of annotators should be used in the training and test set (Geva et al., 2019).

The high variety of annotators' beliefs directly impacts the diversity of their subjective assessments. It often means that there is no single correct label for a given text (Aroyo and Welty, 2013). In such case, Bayesian probabilistic models can be used to estimate consensus level, which can then be converted to categorical values using simple methods, e.g. thresholding (Kara et al., 2015). Another solution is to regard disagreement in annotations as a positive factor that will provide more information about single humans. This ambiguity can be utilized in many ways. Patterns discovered from differences in annotations can be exploited both to group like-minded individuals (Akhtar et al., 2020) and to automatic detect spammers, deliberately introducing noise into their assessments (Raykar and Yu, 2012; Soberón et al., 2013). On the other hand, too high annotations similarity level may be related to the conformity bias, which reflects an excessive influence of the group's beliefs on its members (Gao et al., 2019). Moreover, annotation disagreement can determine the ambiguity of a given text (Aroyo and Welty, 2013). The variability between annotators can also be used to generate soft labels such as inter-annotator standard deviation, which may be an additional feature of a given sample (Eyben et al., 2012). Such soft labels can also be a good source of information about annotators themselves, e.g. to estimate the unanimity of a specific social group in recognizing emotions (Steidl et al., 2005). Another approach is to leverage the ensemble model architecture to incorporate knowledge regarding the subjectivity of emotion recognition (Fayek et al., 2016). In order to reduce the potential noise caused by relying solely on subjective annotations, a hybrid method can be applied mixing both individual ratings and majority voting (Chou and Lee, 2019). The final model consists of multiple sub-models using annotations of individuals separated and combined. All sub-models are fused providing one general and non-personalized decision.

The topic of emotion personalization was explored in the context of social photos (Zhao et al., 2016) or emotions evoked by music (Yang et al., 2007). However, in the context of text analysis, it has not been studied sufficiently yet.

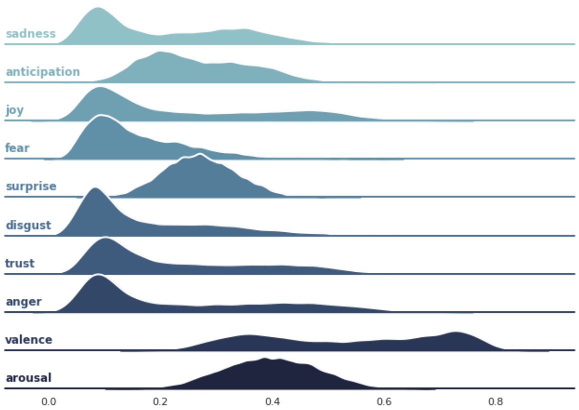


Figure 1: Rating distributions within emotional categories. All values are normalized to the interval [0,1].

3 Dataset and annotation procedure

To create a Sentimenti[1] dataset, a combined approach of different methodologies were used, namely: Computer Assisted Personal Interview (CAPI) and Computer Assisted Web Interview (CAWI) (Kocoń et al., 2019a). Two studies were carried out involving evaluation of: 30,000 word meanings (CAWI1) and 7,000 reviews from the Internet (CAWI2). Reviews cover 3 areas: medicine (3,130 texts), hotels (2,938 texts), and other (936 texts). In this work, we will focus on the use of CAWI2 due to the evaluation of entire documents within the study.

In the CAWI2 study, each text received an average of 50 annotations. To obtain reliable results, the following cross-section of the population was used: 8,853 unique respondents were sampled from the Polish population. Sex, age, native language, place of residence, education level, marital status, employment status, political beliefs and income were controlled, among other factors.

The annotation schema was based on the procedures most widely used in NAWL (Riegel et al., 2015), NAWL BE (Wierzba et al., 2015) and plWordNet-emo (Zaśko-Zielińska et al., 2015; Janz et al., 2017; Kocoń et al., 2018; Kulisiewicz et al., 2015). Therefore, the acquired data consists of ten emotional categories: *valence*, *arousal*, and eight basic emotions: *sadness, anticipation, joy, fear, surprise, disgust, trust* and *anger*. Mean text rating distributions within emotional categories are presented in Figure 1. In total, 7k opinions * average of 53.46 annotators per opinion * 10 categories = 3.74M single annotations were collected.

The annotation process was carried out using the web-based system with an interface designed in collaboration with the team of psychologists to reduce as much as possible the difficulty of handling the annotation process and its impact on grades or their quality (see Figure 2). The collection resulting from the study is copyrighted and we got permission to conduct the research. A sample containing 100 texts with annotations and annotators' metadata with the source code of the experiments are publicly available on GitHub[2].

4 Personal Emotional Bias – PEB and agreement measures

In principle, we assume our collection (Internet review documents) is split into three partitions: *past* (D^{past}), *present*, and *future* (Figure 3). The past texts are used to estimate individual user beliefs and biases. The present documents allow us to train the reasoning model, whereas the future reviews are for the evaluation, test purposes.

