Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022, pages 37–45
Marseille, 20 June 2022

What if Ground Truth is Subjective? Personalized Deep Neural Hate Speech Detection

Kamil Kanclerz, Marcin Gruza, Konrad Karanowski, Julita Bielaniewcz, Piotr Miłkowski, Jan Kocoń, Przemysław Kazienko
Department of Artificial Intelligence
Wrocław University of Science and Technology, Wrocław, Poland
{kamil.kanclerz, marcin.gruza, konrad.karanowski, julita.bielaniewicz, piotr.milkowski, jak.kocon, kazienko}@pwr.edu.pl

Abstract

A unified gold standard commonly exploited in natural language processing (NLP) tasks requires high inter-annotator agreement. However, there are many subjective problems that should respect users' individual points of view. Therefore, in this paper, we evaluate three different personalized methods for the task of hate speech detection. Our user-centered techniques are compared to the generalizing baseline approach. We conduct our experiments on three datasets including single-task and multi-task hate speech detection. For validation purposes, we introduce a new data split strategy, which prevents data leakage between training and testing. To better understand the behavior of the model for individual users, we carried out personalized ablation studies. Our experiments revealed that all models leveraging user preferences in any case provide significantly better results than most frequently used generalized approaches. This supports our general observation that personalized models should always be considered in all subjective NLP tasks, including hate speech detection.

Keywords: NLP, subjective NLP tasks, hate speech, offensive content, human bias, human representation

1. Introduction

At first glance, disagreement and nonregular annotations can be seen as noise that drags the performance of NLP task detection models down. As we know, the ability to think and perceive the environment differently is natural to humans as such. Therefore, it is crucial to include this observation while building predictive models in order to reflect the setup close to reality. As simple as this may seem, it is important to keep in mind that the key ideas behind NLP phenomenon detection, such as gold standard, agreement coefficients, or the evaluation itself need to be thoroughly analyzed and reconsidered especially for subjective NLP tasks like hate speech detection, prediction of emotional elicitation, sense of humor, sarcasm detection, or even sentiment analysis. Such NLP tasks come with each complexity of their own, especially within the aspect of subjectivity, therefore making them difficult to solve compared to non-subjective tasks.

The changes that need to be implemented do not only consist of acquiring of suitable annotated data, but also of the problem definition itself. The vast majority of methods related to hate speech detection focus on one generalized interpretation of the texts, usually called *ground truth* or *gold standard* (Basile, 2020a), that is, an assignment of a single *right* value to the textual content being labeled. This process could be supported by defining specific guidelines or by adding active learning methods (Huang et al., 2017) in order to adequately address the disagreement of annotations. We, however, follow another *personalized* direction, in which model prediction is individualized for every user.

Our contribution is, inter alia, comprehensive experimental studies on hate speech for three datasets (suitable for both multi-task and single-task) and various personalized architectures (section 3). This data diversity helps us to accurately grasp the accuracy in the subjective setup, regardless of the characteristics of the datasets themselves. We have also decided to compare the fine-tuned and non-fine-tuned models in order to uncover possible errors in the assessment of the scores. Another valuable comparison was performed between collaborative filtering and the transformer-based architecture. Data extraction methods were evaluated side by side with information extraction methods based on data related to attention. As the key personalization ideas needed a new definition, we have managed to formulate a new data split and validation strategy, see Fig. 3. Such enhancements in the fundamental processes and concepts of deep neural solutions to NLP tasks turned out to be more accurate in terms of capturing the subjectivity of a single user, performing a legitimate personalization of user opinions in terms of their sensitivity to hate speech, both as a receiver and as an addressee (Fig. 1). Compared to the generalized approach, we have achieved results that greatly exceed the more common process of gratifying the majority, as seen in Section 6. To magnify and secure the scores achieved, we performed an ablation study, as well as a detailed analysis of the lower performance values in our models.

