Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022, pages 46–55
Marseille, 20 June 2022

StudEmo: A Non-aggregated Review Dataset for Personalized Emotion Recognition

Anh Ngo[1], Argi Candri[1], Teddy Ferdinan[1], Jan Kocoń[2], Wojciech Korczyński[2]
Wrocław University of Science and Technology, Wrocław, Poland
[1]{269588, 268894, 268893}@student.pwr.edu.pl
[2]{jan.kocon, wojciech.korczynski}@pwr.edu.pl

Abstract

Humans' emotional perception is subjective by nature, in which each individual could express different emotions regarding the same textual content. Existing datasets for emotion analysis commonly depend on a single ground truth per data sample, derived from majority voting or averaging the opinions of all annotators. In this paper, we introduce a new non-aggregated dataset, namely StudEmo, that contains 5,182 customer reviews, each annotated by 25 people with intensities of eight emotions from Plutchik's model, extended with valence and arousal. We also propose three personalized models that use not only textual content but also the individual human perspective, providing the model with different approaches to learning human representations. The experiments were carried out as a multitask classification on two datasets: our StudEmo dataset and GoEmotions dataset, which contains 28 emotional categories. The proposed personalized methods significantly improve prediction results, especially for emotions that have low inter-annotator agreement.

Keywords: emotion recognition, personalization, non-aggregated dataset, learning human representation

1. Introduction

Emotions play an essential role in human communication. We can observe an increasingly high demand in studies of emotion recognition within natural language processing (NLP) due to its applicability in multiple domains. Emotion perception is naturally subjective and varies regarding each individual due to the differences in personal backgrounds, such as culture, gender, and age, which leads to the problem of low inter-annotator agreement in the existing datasets.

Recent studies have shown that different reviewers may classify the same object differently, but that unnecessarily means they're wrong, as they merely have different sentiments about the same thing (Basile et al., 2021). Those studies also identified the increased demand for new datasets related to personal perspectives on subjective NLP tasks. However, almost all available datasets for emotion recognition provide only limited information on the annotators. Moreover, to solve the problem of low inter-annotator agreement, only a few of them retain multiple annotations per sample. One of the most popular approaches is to use majority voting to obtain a single ground truth for each data sample. Another common approach is to collect the annotation from experts. Both methods consider only one correct label for a given text sample.

Existing solutions for emotion recognition do not consider involving individual perspectives, which rely on using only one ground truth to train the emotion recognizer. In addition, current personalization approaches include human representation generated from personal characteristics. However, these methods do not take into account the relationship between each annotator and the specific features of the text.

In this work, we introduce StudEmo, a non-aggregated dataset of 5,182 customer reviews in English, labeled for eight basic emotions from Plutchik's model, along with valence and arousal. Our dataset provides the annotations from 25 unique annotators who are students from different countries with different cultures, ages, and characteristics. The annotation strategy followed the procedures proposed by Janz et al. (2017) and Zaśko-Zielińska and Piasecki (2018).

Additionally, we propose personalized methods for emotion recognition tasks on textual data that take into account both textual content and how the raters react to that content. The approach is inspired by the idea of involving personal human bias and representation (Kocoń et al., 2021b), which is based on optimizing a multidimensional latent vector that represents the perspective of each annotator in a targeted text. Here, we propose extensions to these models by finetuning the entire architecture, which yields a significant quality improvement over the methods presented in (Kocoń et al., 2021b).

2. Related Work

Recent studies have highlighted the advantages of integrating the opinions and perspectives of individual annotators involved in subjective NLP tasks (Basile et al., 2021; Kocoń et al., 2021b). However, most current methods do not consider involving multiple annotator perspectives, in which neural network models such as CNN, Bi-LSTM, GRU (Abdullah et al., 2018) are combined with a separate model to extract text embeddings, such as transformer-based (Ghosh and Kumar, 2021; Chiorrini et al., 2021; Wang and Tong, 2021); GloVE, and ELMo (Lee et al., 2020). Akhtar et al. (2020a) proposed a stacked ensemble architecture for the recog-

nition of the intensity of emotions, while Li and Xu (2014) involved emotional causes extracted from expert knowledge.

Dealing with tasks related to subjectivity in text perception is difficult due to the high variability in different points of view. One of the common approaches to representing multiple annotators without losing individual perspectives is to utilize a multitask or ensemble architecture that treats predicting annotator decisions as separate subtasks (Fayek et al., 2016; Davani et al., 2022). Another idea is to use the attention mechanism to introduce human representation, which considers personal characteristics, into emotion modeling. Although Li and Lee (2019) used the feature *Linguistic Inquiry Word Count* to create personal profile embeddings, a valuable idea was presented in (Kamran et al., 2021), where the authors demonstrated the correlation between personal cognitive factors and emotions from textual data. Furthermore, Akhtar et al. (2020b) considered a group-based personalized method and tried to maximize the polarity index between two groups.

The problem of the scarcity of non-aggregated datasets is discussed by Basile et al. (2021), since most current datasets for emotion recognition are aggregated by majority voting, best-worst scaling (Mohammad and Bravo-Marquez, 2017), or using a hybrid rule-based automated system (Krommyda et al., 2021). As mentioned by Hernandez et al. (2021) collecting high-quality emotional data is difficult and expensive which limits the availability of generalizable data. Only a few non-aggregated datasets exist, such as Measuring Hate Speech (Kennedy et al., 2020), Offensive Language Datasets with Annotators' Disagreement (Leonardelli et al., 2021). Specifically, we have found only three datasets for the emotion recognition task that preserve each annotator's opinions without combining them, including GoEmotions Datasets (Demszky et al., 2020), Emotion Meanings dataset (Wierzba et al., 2021), and Sentimenti database (Kocoń et al., 2019).