To quantify individual subjective emotional perception of textual content, we introduce a new measure – *Personal Emotional Bias*, $PEB(u, c)$. It describes to what extent the previously known annotations $v_{c,d,u}$ of the given user u differ from the average annotations provided by all others for emotional category c, aggregated over all documents $d \in D^{past}$. Emotional category $c \in C$, where $C = \{sadness, anticipation, joy, fear, surprise, disgust, trust, anger, valence, arousal\}$. Integer values of the emotional annotations $v_{c,d,u}$ come from the study design, Figure 2, i.e. $v_{c,d,u} \in \{-3, -2, -1, 0, 1, 2, 3\}$, if $c = valence$ and $v_{c,d,u} \in \{0, 1, 2, 3, 4\}$ otherwise.

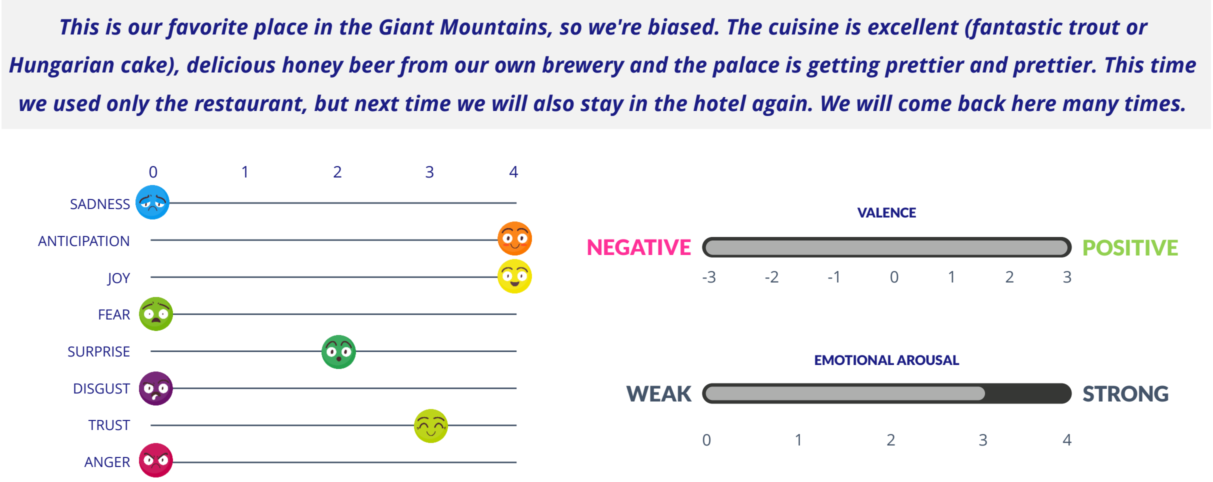
First, we need to compute the mean emotional value $\mu_{c,d}$ of each document $d \in D^{past}$ in each category c over all previously known d's annotations, i.e. provided by users from the train data, $u \in U_d^{train}$:

$$\mu_{c,d} = \frac{\sum_{u \in U_d^{train}} v_{c,d,u}}{|U_d^{train}|}, d \in D^{past}$$

In the next step, we calculate the standard deviation $\sigma_{c,d}$ of each emotional category c for each document d in a similar way:

[1] https://www.sentimenti.com/

[2] https://github.com/CLARIN-PL/personal-bias



This is our favorite place in the Giant Mountains, so we're biased. The cuisine is excellent (fantastic trout or Hungarian cake), delicious honey beer from our own brewery and the palace is getting prettier and prettier. This time we used only the restaurant, but next time we will also stay in the hotel again. We will come back here many times.

Figure 2: Emotional annotations for a real example of the hotel review – the CAWI study. Participants scored eight basic emotions (Plutchik model), arousal and valence on separate scales; varying from 0 to 4 for emotions and arousal and -3 to 3 for valence. Example review was manually translated from Polish to English.

$$\sigma_{c,d} = \sqrt{\frac{\sum_{u \in U_d^{train}} (v_{c,d,u} - \mu_{c,d})^2}{|U_d^{train}|}}, d \in D^{past}$$

Based on the above knowledge, we can estimate the Personal Emotional Bias $PEB(u, c)$ of the user u for the emotional category c. It is an aggregated Z-score, as follows:

$$PEB(u, c) = \frac{\sum_{d \in D_u^{past}} \frac{v_{c,d,u} - \mu_{c,d}}{\sigma_{c,d}}}{|D_u^{past}|}$$

where D_u^{past} is the set of documents $d \in D^{past}$ annotated by user u.

Please note that $PEB(u, c)$ may be calculated for any user, who provided their annotations to any document $d \in D^{past}$. It means that we can estimate PEB for users from the *dev* and *test* set, always aggregated over *past* documents. Nevertheless, components $\mu_{c,d}$ and $\sigma_{c,d}$ are fixed and computed only based on the previously known knowledge, i.e. for users from the *train* set. Obviously, the *train*, *dev*, and *test* sets are different for each out of ten cross-validation folds, which forces the recalculation of all PEB values at each fold.