2. Related work

The number of tasks included in the natural language processing research areas is constantly growing. This phenomenon has potential that will even-

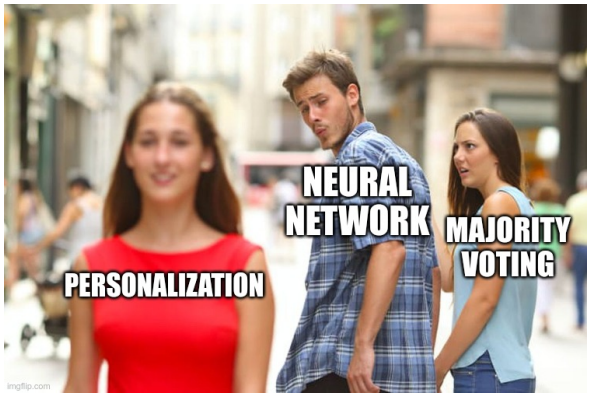

Figure 1: Personalization as an interesting alternative to majority voting.

tually help with the tasks where consumers' opinions will be prioritized. The use of a perspectivist approach performs well in many NLP detection tasks, such as hate speech (Rajadesingan et al., 2015; Zhang et al., 2016; Amir et al., 2016; Gong et al., 2017; Chetty and Alathur, 2018; Fortuna and Nunes, 2018; Gultchin et al., 2019; Kamal and Abulaish, 2019; Kocoń et al., 2021; Mondal and Sharma, 2021). To accurately grasp the idea behind uncovering the universal emotional characterization of the data annotated by users, we first need to define what that gold standard truly is in our case. The authors of the work (Aroyo and Welty, 2015) claim that the truth is completely relative and is more closely related to agreement and consensus. In *Seven Myths*, the myth of One Truth is debunked through various examples, indicating that the *correct* interpretation of the sentence is a matter of opinion, and therefore there is no one true interpretation. This statement is a high-level look at the domain of NLP. However, there are other approaches. As such, the most common is represented through the generalized approach. This method suggests that the majority is the gold standard and the authors of the work (Liu et al., 2019) imply that specific label aggregation methods can help provide reliable representative semantics at the population level. In the domain of detecting and labeling hate speech, recent work (Akhtar et al., 2020) presents an approach that creates different gold standards, one per chosen group. Experiments indicate that supervised models that include different perspectives on a certain topic outperform a baseline model that was trained on fully aggregated data. Similar results exposing these phenomena were presented in (Weerasooriya et al., 2020), which included the size of each group. The authors processed the annotation collection for each data item as a sample of the opinions of a population of human annotators. Among each group of individuals, disagreement was a natural and expected occurrence. Therefore, a standard training set may contain a large number of very small samples, one for each data item, none of which, by itself, is large enough to be considered representative of the beliefs of the underlying population about each topic. Another crucial aspect in the phenomena detection in texts, is the agreement coefficients. Some of them were shown in the work (Artstein and Poesio, 2008) in which the authors exposed the underlying assumptions of the agreement coefficients, covering Krippendorff's alpha, Scott's pi, and Cohen's kappa. They discussed the use of coefficients in various annotation tasks and argued that weighted alpha-like coefficients, traditionally less used than kappa-like measures in computational linguistics, may be more appropriate for many corpus annotation tasks. However, a certain problem with Cohen's Kappa has been found, as described in (Powers, 2012). Deploying a system in a context which has the opposite skew from its validation set can be expected to approximately negate Fleiss's Kappa and halve Cohen's Kappa, but leave Powers Kappa unchanged. For most performance evaluation purposes, the latter is, therefore, most appropriate. Some annotators choose bad labels to maximize their pay. To avoid manual identification, a response model item named MACE (Multi-Annotator Competence Estimation) was introduced in (Hovy et al., 2013). It learns in an unsupervised fashion to identify which annotators are trustworthy and predict the correct underlying labels. The process of matching the performance of more complex state-of-the-art systems performs well even under adversarial conditions. On the other hand, a low level of agreement between annotators can have a positive effect on the performance of the models (Leonardelli et al., 2021). (Plank et al., 2014) present an empirical analysis of part-of-speech annotated data sets that suggests that disagreements are systematic across domains and, to some extent, also across languages. A quantitative analysis of tag confusions reveals that most disagreements are due to linguistically debatable cases rather than annotation errors. And the final key element is the evaluation itself. Although not largely analyzed, it may expose some of the less obvious issues. The work (Basile, 2020b) suggests that majority-driven gold standards can be undone in time, and the coming progress in NLP is headed towards an inclusive approach that may preserve the personal opinions and perspectives of annotators. The same author appeared in the work (Basile et al., 2021) and expressed disagreement with practices such as minimizing disagreement or creating cleaner datasets. That simplification is said to result in oversimplified models for end-to-end tasks. Therefore, there exists a need for improvement evaluation practices in order to better grasp such a disagreement.