3. Datasets

3.1. StudEmo Dataset

Our dataset consists of 5,182 reviews in English. It is available on the DSpace CLARIN-PL repository under a CC BY-NC-ND 4.0 license[1]. These reviews were acquired from the MultiEmo dataset (Kocoń et al., 2021), which is a benchmark dataset for multilingual sentiment analysis containing consumer reviews from four different domains: hotels, medicine, products, and university. Since the original texts were written in Polish, the translation to English was performed using DeepL which is a translation system based on deep neural networks. The tool's producers present it as the best existing translation system[2]. Its superiority or similar performance in comparison to other existing tools, e.g, Google Translate, is proved by some recent studies: (Cambedda et al., 2021), (Hidalgo-Ternero, 2021) and (Bellés-Calvera and Quintana, 2021).

It is not easy to determine if emotions and sentiment are preserved after translation. Nevertheless, sentiment classification results for the original and translated texts (Kocoń et al., 2021) are very similar what suggests that translation quality is good enough to express similar sentiment.

The texts are annotated by 25 unique English-speaking annotators who are international students from different countries and cultural backgrounds studying at the master's degree level. They were not remunerated, annotators were only graded based on number of annotations during one of the study tutorials. The annotation schema was based on the procedures used in (Janz et al., 2017; Zaśko-Zielińska and Piasecki, 2018). Each annotator received a subset of 400 reviews and was asked to annotate it according to their own personal emotional reaction to the given text. Each annotator was allowed to annotate a given text with multiple emotion labels.

The resulting annotated data consist of ten categories: eight basic emotion categories from Plutchik's Wheel of Emotions: joy, trust, anticipation surprise, fear, sadness, anger, and disgust. Two additional dimensions were valence and arousal. Each basic emotion category and arousal has an intensity range of [0,3]. Meanwhile, valence has a range of [-3,+3]. Finally, a total of 7,463 annotations were acquired. The average annotation distribution for each basic emotion category is shown in Figure 1.

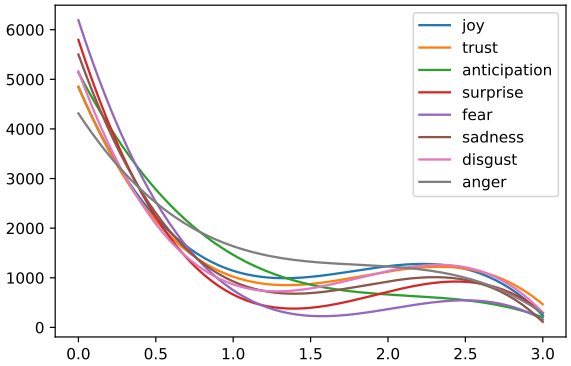

Figure 1: Data distribution of basic emotions in the StudEmo dataset. The x-axis is the intensity levels of emotions, while the y-axis is the number of annotations.

Of the 5,182 texts in the dataset, 2,901 were annotated by one annotator, and 2,281 were annotated by two annotators. There are 1,701 texts in which both annotators agree on the existence of at least one emotion category regardless of the intensity level. If the intensity level is considered, there are 1,011 texts where both

[1] http://hdl.handle.net/11321/895

[2] https://www.deepl.com/en/blog/20200206

annotators agree on at least one emotion category with the same intensity level.

On texts that received two annotations, the inter-annotator agreement was measured using the weighted Cohen's kappa coefficient to take into account the degree of disagreement, as the intensity levels in each category are ordered. The average weighted Cohen's kappa is 0.26. The weighted Cohen's kappa value for each category is as follows: Joy 0.33, Trust 0.33, Anticipation 0.22, Surprise 0.09, Fear 0.08, Sadness 0.21, Disgust 0.25, Anger 0.40, Valence 0.52, Arousal 0.12.

3.2. GoEmotions Dataset

The GoEmotions dataset from (Demszky et al., 2020) consists of 58,011 texts with 28 labels (27 emotion categories and 1 neutral category). The texts were carefully selected from Reddit. Each emotion category only has two possible values, 0 or 1. However, the texts are multi-labeled so that a given text may be annotated with more than one emotion category.

The texts were annotated by 82 unique annotators, each of them having 1-5 annotations. A total of 211,225 annotations are available; the average annotation distribution for each emotion category is shown on Figure 2.

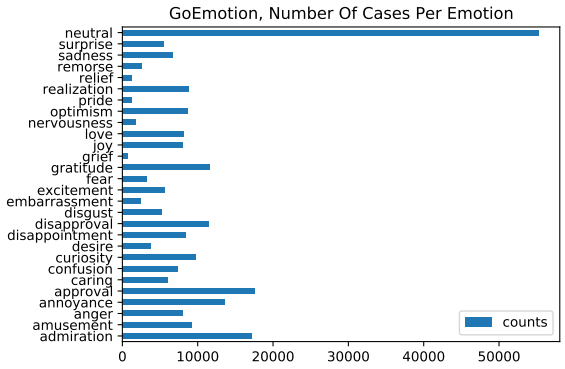

Figure 2: Data distribution of emotion categories in the GoEmotions dataset. The x-axis is the emotion categories, while the y-axis is the number of annotations.

The inter-annotator agreement in this dataset is somewhat high. There are 54,263 (94%) texts in which two or more annotators agree on at least one emotion category. However, there are only 17,763 (31%) texts in which three or more annotators agree on at least one emotion category. One reason for the relatively high inter-annotator agreement is that this dataset does not consider the intensity levels of the emotions, only their presence.

4. Dataset Splitting

Our dataset splitting strategy is based on (Miłkowski et al., 2021) and is depicted by Figure 3. The dataset was divided into columns (texts) and rows (annotators/users). The dataset was then partitioned with respect to the *texts* axis into Past (15%), Present (55%), Future1 (15%), and Future2 (15%). Meanwhile, the user-based split into the train, dev, and test sets was performed with the 10-fold cross-validation schema. All of the annotators/users are seen, which means that the models already learned all users before making the predictions.