The PEB measure provides us information about the unique views and preferences of the individual user. We suspect PEB to be more informative in the case of ambiguous texts with relatively low agreement among the annotators. To measure this agreement we leveraged two different document controversy measures: (1) the averaged Krippendorff's alpha coefficient α^{int} (Krippendorff, 2013) and (2) the general $contr^{std}$ controversy measure. The former is commonly used; it is resistant to missing annotations (Al Kuwatly et al., 2020; Wich et al., 2020a; Binns et al., 2017). According to our data, we used the variant of Krippendorff's alpha coefficient α^{int} with the interval difference function $\delta^{interval}(v_{c,d,u}, v_{c,d,u'})$ which calculates the distance between the two annotations $v_{c,d,u}$ and $v_{c,d,u'}$ for document d provided by two different users u and u' regarding emotional category c:

$$\delta^{interval}(v_{c,d,u}, v_{c,d,u'}) = (v_{c,d,u} - v_{c,d,u'})^2$$

Our first emotional controversy measure is expressed by the Krippendorff's alpha coefficient α_c^{int} separately calculated for the specified emotional category $c \in C$.

The alternative second measure $contr^{std}(d)$ was also used to analyze the controversial nature of any document d. It is the standard deviation of user ratings averaged over all emotional categories $c \in C$:

$$contr^{std}(d) = \frac{\sum_{c \in C} \sigma_{c,d}}{|C|}$$

5 Experimental plan, scenarios

All experiments were performed for two types of machine learning tasks, Figure 4:

- **Multi-task classification** - where each task was to predict an accurate discrete answer for each emotional category, i.e. one of the

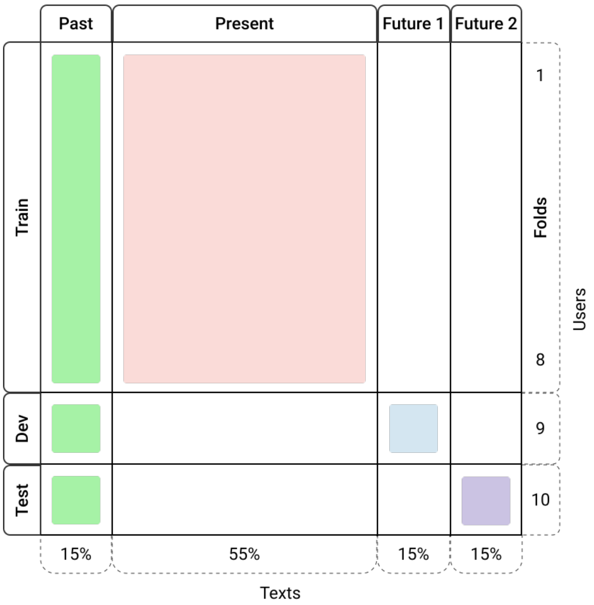


Figure 3: The CAWI2 collection was divided by the texts (columns) and the users/annotators (rows). The *past* texts (15% of all) were used to compute the PEB measure. The models were trained on 55% of the *present* texts and 80% of all users. They are verified with the *dev* set (disjoint from *train*) and tested on the *test* set - both containing 10% of users and 15% of texts each. The aforementioned proportions were chosen so that there were at least 1000 texts and more than 500 annotators in each section. The user-based split into *train*, *dev* and *test* is performed in the 10-fold cross-validation schema.

five classes {0, 1, 2, 3, 4} for eight emotions and arousal, and one out of seven classes for valence. Due to data imbalance ('0' was the dominating class for most emotions), the F1-macro measure was used to estimate the model performance;

- **Multivariate regression** - where the task was to estimate the numerical value of each emotional category. Such approach takes into account the distances between user ratings. R-squared measure was applied to compute the model quality.

In order to investigate the effect of PEB on emotion recognition for individual annotators, the following scenarios of the input data were considered:

- **AVG** - mean value of the annotation (regression) or most common class (classification) for all texts compared to the target values; this scenario is treated as initial baseline;

- **TXT** - text embeddings; it was the main baseline;

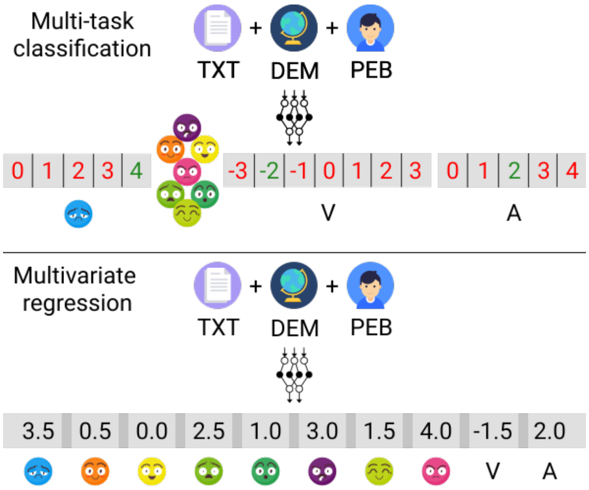


Figure 4: Two approaches to reasoning: (1) 10-task classification and (2) multivariate regression. In (1), the output contains 10 out of 52 classes. In (2), the output contains 10 real values, one for each emotional category. V – valence, A – emotional arousal.

- **TXT+DEM** - text embeddings and annotator demographic data;

- **TXT+PEB** - text embeddings and annotator's PEB;

- **ALL** - text embeddings, demographic data and PEB;

Additional **SIZE** scenario was performed to examine the impact of the number of annotated texts in PEB on the emotion recognition quality.