3. Datasets

The data we used were collected from three datasets: Measuring Hate Speech, Wikipedia Detox Aggression, and Unhealthy Conversations. All datasets contain texts that are related to offensive speech, yet

differ significantly from each other to a degree that accurately displays the universal nature of the evaluated methods; see Tab. 1 for a detailed data profile.

3.1. Measuring Hate Speech (MHS) dataset

The Measuring Hate Speech dataset (Kennedy et al., 2020) consists of 39,565 comments acquired from YouTube, Twitter, and Reddit. These comments are annotated by 7,912 Amazon Mechanical Turk workers from the United States. The annotators focused on measuring the intensity of various types of offensiveness. It means that a given user annotated a text with the level of each of ten types: (1) disrespect, (2) insult, (3) humiliation, (4) sentiment, (5) attacking or defending nature of the post, (6) dehumanization, (7) inferiority of the status, (8) hate speech, (9) violence, and (10) genocide. Each type was treated by us as another NLP task – a distinct output of the model. The correlations between the annotations for the different types (tasks) are shown in Fig. 2.

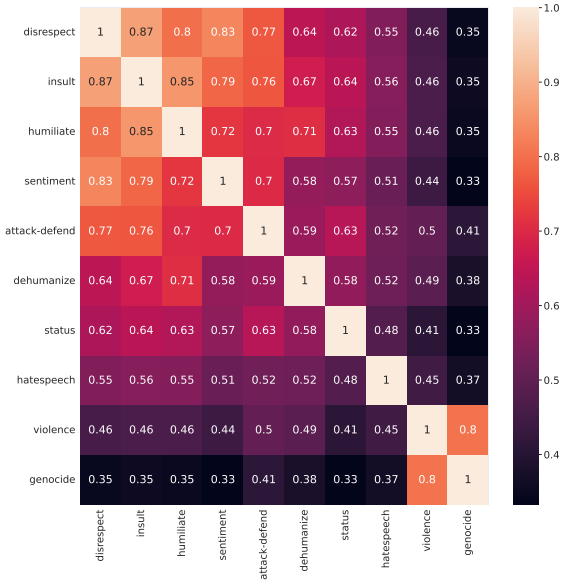

Figure 2: Correlation between real values of the hate speech types (tasks) for the same text in the MHS dataset

3.2. Wikipedia Detox: Aggression

The data available in Wikipedia Detox: Aggression dataset was accumulated during the Wikipedia Detox project [1] that took place between 2001 and 2015. It consists of 116k texts from the Wikipedia forum that were labeled by more than 4k annotators. Each human annotator marked the level of aggression from -3 to 3, where the value -3 defines a highly aggressive text and 3 implies a complete lack of aggression in the labeled text. We have simplified the values to range from -1 to 1, where negative or zero values correspond to the highly aggressive label, whereas the values greater than 0 to the non-aggressive one.

[1] https://meta.wikimedia.org/wiki/Research:Detox/Data_Release

3.3. Unhealthy Conversations

The Unhealthy Conversations dataset (Price et al., 2020) was made publicly available in October 2020. It contains 44k unique comments of 250 characters or less from Globe and Mail opinion articles sampled from the Simon Fraser University Opinion and Comments Corpus dataset (Kolhatkar et al., 2020). Each comment was coded by at least three annotators with at least one of the following class labels: *antagonize*, *condescending*, *dismissive*, *generalization*, *generalization unfair*, *healthy*, *hostile*, and *sarcastic*. The comments were presented in isolation to the annotators, without the surrounding context of the news article and other comments, thus possibly reducing bias.

4. Methods

To investigate the impact of subjectivity on the modeled tasks, we compare four different neural-based models: one non-personalized (TXT-Baseline) and three personalized (HuBi-Formula, HuBi-Medium and UserId). All the described models are neural networks trained using a backpropagation algorithm.