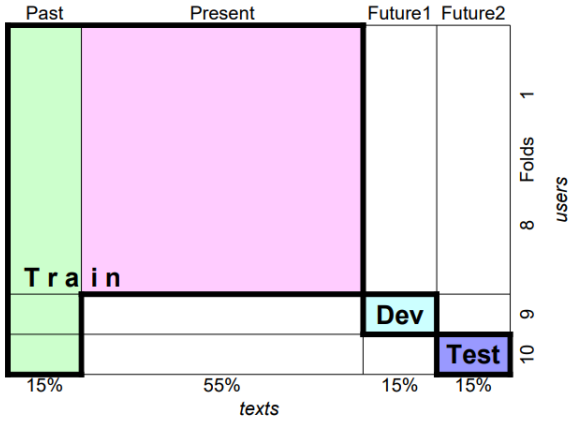

Figure 3: Dataset splitting strategy.

The dataset is split into *Past*, *Present*, and *Future* partitions to simulate data that is available in a working emotion prediction system. We assume the *Past* partition as texts that users have previously annotated (i.e. when these users started using the system, they were asked to annotate several texts for the purpose of calibrating the system); this *Past* partition is used to estimate individual user beliefs and biases. The *Present* partition represents texts and annotations that come up during the usage of the emotion prediction system, and it allows us to train the reasoning model. The *Future* partition is used for evaluation and test purposes.

The models were trained on the *Past* partition of 100% of all users and the *Present* partition of 80% of all users. In the case of personalization methods, the *Past* partition signifies some background knowledge about the users, and it was used to calculate the Human Bias (HuBi) measure of each user. On the other hand, *Present* partition signifies the general view of the texts and was used to train the reasoning of a personalized model. In the case of the baseline methods, both the *Past* and *Present* partitions were used for training but without considering the biases of the users.

The models were validated with the *dev* split, which uses the *Future1* partition of a different user fold. Therefore, *dev* contains about 10% of all users and 15% of all texts. It is important to note that *dev* is disjoint from *train*, which means that the models were validated on annotations never seen before.

The models were tested with the *test* split, which uses the *Future2* partition of yet another different user fold. Hence, the *test* split also contains about 10% of

all users and 15% of all texts. Similarly, *test* is disjoint from *train* and *dev*, which means that the models were tested on annotations never seen before during training or validation.

5. Models

With the objective of emotion recognition based on individual's perspectives, we decided to exploit four different sources of information about annotators and text, including embedding of the considered text, user id, embeddings of annotated texts with annotations, and human bias. The text embeddings are generated by the pre-trained Transformer language model, in which the parameters are either finetuned or frozen during the training process. We started with the two variants, AVG-ANN and SINGLE-ANN, of the baseline models, which used only text embeddings as input. Next, we proposed and compared three new deep learning architectures that utilized the annotator's information, including the following:

1. User-ID – modeling the user id as a special token in text embedding,
2. Past-Embedding – the model uses embeddings of a few texts from Past split with user annotations,
3. HuBi-medium – the model using learned human embedding and word biases.

5.1. AVG-ANN Baseline

The AVG-ANN Baseline model adapts a simple approach in which it receives the evaluated text embeddings as input and compares the mean value of annotations of all texts to the target values. The method is similar to the majority voting calculation in which the annotations are also aggregated.

5.2. SINGLE-ANN Baseline

The SINGLE-ANN Baseline model implements a commonly investigated approach known in NLP with one unified output for all users. The model receives the evaluated text embeddings as input and trained on each users annotation.

5.3. User-ID

User-ID is a simpler personalization approach that is adapted from (Kocoń et al., 2021a). This approach was briefly mentioned in Dudy et al. (2021), in which it was argued that user-level personalization on language models can be done by conditioning textual generation on different users. With the User-ID approach, the annotator was simply represented as a one-hot vector that was concatenated to the text embeddings. However, one potential issue with that approach is that the dimension of the vector can become quite large with an increase in the number of annotators. Hence, in User-ID method, the annotator is represented by a special token that is added to the text embedding; and in the case of BERT, the special token gets its own embedding.

5.4. Past-Embedding

In Past-Embedding model, personalization is ensured by adding an extra input composed of embeddings of a few texts from Past split along with their annotations given by a user. It is an adaptation of the *Class-based* model from (Kanclerz et al., 2021). These embeddings and annotations form a vector that constitutes a representation of the user beliefs. It is concatenated with an embedding of a currently processed text. This concatenation forms an input to the final classification layer. Embeddings of annotated past texts come from frozen pre-trained language model.

5.5. HuBi-medium: Learned Human Embedding Model

HuBi-medium model is derived from the approach introduced by Kocoń et al. (2021b), in which the multi-dimensional latent vector of an annotator is optimized for multi-dimensional modeling user subjectivity. This approach is based on the concept of Neural Collaborative Filtering (NFC) in recommender systems (He et al., 2017). A typical issue when directly applying NFC to personal perspective modeling is a cold start, which is a consequence of the small number of annotations assigned for each text, making it difficult to obtain a good representation from scratch. To deal with this problem, we propose an alternative hybrid model that utilizes text representations from language models and optimizes only the annotator's latent vector. Figure 4 illustrates the HuBi-medium architecture to capture the relationship between the annotator and the targeted text, in which the product of element-wise multiplication between the annotator embedding and the text embedding is passed on to the fully connected layer for the final prediction. The prediction is defined as follows:

$$y(t,a) = W_{TU}(a(W_T x_t) \otimes a(W_U x_u)) + \sum_{word \in t} b_{word}$$

where t and u: evaluated text and user; b: a vector of biases indexed with words; x_t, x_u: text embedding of the evaluated text t and embedding of user u, respectively; W_{TA}, W_T, W_A: weights of the fully-connected layers; a: the activation function.