As a source of text embeddings the following models for Polish were used: (1) HerBERT, (2) XLM-RoBERTa, (3) fastText and (4) RoBERTa. The first one – HerBERT is currently considered state of the art according to the KLEJ benchmark (Rybak et al., 2020). Two neural network architectures were used to perform the experiments: (1) multi-layer perceptron (MLP) for transformer-based text embeddings (2) LSTM for fastText-based word embeddings (with 32 hidden units and a dropout equal to 0.5) with MLP to combine LSTM output with additional features. In both cases, the size of the input depends on the input embedding size. MLP output for classification is a multi-hot vector of length 52 (8 emotions x 5 possible ratings, 7 possible valence ratings, and 5 possible emotional arousal ratings), and for regression – a vector of size 10 containing real values ranging from 0 to 1 for each emotion dimension.

Ten fold cross-validation was applied as randomized non-overlapping partition of users and one division of texts, Figure 3. Such an approach is in line with leave-one-subject-out (LOSO) cross-validation where data is also split according to participants (subjects), i.e. data on one or more users are separated in the test set. Recently, it is commonly treated as SOTA approach in emotion recognition (Barlett et al., 1993; Schmidt et al., 2019)

In the SIZE scenario, we verified what incremental gain in model evaluation score we would achieve by increasing the number of texts in PEB (Figure 5 and Figure 6). The PEB measure denotes how much emotional perception of a given user differs from opinions of other users. To examine the significance of PEB for different emotional dimensions, we calculated the correlation between the PEB model results (R-squared) and the Krippendorff's alpha coefficient α_c^{int} for each emotional category $c \in C$.

To investigate the impact of PEB also for multiple languages, we translated Polish texts automatically into 8 languages using DeepL[3]. According to our manual tests and evaluation of translation quality, DeepL is characterized by better context matching of the target language utterances than other solutions available on the market. We applied the original annotations to the translated texts and then prepared dedicated models using XLM-RoBERTa. The training, test and validation sets were identical for all languages. The results are in Table 5 for classification and Table 6 for regression.

In order to verify the significance of differences between the evaluation results of each model in each scenario, we performed the independent samples t-test with the Bonferroni correction, as we tested more than two different models. We also checked the normality assumptions before its execution using Shapiro-Wilk test. If a sample did not meet them, we used the non-parametric Mann-Whitney U test.

6 Results

The results for all experimental scenarios and models, averaged collectively over ten folds are presented in Table 1 for classification and Table 2 for regression. The performance for each emotional category for all experimental variants for the best model (HerBERT), is specified in Table 3 for classification and Table 4 for regression. The results of multilingual model (XLM-RoBERTa) trained on sets translated into 8 languages can be seen in Table 5 for classification and Table 6 for regression.

	AVG	TXT	TXT+DEM	PEB	TXT+PEB	ALL
(1) HerBERT	5.97	**17.69**	**21.94**	32.02	**38.42**	**38.81**
(2) XLM-RoBERTa	5.97	**17.30**	**21.29**	31.91	**38.20**	**38.44**
(3) fastText+LSTM	5.97	16.48	20.52	32.09	37.25	**38.36**
(4) Polish RoBERTa	5.97	**17.01**	20.39	32.05	37.10	37.38

Table 1: Classification performance: F1-macro (%) averaged over ten folds. The best model for a specified scenario (column) is marked in **bold**; the best scenario for a given model (row) is underlined. More than one marked value means statistical insignificance between them.

	AVG	TXT	TXT+DEM	PEB	TXT+PEB	ALL
(1) HerBERT	-0.17	**13.16**	**14.37**	32.27	**45.96**	**45.64**
(2) XLM-RoBERTa	-0.17	**12.11**	**13.08**	32.24	**44.76**	**44.49**
(3) fastText+LSTM	-0.17	10.93	11.70	32.45	**43.74**	**43.50**
(4) Polish RoBERTa	-0.17	9.92	10.53	32.26	42.45	42.29

Table 2: Performance of regression models: R-squared averaged over folds. The best model in a given scenario (column) is in **bold**; the best scenario for a model (row) is underlined. More than one value highlighted means statistical insignificance between them.

Figure 5 presents R-squared results of reasoning for the TXT+PEB scenario and HerBERT model in relation to the number of texts from the *past* set used to estimate personal bias $PEB(u, c)$; averaged over all emotional categories and all users u. The past texts d annotated by user u are either randomly selected or starting from the most controversial, i.e. with the greatest $contrstd(d)$ value among all annotated by u in the past. The component results for each emotion and only for random selection are in Figure 6.

Figure 7 depicts the correlation between the annotation consistency counted using Krippendorff's alpha and the prediction performance in the regression task on the best model – HerBERT.

7 Discussion

The best results for each model were observed in the TXT+PEB scenario. The use of demographic data as additional user characteristics apart from the PEB measure in the ALL scenario did not provided significantly better results. HerBERT model achieved the best results, but differences between models are not statistically significant (except for the Polish RoBERTa).