6. Experimental Setup

We formulated all experiments as a multitask classification, in which each task was to predict an accurate label for each emotional category, including one over four classes for arousal and eight emotion types, and one over seven classes for valence. To handle the class imbalance problem, in which other labels are dominated by label '0', the macro F1-score was used for model evaluation. The 10-fold cross-validation is applied to randomly divide the dataset into 10 subsets of the same size.

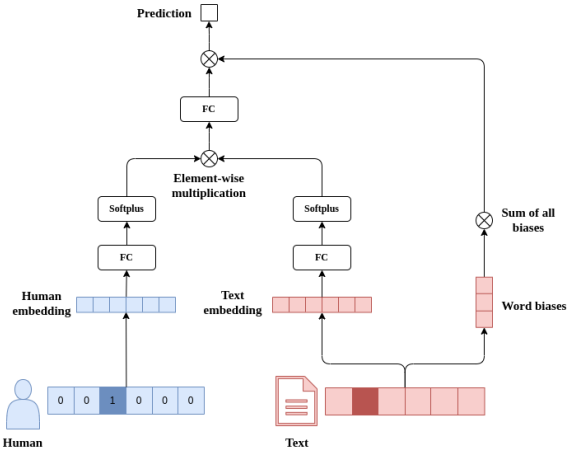

Figure 4: HuBi-medium: learned human embedding model architecture.

6.1. Language Models

A proposed architectures utilize RoBERTa (Liu et al., 2019), a Transformer-based language model, to obtain a representation of text. RoBERTa is an extension of BERT (Devlin et al., 2019) with additional key modifications introduced above BERT's pretraining procedure, including removing the next sentence prediction objective and changing the masking pattern applied to training data dynamically.

All experiments were performed on both the *original* RoBERTa model (*non-finetuned*) and the *finetuned* model. In the non-finetuned scenario, the text embeddings are generated from the pre-trained Transformer's RoBERTa, while in finetuning, the entire pre-trained model was unfrozen, and the entire pre-trained weights are updated during further training on our dataset.

6.2. Hyperparameter Settings

For both scenarios (non-finetuned and finetuned), the optimal values for hyperparameters were obtained for each model separately, in which the optimal learning rate for two baselines was 5e-5, for both User-ID and HuBi-medium was 3e-5, and 1e-5 for Past-Embedding. We used the Adam optimizer and cross-entropy as a loss function. For the *finetuning* scenario, the weight decay was 0.01, and we used the learning rate schedule with a warm-up proportion of 0.1. All models were trained for 20 epochs in both training scenarios, except for finetuned Past-Embedding, where the trained epochs were 10.

In addition, Past-Embedding requires the parameter to control the number of texts in the annotator's past embedding, which is equal to 4. Since the HuBi-medium model extends the standard architecture with human embedding, it requires additional hyperparameters, including the annotator embedding size of 50 and the hidden size of 100 for the classifier's last fully connected layer. A dropout layer with a rate of 0.25 was added to prevent overfitting.

Similar experiments were performed on the GoEmotions dataset, in which we utilized the same parameters, except for learning rates and the number of trained epochs. While the learning rate for both baseline models and HuBi-medium was 3e-5, User-ID was trained with a learning rate of 1e-3, in both non-finetuning and finetuning scenarios. For Past-Embedding, they are 1e-4 and 1e-5, respectively. The epoch number was 10, since it preserves a stable learning curve for all models, except for Past-Embedding, in which we employed 20 epochs without finetuning and 5 epochs on finetuning.

6.3. Statistical Testing

To determine the significance of the differences found in the models' results, statistical tests are performed. The normality of the distribution of the results is checked using Q-Q plots and Shapiro-Wilk test with significance level $\alpha = 0.05$. Depending on that, an appropriate statistical hypothesis test is chosen.

For data with a normal distribution, *independent samples t-test* is used. Since the results are acquired from different models, the assumption that the groups are independent is fulfilled. The homogeneity of the variance is tested using the Levene test. In case the data do not have homogeneous variances, the independent samples t-test is performed using the Welch-Satterthwaite adjusted method.

The independent samples t-test is performed with $\alpha = 0.1$ on results for each emotional category. If $p\_value > \alpha$, then we cannot reject the null hypothesis, which means that there is no significant difference between the results of the two models. If $p\_value \leq \alpha$, then the null hypothesis is rejected, which means that there is a significant difference between the results of the two models.

7. Results

The results of the StudEmo experiment scenarios for each emotional category are presented in Table 1-p.6 for the *non-finetuning* scenario and Table 2-p.6 for the *finetuning* scenario. The results of GoEmotions experiment scenarios for each emotional category are presented in Table 3-p.7 for the *non-finetuning* scenario and in Table 4-p.7 for the *finetuning* scenario. Figure 5-p.6 presents boxplots of the averaged macro F1-scores among all categories for all experiment scenarios on the StudEmo dataset. The analogous plot for the GoEmotion experiments is shown in Figure 6-p.7.

Generally, the differences in results for HuBi-medium and Past-Embedding are not significant in most cases. For StudEmo dataset, the latter achieved slightly better results and vice-versa for GoEmotions dataset. However, the difference is not drastic, only about 1.3 - 1.8 pp, which shows the stability in the performance of Past-Embedding. The only exception is observed in non-finetuned models on StudEmo, in which HuBi-medium is 8.4% behind Past-Embedding. This phenomenon may arise because the HuBi medium

benefits more from finetuning and a larger dataset. For larger datasets like GoEmotions, Past-Embedding also took advantage of finetuning considerably much higher than with small datasets like StudEmo, in which there is no significant difference between the two strategies.

7.1. StudEmo

Overall, the best results were obtained for the Past-Embedding method in both non-finetuned and finetuned scenarios, with the mean macro F1-score of 34.4% and 34.3%, respectively (Table 1, Table 2). Statistical tests reveal no statistical significance between these two scenarios, which shows that Past-Embedding does not benefit from finetuning. Additionally, Figure 5 shows that the finetuned Past-Embedding results have a broader range than the original accompanying slightly positive skewing and outliers. It indicates a larger data dispersion and instability for the finetuned variant of Past Embedding.

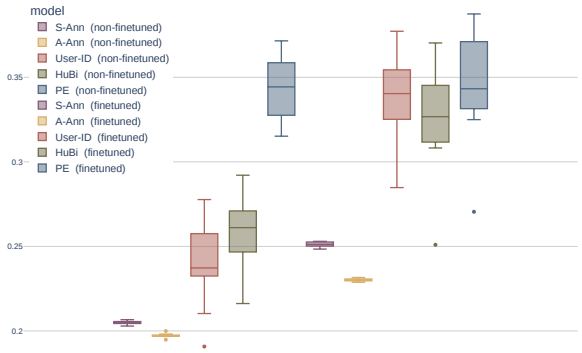

Figure 5: Test macro F1 mean results from non-finetuned and finetuned models, run on StudEmo dataset. S-Ann: SINGLE-ANN, A-Ann: AVG-ANN, HuBi: HuBi-medium, PE: Past-Embedding.

Emotions	S-Ann	A-Ann	User-ID	HuBi	PE
anger	19.1%	19.5%	20.7%	29.4%	**40.7%**
anticipation	20.3%	20.6%	21.3%	22.4%	**29.8%**
arousal	20.7%	19.5%	23.5%	26.6%	**37.8%**
disgust	20.9%	20.9%	20.5%	21.7%	**30.9%**
fear	22.7%	23.0%	29.8%	29.8%	**33.6%**
joy	19.5%	19.6%	20.9%	23.4%	**34.1%**
sadness	21.5%	21.7%	24.5%	24.5%	**30.0%**
surprise	21.9%	21.8%	21.6%	21.6%	**24.6%**
trust	19.4%	19.2%	19.7%	21.0%	**33.6%**
valence	18.8%	11.5%	36.5%	39.1%	**49.5%**
Mean	20.5%	19.7%	23.9%	26.0%	**34.4%**

Table 1: Test macro F1 results for models in *non-finetuned* scenario run on StudEmo dataset. S-Ann: SINGLE-ANN, A-Ann: AVG-ANN, HuBi: HuBi-medium, PE: Past-Embedding. The best model for a specified emotion is marked in bold.

For the *non-finetuned* scenario, there are remarkable differences between the three personalized models. The gap between the best (Past-Embedding) and the worst (User-ID) is approximately 10.5 pp. HuBi-medium with 26% of macro F1-score on average is situated between them. Table 1 demonstrates that Past-Embedding and HuBi-medium outperformed User-ID on all emotions, except fear, sadness, and surprise, for which HuBi-medium and User-ID resulted similarly.

Emotions	S-Ann	A-Ann	User-ID	HuBi	PE
anger	30.9%	24.0%	**45.2%**	43.9%	44.3%
anticipation	23.4%	20.9%	**29.2%**	26.9%	28.8%
arousal	27.2%	28.9%	29.1%	27.5%	**30.2%**
disgust	22.7%	20.9%	**31.8%**	30.0%	30.9%
fear	22.7%	23.0%	28.2%	**29.8%**	**29.8%**
joy	25.9%	21.9%	36.9%	35.5%	**39.3%**
sadness	21.5%	21.7%	28.4%	24.6%	**29.6%**
surprise	**21.9%**	21.8%	21.6%	21.6%	21.6%
trust	22.3%	19.7%	**43.0%**	38.8%	40.0%
valence	32.5%	27.4%	46.6%	46.1%	**48.6%**
Mean	25.1%	23.0%	34.0%	32.5%	**34.3%**

Table 2: Test macro F1 results for models in *finetuned* scenario run on StudEmo dataset. S-Ann: SINGLE-ANN, A-Ann: AVG-ANN, HuBi: HuBi-medium, PE: Past-Embedding. The best model for a specified emotion is marked in bold.

In contrast, an interesting phenomenon was observed for the *finetuned* scenario, in which both User-ID and HuBi-medium took advantage of finetuning. User-ID achieved 34% of macro F1-score on average, which is 10.1 pp higher than the non-finetuned User-ID and only 0.3 pp lower than Past-Embedding, followed by HuBi-medium, which increased from 26% to 32.5%. Furthermore, statistical tests showed almost no significance in the differences between these three finetuned personalized models, indicating that they are all comparable.

However, Figure 5 exhibits a moderately wide range in User-ID's macro F1-score distribution compared to the other personalized models, implying a broader dispersion of predictions. In terms of that comparison, Past-Embedding, and HuBi-medium have shown more stable and less scattered predictions.

Detailed studies of the results for particular emotions demonstrate some differences among personalized methods, even though they are relatively comparable on average. Past-Embedding outperformed the other models in predicting four emotions, including arousal, joy, sadness, and valence. Meanwhile, User-ID achieved the best results in predicting anger, anticipation, disgust, and trust. Both HuBi-medium and Past-Embedding got the same score on *fear*. Exceptionally, the best result for *surprise* came from the SINGLE-ANN baseline with 21.9%, while all personalized methods got slightly lower at 21.6%. Interestingly, except for the original Past-Embedding, which achieved 24.6% for *surprise*, all other experiments got almost identical results of approximately 21% on that emotion. The high imbalance of classes distribution for that emotion (value 0 is nearly 20 times more frequent

than value 3), together with the low annotator agreement of 0.09 on the Cohen's kappa coefficient, could be the reason to explain this phenomenon. A similar case can also be be seen for *fear*, in which all the non-finetuned and finetuned baselines resulted in the same score of 23%, while finetuned HuBi-medium could achieve 29.8%, and the non-finetuned Past-Embedding obtained 33.6%. *Fear* is also a contentious emotion that got only 0.08 of Cohen's kappa coefficient and was affected by a high imbalance. These phenomena strengthen the benefits of the personalized methods on high-controversial emotions, such as *fear* and *surprise*.