The performance improvement related to demographic data about individual users was considered in the TXT+DEM scenario. Demographic features encode bias for social groups. However, once we have individual biases (the PEB measure), demographics becomes redundant and negatively affects

[3] https://www.deepl.com/

	AVG	TXT	TXT+DEM	PEB	TXT+PEB	ALL	std	α_c^{int}
sadness	6.28±0.18	16.47±0.90	21.91±1.08	29.93±2.12	37.85±1.26	37.68±0.94	1.18	0.18
anticipation	6.11±0.32	13.43±0.26	19.14±1.12	36.21±2.00	38.58±1.35	38.68±1.61	1.32	0.06
joy	5.64±0.26	20.58±1.36	25.58±1.22	30.69±1.65	39.13±1.24	39.62±1.74	1.28	0.24
fear	5.20±0.23	16.07±0.29	18.57±1.30	34.58±1.65	38.80±1.25	39.22±1.88	1.07	0.09
surprise	6.45±0.28	13.05±0.31	16.73±1.28	35.07±1.15	36.23±1.04	37.52±1.37	1.30	0.02
disgust	5.22±0.31	17.32±0.80	20.13±1.37	30.31±1.69	36.25±1.07	36.75±0.94	1.13	0.16
trust	5.36±0.27	17.11±0.76	22.71±1.43	30.02±1.45	37.07±1.00	38.94±1.56	1.26	0.19
anger	5.33±0.21	21.09±0.79	24.42±1.30	29.90±1.71	37.91±1.32	38.12±1.19	1.31	0.25
arousal	7.99±0.18	18.80±1.63	24.42±1.30	42.08±1.31	45.48±0.98	44.45±0.72	1.28	0.05
valence	6.10±0.21	23.00±1.42	25.75±1.12	21.45±1.39	36.89±0.82	37.15±1.26	1.58	0.38

Table 3: Classification performance – F1-macro for HerBERT model; last two columns are (1) aggregated standard deviation (std) and (2) Krippendorff's alpha coefficient α_c^{int}.

	AVG	TXT	TXT+DEM	PEB	TXT+PEB	ALL	std	α_c^{int}
sadness	-0.14±0.13	14.08±1.85	14.73±2.27	30.24±3.37	44.93±2.46	44.40±2.74	1.18	0.18
anticipation	-0.12±0.13	5.03±0.77	6.60±2.10	44.24±2.66	49.50±2.27	49.21±2.43	1.32	0.06
joy	-0.13±0.15	20.20±2.21	21.41±2.19	26.82±2.92	47.66±2.00	47.50±1.97	1.28	0.24
fear	-0.22±0.30	6.89±1.41	8.75±1.67	38.77±4.08	46.34±3.38	46.05±3.46	1.07	0.09
surprise	-0.14±0.17	1.00±0.55	2.82±2.62	43.20±2.75	44.96±2.58	44.42±2.72	1.30	0.02
disgust	-0.25±0.29	12.93±1.58	14.03±1.70	29.38±3.43	43.06±3.02	42.84±3.25	1.13	0.16
trust	-0.13±0.21	15.92±1.50	16.81±1.73	29.72±3.25	45.69±2.36	45.57±2.25	1.26	0.19
anger	-0.17±0.15	20.04±2.15	20.51±2.31	23.72±2.95	44.61±2.27	44.41±2.29	1.31	0.25
arousal	-0.20±0.21	3.05±1.10	4.70±1.28	47.30±1.98	50.87±1.52	50.37±1.70	1.28	0.05
valence	-0.16±0.13	32.44±2.75	33.35±2.56	9.32±2.22	41.98±1.61	41.68±1.49	1.58	0.38

Table 4: Regression performance – R-squared for HerBERT model; last two columns are (1) aggregated standard deviation (std) and (2) Krippendorff's alpha coefficient α_c^{int}.

	AVG	TXT	TXT+DEM	PEB	TXT+PEB	ALL
Dutch	5.97	17.44	20.83	32.03	37.88	38.24
English	5.97	17.47	21.19	32.20	37.75	38.32
French	5.97	17.13	21.08	32.23	37.48	38.19
German	5.97	17.13	21.04	32.14	37.85	38.13
Italian	5.97	17.12	20.84	31.73	37.66	38.24
Portuguese	5.97	17.35	21.03	31.99	37.70	38.29
Russian	5.97	17.23	21.30	32.32	37.75	38.27
Spanish	5.97	17.42	21.35	32.19	37.75	38.35

Table 5: Classification results (F1-macro, XLM-RoBERTa) for the texts translated into eight languages.

	AVG	TXT	TXT+DEM	PEB	TXT+PEB	ALL
Dutch	-0.17	11.76	12.75	32.29	44.41	44.11
English	-0.17	12.04	12.91	32.23	44.70	44.33
French	-0.17	11.79	12.67	32.26	44.44	44.13
German	-0.17	11.76	12.50	32.30	44.42	44.04
Italian	-0.17	11.69	12.75	32.20	44.39	44.11
Portuguese	-0.17	11.74	12.60	32.31	44.46	44.11
Russian	-0.17	11.74	12.33	32.22	44.35	44.07
Spanish	-0.17	11.79	12.66	32.26	44.43	44.08

Table 6: Regression results (R-squared, XLM-RoBERTa) for the texts translated into eight languages.

the results: compare TXT+PEB vs. ALL.

The PEB measure quantifies the difference in opinions of a particular user with respect to the others. In addition to beliefs, user decisions are also influenced by UI design. Several emotional categories could prove to be incomprehensible to individual users, so that their annotations do not reflect their opinions. Moreover, the scale of values could be misunderstood by some annotators who could mark the middle value when they were unsure whether a given emotional category was present in the analyzed text at all.

The use of simple statistical methods based on the averaged opinion about the text presented in the AVG scenario performs much worse than language models combined with MLP. Predicting the user's opinion solely upon the text in the TXT scenario (our baseline) results in poor performance. Therefore, there is a need to exploit personalized user data. The phenomenon of improving inference thanks to personalization is the same for each of the four considered models. It means that the proper personalization carried out at the stage of input data is much more important than the language model

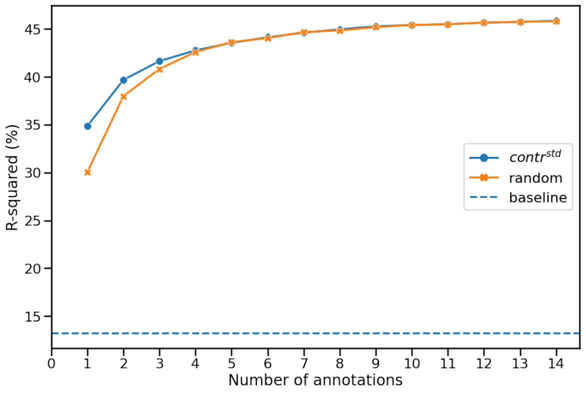


Figure 5: R-squared results on TXT+PEB scenario and HerBERT model in relation to the number of texts from the *past* set used to compute $PEB(u, c)$ values for a given user u, averaged over all emotional categories and all users. Two text selection procedures were considered: random and the most controversial – $contr^{std}(d)$. The baseline is the TXT scenario. The results for emotion categories and random selection are in Figure 6.

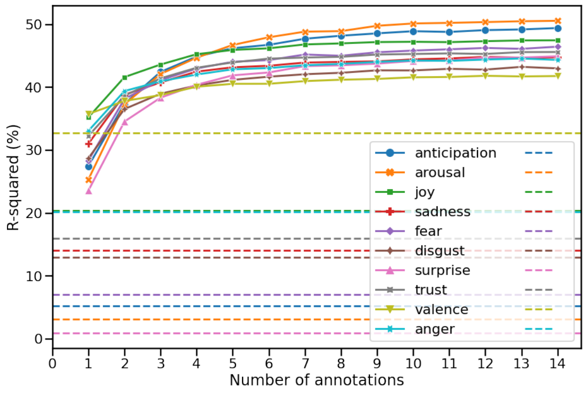


Figure 6: R-squared results on TXT+PEB scenario and HerBERT model in relation to the number of texts from the *past* set, randomly selected to compute $PEB(u, c)$ averaged over all users u – the solid lines. The dotted lines of the same color is the baseline for a given category (the TXT scenario).

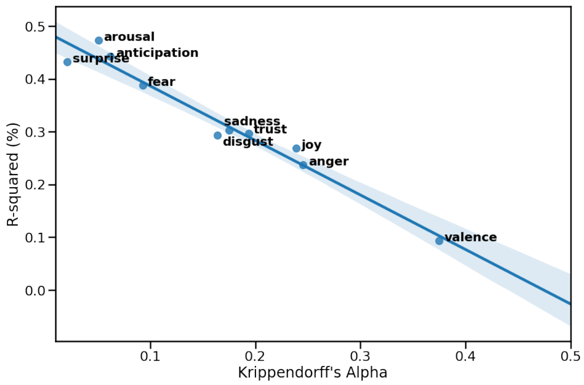


Figure 7: R-squared results on PEB scenario and HerBERT model in relation to Krippendorff's Alpha. Each data point corresponds to a separate emotional category from Table 4.

or inference model.

In the case of regression models, the complementary nature of the PEB measure and the text itself is clearly visible, see the PEB and TXT scenarios in Table 2, Table 4, and Table 6. This is manifested in a large number of cases in which a higher quality of inference from the text (TXT scenario) corresponds to the lower quality of the PEB-based inference (PEB scenario) and vice versa. In turn, their combination provides very good results. We calculated the correlation value for the results of evaluation over each emotional category and they are equal to -0.558 and -0,970 for the results in Table 3 and Table 4, respectively. We also analyzed the correlation between two values: (1) the sum of the results in the TXT and PEB scenarios and (2) the result in the TXT + PEB scenario. For the regression models, correlations are 0.999, 0.995, 0.896 for the results in Table 2, Table 4 and Table 6, respectively. In a similar way, we computed the correlation values for the results of the classification models; they reach: 0.802, 0.931, 0.257, for data from Table 1, Table 3 and Table 5, respectively.

The performance in the PEB scenario is the lowest for the valence category, which may result from the highest agreement level ($\alpha_c^{int} = 0.38$) and more flat distribution, Figure 1. Simultaneously, the reasoning based on text only (TXT scenario) demonstrated an opposite dependency: its performance is greatest for the highest agreement (valence) and lowest for low agreements (surprise, arousal and anticipation). It means that the more users disagree, to the greater extent we should rely on personal biases rather than solely on the textual content.

Even only one document annotated by a user utilized to estimate PEB can boost the reasoning, Figure 5. Moreover, only about 5-7 texts provided in the past are enough to capture the personal user beliefs. Later on, the gains are much smaller. This is valid for all emotional categories, Figure 6. The benefit is greater if PEB is computed for 1-3 most controversial texts ($contr^{std}$) annotated by a given user.