The highest results were obtained for *valence* (49.5% with non-finetuned Past-Embedding method), *anger* (45.2% with finetuned User-ID), and *trust* (43% with finetuned User-ID). The Cohen's kappa coefficient for *valence* is relatively high and equals 0.52, which can explain the higher performance. However, in contrast to personalized approaches, without finetuning, two baselines performed the worst on valence, especially the AVG-ANN, which got to the bottom at 11.5% on predicting valence (Table 1). It demonstrates that even for less controversial emotions, the application of the personalized methods give performance gain.

7.2. GoEmotions

In the case of GoEmotions, the best performing baseline model is the finetuned AVG-ANN, with an average macro F1-score of 50.9%. Meanwhile, the best personalized model is the finetuned HuBi-medium, with an average macro F1-score of 66.1%. In Figure 6, we can see that the best personalized model outperformed the best baseline in both non-finetuned and finetuned scenarios. Statistical testing also proved that the differences between the best personalized model and the best baseline are statistically significant.

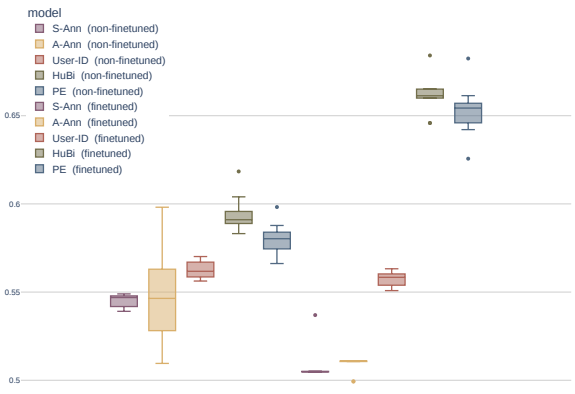

Figure 6: Test macro F1 mean results from non-finetuned and finetuned models, run on GoEmotions dataset. S-Ann: SINGLE-ANN, A-Ann: AVG-ANN, HuBi: HuBi-medium, PE: Past-Embedding.

The non-finetuned baseline AVG-ANN model exhibited an interesting behavior, which can be seen in Table 3., and *relief* emotions. For *nervousness*, *pride*, *remorse*, *desire*, and *grief* it obtained macro F1-score median of around 0.49, but outliers could reach a macro F1-score of 1.0. In the case of *grief* and *relief*, the macro F1-score median was 1.0 with a mean value of about 0.8, yet the distribution was very wide. Consequently, the mean of the macro F1-score of the non-finetuned AVG-ANN became abnormally high in these emotions.

Emotions	S-Ann	A-Ann	User-ID	HuBi	PE
desire	52.7%	55.0%	54.9%	**59.4%**	56.3%
nervousness	49.8%	**65.0%**	50.7%	50.3%	50.9%
pride	49.9%	**65.0%**	51.1%	52.0%	52.3%
remorse	57.8%	55.0%	59.9%	61.8%	**63.3%**
grief	50.3%	**85.0%**	55.2%	51.6%	55.9%
relief	49.8%	**80.0%**	50.4%	50.2%	50.4%
gratitude	82.2%	78.8%	84.1%	**87.9%**	83.5%
fear	54.5%	49.9%	59.1%	**60.6%**	59.1%
embarrassment	49.7%	50.0%	50.2%	50.0%	**50.4%**
joy	52.0%	49.9%	54.7%	**59.2%**	55.5%
disgust	51.0%	49.9%	52.6%	**55.2%**	52.9%
sadness	53.9%	49.9%	55.9%	**58.8%**	56.2%
surprise	52.5%	49.9%	54.9%	**56.7%**	55.0%
Mean	54.5%	54.7%	56.3%	**59.4%**	58.1%

Table 3: Test macro F1 results for models in *non-finetuned* scenario run on GoEmotions dataset. S-Ann: SINGLE-ANN, A-Ann: AVG-ANN, HuBi: HuBi-medium, PE: Past-Embedding. The best model for a specified emotion is marked in bold.

Furthermore, it was found that the model was never detecting the considered emotions; it always predicts the lack of these emotions. Meanwhile, there is a remarkably high imbalance in these categories (class *0* is present 312 times more frequently than class *1*). These emotion categories are so rare that they are not available in some test folds, giving a perfect F1-score even though the model was always predicting zero. However, the finetuned baseline AVG-ANN showed a much more stable behavior and greatly reduced outliers.

Emotions	S-Ann	A-Ann	User-ID	HuBi	PE
desire	49.7%	50.0%	54.3%	63.9%	**65.2%**
nervousness	49.8%	50.0%	50.7%	**55.5%**	53.6%
pride	49.8%	50.0%	51.5%	**58.8%**	52.3%
remorse	50.6%	50.0%	59.7%	67.8%	**73.1%**
grief	49.9%	50.0%	53.8%	**57.2%**	50.5%
relief	49.8%	50.0%	50.5%	**55.8%**	51.2%
gratitude	74.9%	74.0%	84.1%	89.8%	**90.3%**
fear	49.7%	49.9%	58.2%	72.8%	**73.7%**
embarrassment	49.7%	50.0%	50.0%	61.9%	**62.1%**
joy	49.2%	49.9%	54.2%	64.5%	**64.9%**
disgust	49.4%	49.9%	51.4%	**63.5%**	61.8%
sadness	49.6%	49.9%	55.3%	66.9%	**68.7%**
surprise	49.5%	49.9%	54.4%	67.9%	**68.6%**
Mean	50.8%	50.9%	55.8%	**66.1%**	65.3%

Table 4: Test macro F1 results for models in *finetuned* scenario run on GoEmotions dataset. S-Ann: SINGLE-ANN, A-Ann: AVG-ANN, HuBi: HuBi-medium, PE: Past-Embedding. The best model for a specified emotion is marked in bold.