We have discovered a nearly linear negative correlation between annotators' agreement level (Krippendorff's alpha coefficient) and performance of the regression model based only on the personal bias (PEB), Figure 7.

8 Conclusions

Summarizing the experiments performed, we can draw several conclusions related to additional data that can be gathered during the annotation process. By means of them, we are able to significantly improve reasoning about emotional categories, i.e. prediction of emotions evoked by the given textual opinion in different people.

The most important conclusion is that the use of our proposed Personal Emotional Bias measure allows for a tremendous gain in prediction scores for the particular annotator. Thus, we have shown that using the current state-of-the-art methods for embedding texts and data from just a few annotations made by an individual user, we can infer the user's perception of emotions with much greater effectiveness. This opens up the possibility of creating dedicated and personalized solutions targeted at specific social groups and individuals we want to reach with a given message.

We have shown that demographic data of annotators have a positive impact on predicting their reactions, however not as much as the answers they provided during the survey itself. In addition, the combination of text content, demographic data and the single PEB feature built on the basis of their historical ratings is even several times better than the quality of responses given by the system based on text data alone.

Such a great influence on the outcome of single-individual data reveals a completely new direction. The NLP solutions should focus more on good design of the annotation process, its flow and single text-annotation sets rather than on post-processing and generalization of data, i.e. common class labels received by majority voting. The best proof of this thesis is the fact that we are able to successfully ignore the problem of annotator disagreement within a given text and fill in these gaps with human information.

In future work, we want to investigate the effect of individual PEB vector components on recognition quality. Additionally, we want to extend the PEB with information about the averaged annotation value of texts. Finally, the quality of dedicated models for individual emotional dimensions can be compared to the multi-task model presented in this work.

Acknowledgments

This work was supported by (1) the National Centre for Research and Development, Poland, grant no. POIR.01.01.01-00-0472/16 – *Sentimenti*; (2) the National Science Centre, Poland, project no. 2020/37/B/ST6/03806; (3) the Polish Ministry of Education and Science, CLARIN-PL Project; (4) 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.

References

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2020. Modeling annotator perspective and polarized opinions to improve hate speech detection. In *Proceedings of the Eighth AAAI Conference on Human Computation and Crowdsourcing*, pages 151–154.

Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 184–190, Online. Association for Computational Linguistics.

Nourah Alswaidan and Mohamed El Bachir Menai. 2020. A survey of state-of-the-art approaches for emotion recognition in text. *Knowledge and Information Systems*, pages 1–51.

Lora Aroyo and Chris Welty. 2013. Harnessing disagreement in crowdsourcing a relation extraction gold standard. Technical report, Technical Report.

M Barlett, G Littlewort, M Frank, C Lainscse, I Fasel, and J Movellan. 1993. Automatic recognition of spontaneous facial actions. *American Psychologist*, 48:384–392.

Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. 2017. Like trainer, like bot? inheritance of bias in algorithmic content moderation. *Social Informatics*, page 405–415.

Su Lin Blodgett and Brendan T. O'Connor. 2017. Racial disparity in natural language processing: A case study of social media african-american english. *ArXiv*, abs/1707.00061.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4356–4364, Red Hook, NY, USA. Curran Associates Inc.

H. Chou and C. Lee. 2019. Every rating matters: Joint learning of subjective labels and individual annotators for speech emotion classification. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5886–5890.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 67–73, New York, NY, USA. Association for Computing Machinery.

Paul Ekman and Wallace V Friesen. 1976. Measuring facial movement. *Environmental psychology and nonverbal behavior*, 1(1):56–75.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2012. A multitask approach to continuous five-dimensional affect sensing in natural speech. *ACM Trans. Interact. Intell. Syst.*, 2(1).

H. M. Fayek, M. Lech, and L. Cavedon. 2016. Modeling subjectiveness in emotion recognition with deep neural networks: Ensembles vs soft labels. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 566–570.

Yang Gao, Steffen Eger, Ilia Kuznetsov, Iryna Gurevych, and Yusuke Miyao. 2019. Does my rebuttal matter? insights from a major NLP conference. *CoRR*, abs/1903.11367.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166, Hong Kong, China. Association for Computational Linguistics.

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

Arkadiusz Janz, Jan Kocoń, Maciej Piasecki, and Zaśko-Zielińska Monika. 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.

Kamil Kanclerz, Alicja Figas, Marcin Gruza, Tomasz Kajdanowicz, Jan Kocoń, Daria Puchalska, and Przemyslaw Kazienko. 2021. Controversy and conformity: from generalized to personalized aggressiveness detection. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*. Association for Computational Linguistics.

Kamil Kanclerz, Piotr Miłkowski, and Jan Kocoń. 2020. Cross-lingual deep neural transfer learning in sentiment analysis. *Procedia Computer Science*, 176:128–137.

Yunus Emre Kara, Gaye Genc, Oya Aran, and Lale Akarun. 2015. Modeling annotator behaviors for crowd labeling. *Neurocomput.*, 160(C):141–156.

Jan Kocoń, Arkadiusz Janz, and Maciej Piasecki. 2018. Classifier-based polarity propagation in a wordnet. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Jan Kocoń and Marek Maziarz. 2021. Mapping wordnet onto human brain connectome in emotion processing and semantic similarity recognition. *Information Processing & Management*, 58(3):102530.