HuBi-medium benefits significantly from finetuning in the case of the GoMEotions dataset. Statistical testing showed that there is a significant difference between the non-finetuned and the finetuned model for every category, with a macro F1-score difference of about 6.7 pp on average.

Past-Embedding also conveyed good performance on GoEmotions. Without finetuning, it reached an average macro F1-score of 58.1%, only 1.3% behind HuBi-medium. With finetuning, it reached 65.3%, 1.2% behind the HuBi-medium. It shows the model benefits a lot from finetuning. Nevertheless, despite the relatively small difference with HuBi-medium, statistical testing showed that the difference is significant.

In the finetuned scenario, Past-Embedding is actually the best-performing model for the most of emotions. However, the differences with comparison to HuBi-medium are minimal. On the other hand, HuBi-medium is the best model for the remaining categories with considerable advantages for some of them.

The other personalized method, i.e. User-ID was not as good as HuBi-medium or Past-Embedding, and it was greatly outperformed by HuBi-medium and Past-Embedding on almost every category. However, it still was statistically better compared to both baselines.

We assume User-ID method was struggling more than the other personalized models because of the high number of annotators in the dataset. There are 82 annotators, which is three times more than in the StudEmo dataset. Thus it is harder for the model to learn the user special tokens. It would require more time to learn them properly. Having too many tokens without enough training may lead to a generalization problem, hence the lower performance.

Nevertheless, there are a few categories where User-ID and Past-Embedding performed almost similarly, namely *nervousness* (2.9 pp difference), *pride* (0.8 pp difference), and *relief* (0.7 pp difference). These categories are affected by high data imbalance. It appears that HuBi-medium is able to deal with the high data imbalance the best, while User-ID and Past-Embedding are less efficient in dealing with the issue.

The experiments for the GoEmotions dataset revealed again that the performance of the personalized approaches is much better than for the baselines. In that case, the best model was the HuBi-medium.

8. Conclusions and Future Work

In this work, we present StudEmo, a non-aggregated, manually annotated review dataset for personalized emotion recognition. We also provide detailed information about the source of the texts and annotations, along with the data characteristics including data distribution, number of annotators, and inter-annotator agreement. The dataset keeps all the decisions of the annotators without aggregating or combining them in any way. Thanks to that, it can be used as a benchmark for personalized NLP methods.

That dataset was used to compare the personalized methods with non-personalized baselines. Additional experiments were also performed on the GoEmotions dataset. Two baseline methods were considered: the AVG-ANN baseline which represents the aggregated approach, and the SINGLE-ANN baseline, which represents the non-personalized approach where the model learns individual annotations without any further information about the annotators. Three personalized methods were analyzed: User-ID, where the model is provided with information about the user in the form of a special token; Past-Embedding, where the user beliefs are represented by a vector of the text embeddings and annotations, and HuBi-medium, where additional human embeddings and word biases are learned. For both datasets, the results showed that the personalized methods deliver significantly higher performance compared to baselines.

In StudEmo, the Past-Embedding method featured the highest performance. Without finetuning, it was considerably better compared to not only the baselines, but also the other two personalized models. However, with finetuning, there is no significant difference in the results from User-ID, Past-Embedding, and HuBi-medium. It was shown that finetuning leads to large performance gain for HuBi medium and User-ID methods. The bigger difference between the personalized and non-personalized methods is observed for some controversial emotions. Extra knowledge about user beliefs allows the model to make more appropriate and personalized decisions.

On GoEmotions, HuBi-medium showed the greatest performance with a significant margin. It is slightly better than Past-Embedding, and remarkably better than User-ID and the baselines. We assume that User-ID did not perform as well because a large number of special tokens were injected into the language model. HuBi-medium and Past-Embedding benefit significantly from finetuning.

In future work, the effect of the number of texts in the *Past* split needs to be investigated further because it determines how much knowledge about a user is known to the model. We also would like to see if some ordering of these past texts, such as ranking them by controversy, can further improve the performance.

9. Acknowledgements

This work was financed by (1) the National Science Centre, Poland, project no. 2019/33/B/HS2/02814 and 2021/41/B/ST6/04471; (2) the Polish Ministry of Education and Science, CLARIN-PL; (3) the European Regional Development Fund as a part of the 2014-2020 Smart Growth Operational Programme, CLARIN – Common Language Resources and Technology Infrastructure, project no. POIR.04.02.00-00C002/19; (4) the statutory funds of the Department of Artificial Intelligence, Wroclaw University of Science and Technology.

10. Bibliographical References

Abdullah, M., Hadzikadicy, M., and Shaikhz, S. (2018). Sedat: Sentiment and emotion detection in arabic text using cnn-lstm deep learning. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 835–840.

Akhtar, M. S., Ekbal, A., and Cambria, E. (2020a). How intense are you? predicting intensities of emotions and sentiments using stacked ensemble [application notes]. *IEEE Computational Intelligence Magazine*, 15(1):64–75.

Akhtar, S., Basile, V., and Patti, V. (2020b). Modeling annotator perspective and polarized opinions to improve hate speech detection. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 8(1):151–154, Oct.

Basile, V., Cabitza, F., Campagner, A., and Fell, M. (2021). Toward a perspectivist turn in ground truthing for predictive computing. *ArXiv*, abs/2109.04270.

Bellés-Calvera, L. and Quintana, R. C. (2021). Audiovisual translation through nmt and subtitling in the netflix series 'cable girls'. *Proceedings of the Translation and Interpreting Technology Online Conference TRITON 2021*.

Cambedda, G., Nunzio, G. M. D., and Nosilia, V. (2021). A study on automatic machine translation tools: A comparative error analysis between deepl and yandex for russian-italian medical translation.

Chiorrini, A., Diamantini, C., Mircoli, A., and Potena, D. (2021). Emotion and sentiment analysis of tweets using bert. In *EDBT/ICDT Workshops*, 03.

Davani, A. M., Díaz, M., and Prabhakaran, V. (2022). Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110, 01.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Dudy, S., Bedrick, S., and Webber, B. (2021). Refocusing on relevance: Personalization in nlg. *EMNLP 2021 main conference*.

Fayek, H., Lech, M., and Cavedon, L. (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, 07.

Ghosh, S. and Kumar, S. (2021). Cisco at SemEval-2021 task 5: What's toxic?: Leveraging transformers for multiple toxic span extraction from online comments. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 249–257, Online, August. Association for Computational Linguistics.

He, X., Liao, L., Zhang, H., Nie, L., Hu, X., and Chua, T.-S. (2017). Neural collaborative filtering. *Proceedings of the 26th International Conference on World Wide Web*.

Hernandez, J., Lovejoy, J., McDuff, D., Suh, J., O'Brien, T., Sethumadhavan, A., Greene, G., Picard, R., and Czerwinski, M. (2021). Guidelines for assessing and minimizing risks of emotion recognition applications. *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*.

Hidalgo-Ternero, C. M. (2021). Google translate vs. deepl: analysing neural machine translation performance under the challenge of phraseological variation. *MonTI. Monographs in translation and interpreting*, pages 154–177.

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

Kamran, S., Zall, R., Kangavari, M. R., Hosseini, S., Rahmani, S., and Hua, W. (2021). Emodnn: Understanding emotions from short texts through a deep neural network ensemble.

Kanclerz, K., Figas, A., Gruza, M., Kajdanowicz, T., Kocon, J., Puchalska, D., and Kazienko, P. (2021). Controversy and conformity: from generalized to personalized aggressiveness detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5915–5926, Online, August. Association for Computational Linguistics.

Kocoń, J., Figas, A., Gruza, M., Puchalska, D., Kajdanowicz, T., and Kazienko, P. (2021a). Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach. *Information Processing & Management*, 58(5):102643.

Kocoń, J., Gruza, M., Bielaniewicz, J., Grimling, D., Kanclerz, K., Miłkowski, P., and Kazienko, P. (2021b). Learning personal human biases and representations for subjective tasks in natural language processing. In *2021 IEEE International Conference on Data Mining (ICDM)*, pages 1168–1173, 12.

Krommyda, M., Rigos, A., Bouklas, K., and Amditis, A. (2021). An experimental analysis of data annotation methodologies for emotion detection in short text posted on social media. *Informatics*, 8(1).

Lee, J.-H., Kim, H.-J., and Cheong, Y.-G. (2020). A multi-modal approach for emotion recognition of tv drama characters using image and text. In *2020 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 420–424, 02.

Li, J.-L. and Lee, C.-C. (2019). Attentive to Individual: A Multimodal Emotion Recognition Network

with Personalized Attention Profile. In *Proc. Interspeech 2019*, pages 211–215.

Li, W. and Xu, H. (2014). Text-based emotion classification using emotion cause extraction. *Expert Systems with Applications*, 41(4, Part 2):1742–1749.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Miłkowski, P., Gruza, M., Kanclerz, K., Kazienko, P., Grimling, D., and Kocoń, J. (2021). Personal bias in prediction of emotions elicited by textual opinions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 248–259, 08.

Mohammad, S. M. and Bravo-Marquez, F. (2017). Emotion intensities in tweets. In *Proceedings of the sixth joint conference on lexical and computational semantics (*Sem)*, Vancouver, Canada.

Wang, X. and Tong, Y. (2021). Application of bert+attention model in emotion recognition of metizens during epidemic period. *Journal of Physics: Conference Series*, 1982(1):012102, jul.

Zaśko-Zielińska, M. and Piasecki, M. (2018). Towards emotive annotation in plWordNet 4.0. In *Proceedings of the 9th Global Wordnet Conference*, pages 153–162.

11. Language Resource References

Demszky, Dorottya and Movshovitz-Attias, Dana and Ko, Jeongwoo and Cowen, Alan and Nemade, Gaurav and Ravi, Sujith. (2020). *GoEmotions: A Dataset of Fine-Grained Emotions*. arXiv.

Kennedy, Chris J and Bacon, Geoff and Sahn, Alexander and von Vacano, Claudia. (2020). *Constructing interval variables via faceted Rasch measurement and multitask deep learning: a hate speech application*.

Kocoń, Jan and Miłkowski, Piotr and Kanclerz, Kamil. (2021). *MultiEmo: Multilingual, Multilevel, Multidomain Sentiment Analysis Corpus of Consumer Reviews*. Springer International Publishing.

Kocoń, Jan and Janz, Arkadiusz and Miłkowski, Piotr and Riegel, Monika and Wierzba, Małgorzata and Marchewka, Artur and Czoska, Agnieszka and Grimling, Damian and Konat, Barbara and Juszczyk, Konrad and Klessa, Katarzyna and Piasecki, Maciej. (2019). *Recognition of emotions, valence and arousal in large-scale multi-domain text reviews*.

Leonardelli, Elisa and. Menini, Stefano and Palmero Aprosio, Alessio and Guerini, Marco and Tonelli, Sara. (2021). *Agreeing to Disagree: Annotating Offensive Language Datasets with Annotators' Disagreement*. Association for Computational Linguistics.

Wierzba, Małgorzata and Riegel, Monika and Kocoń, Jan and Miłkowski, Piotr and Janz, Arkadiusz and Klessa, Katarzyna and Juszczyk, Konrad and Konat, Barbara and Grimling, Damian and Piasecki, Maciej and Marchewka, Artur. (2021). *Emotion norms for 6000 Polish word meanings with a direct mapping to the Polish wordnet*.