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

Jan Kocoń, Alicja Figas, Marcin Gruza, Daria Puchalska, Tomasz Kajdanowicz, and Przemysław Kazienko. 2021. Offensive, aggressive, and hate speech analysis: from data-centric to human-centred approach. *Information Processing & Management*.

Jan Kocoń, Arkadiusz Janz, Piotr Miłkowski, Monika Riegel, Małgorzata Wierzba, Artur Marchewka, Agnieszka Czoska, Damian Grimling, Barbara Konat, Konrad Juszczyk, Katarzyna Klessa, and Maciej Piasecki. 2019a. Recognition of emotions, valence 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 10^{th} International Global Wordnet Conference (GWC'19)*.

K. Krippendorff. 2013. *Content Analysis: An Introduction to Its Methodology*. SAGE Publications.

Marcin Kulisiewicz, Tomasz Kajdanowicz, Przemysław Kazienko, and Maciej Piasecki. 2015. On sentiment polarity assignment in the wordnet using loopy belief propagation. In *International Conference on Hybrid Artificial Intelligence Systems*, pages 451–462. Springer.

Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. 2009. Compositionality principle in recognition of fine-grained emotions from text. In *Third International AAAI Conference on Weblogs and Social Media*.

Robert Plutchik. 1982. A psychoevolutionary theory of emotions. *Social Science Information*, 21(4-5):529–553.

Vikas C. Raykar and Shipeng Yu. 2012. Eliminating spammers and ranking annotators for crowdsourced labeling tasks. *J. Mach. Learn. Res.*, 13(1):491–518.

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.

James A Russell and Albert Mehrabian. 1977. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11(3):273 – 294.

Piotr Rybak, Robert Mroczkowski, Janusz Tracz, and Ireneusz Gawlik. 2020. Klej: comprehensive benchmark for polish language understanding. *arXiv preprint arXiv:2005.00630*.

J. Salminen, F. Veronesi, H. Almerekhi, S. Jung, and B. J. Jansen. 2018. Online hate interpretation varies by country, but more by individual: A statistical analysis using crowdsourced ratings. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 88–94.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Philip Schmidt, Attila Reiss, Robert Dürichen, and Kristof Van Laerhoven. 2019. Wearable-based affect recognition—a review. *Sensors*, 19(19):4079.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii. Association for Computational Linguistics.

Guillermo Soberón, Lora Aroyo, Chris Welty, Oana Inel, Hui Lin, and Manfred Overmeen. 2013. Measuring crowd truth: Disagreement metrics combined with worker behavior filters. In *Proceedings of the 1st International Conference on Crowdsourcing the Semantic Web - Volume 1030*, CrowdSem'13, page 45–58. CEUR-WS.org.

S. Steidl, M. Levit, A. Batliner, E. Noth, and H. Niemann. 2005. "of all things the measure is man" automatic classification of emotions and inter-labeler consistency [speech-based emotion recognition]. In *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 1, pages I/317–I/320 Vol. 1.

Rachael Tatman. 2017. Gender and dialect bias in YouTube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, Valencia, Spain. Association for Computational Linguistics.

Zeerak Waseem. 2016. Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas. Association for Computational Linguistics.

Maximilian Wich, Hala Al Kuwatly, and Georg Groh. 2020a. Investigating annotator bias with a graph-based approach. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 191–199, Online. Association for Computational Linguistics.

Maximilian Wich, Jan Bauer, and Georg Groh. 2020b. Impact of politically biased data on hate speech classification. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 54–64, Online. Association for Computational Linguistics.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of Abusive Language: the Problem of Biased Datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 602–608, Minneapolis, Minnesota. Association for Computational Linguistics.

M. Wierzba, M. Riegel, M. Wypych, K. Jednorwóg, P. Turnau, A. Grabowska, and A. Marchewka. 2015. Basic emotions in the nencki affective word list (NAWL be): New method of classifying emotional stimuli. *PLoS ONE*, 10(7).

Michael Wojatzki, Tobias Horsmann, Darina Gold, and Torsten Zesch. 2018. Do women perceive hate differently: Examining the relationship between hate speech, gender, and agreement judgments. In *KONVENS*.

Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting racial bias in hate speech detection.

In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 7–14, Online. Association for Computational Linguistics.

Ashima Yadav and Dinesh Kumar Vishwakarma. 2020. Sentiment analysis using deep learning architectures: a review. *Artificial Intelligence Review*, 53(6):4335–4385.

Yi-Hsuan Yang, Ya-Fan Su, Yu-Ching Lin, and Homer H. Chen. 2007. Music emotion recognition: The role of individuality. In *Proceedings of the International Workshop on Human-Centered Multimedia*, HCM '07, page 13–22, New York, NY, USA. Association for Computing Machinery.

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.

Sicheng Zhao, Hongxun Yao, Yue Gao, Rongrong Ji, Wenlong Xie, Xiaolei Jiang, and Tat-Seng Chua. 2016. Predicting personalized emotion perceptions of social images. In *Proceedings of the 24th ACM International Conference on Multimedia*, MM '16, page 1385–1394, New York, NY, USA. Association for Computing Machinery.