- **TXT-Baseline** (Kocoń et al., 2021) – the baseline model that uses only the language model vector representation for the prediction. This model is used in most NLP tasks, where it is assumed that there is only one ground truth for each text and the prediction is not dependent on the person. The model consists of one linear layer that projects the text vector representation into the desired prediction dimension.
- **HuBi-Formula** (Kocoń et al., 2021) – the simplest personalization model which uses additional statistical features of a person to improve the quality of model predictions. The features of the person are their Z-scores of annotations for each class calculated from the training dataset. The person's Z-score can be interpreted as their standarized deviation from mean labels of texts that he annotated, which allows the model to learn that the person is more or less likely to annotate given label. The architecture of the model is similar to TXT-Baseline, with the difference that Z-scores are concatenated to textual vector representations before the projecting linear layer.
- **HuBi-Medium** (Kocoń et al., 2021) – inspired by collaborative filtering methods, this model learns a personal latent vector which captures personal beliefs about the modeled task. As in the neural collaborative filtering model (He et al., 2017), the personal latent vector is multiplied element-wise with the textual vector, and the resulting vector is further fed to linear layers. Vectors are initialized randomly and learned through backpropagation.

	Measuring Hate Speech	Wiki Detox Aggression	Unhealthy Conversations
Textual content profile	comments	comments & discussions	comments & discussions
Tasks	disrespect, insult, humiliate, sentiment, attack-defend, dehumanize, status, hatespeech, violence, genocide	aggression	antagonize, condescending, dismissive, generalization, unfair generalization, healthy, hostile, sarcastic
Labels / values	$\{0, \ldots, 4\}$	$\{0, 1\}$	$\{0, 1\}$
Output / ML task	10*regression	binary classification	8*binary classification
Number of texts	39,565	115,864	44,355
Number of annotations	135,556	1,365,217	244,468 (227,975 valid)
Number of annotators	7,912	4,053	558
Avg. annotations per text	3.43	11.78	4.66
Avg. annotations per annotator	17.13	336.84	387.71
Language	English	English	English

Table 1: Dataset profiles.

- **UserId** (Kocoń et al., 2021) – this model encodes the information about a person by appending a user ID token to the beginning of the annotated text. The text with the user ID is then encoded with the transformer model into a vector representation. As an extension of the original model, to prevent the tokenizer from splitting the user ID tokens, we manually add them to models' special tokens set. In this model, the transformer weights are trained with the whole model to learn the dependencies between the user and the text.

5. Experimental Setup

To provide a comparison between the generalized approach and personalized methods, we choose the TXT-Baseline architecture as our baseline. It provides the same unified prediction for a given text. It does not take into account the existence of individual users at all. However, to enable comparability of the results, we trained the baseline model in the same setup as the personalized architectures, i.e. treating each annotation concerning a given text and made by a specific user as a separate training sample.

To counteract the possible imbalance between text relevance, we applied the text-based data split and the 10-fold cross-validation shown in Fig. 3.

Due to the various text lengths in each dataset described in Sec. 3 we limited each text to 128 tokens. The WikiDetox Aggression dataset required additional preprocessing, including the removal of the new-line sign from each text. On the other hand, we used multi-objective regression for the MHS dataset and scaled the sample labels to the range $[0, 1]$.

To obtain the vector representations of the texts in each dataset, we leveraged the XLM-RoBERTa (XLM-R) (Conneau et al., 2020) model and its tokenizer. We used the implementation provided by the HuggingFace library (Wolf et al., 2020).

For the TXT-Baseline, HuBi-Formula, and HuBi-Medium models, our experimental setup consists of two phases: generating embeddings and training classifiers. The first phase involves splitting the texts of the training samples into tokens and then generating their embeddings via the language model. On the contrary, we could include the language model in the training process. This would improve the performance of each model, but also significantly increase the learning time, because of the performing the forward and the backward propagation through the layers of the language model, which in our case consists of a very large number of parameters. This setup would be too expensive, taking into account multiple model architectures and the 10-fold cross-validation. The main objective of our work is to show the impact of personalization on the performance of reasoning methods. Another advantage of this approach is a more robust comparison of different model architectures, highlighting the best extraction of user knowledge.

To obtain a vector representation of the text, we averaged the embeddings of all tokens. Our technique differs from the standard approach of focusing on a CLS token that contains a representation of the entire text. During the initial experiments, we found that embedding of the entire text based on the averaged vector representation of the tokens yields better results than the standard technique using the CLS token embedding.

In the case of the UserId model, each text is tokenized and encoded with the transformer in each epoch during the training procedure. This approach results in significantly increased training time. However, it enables fine-tuning of the transformer weights in order to achieve a better quality of the predictions.

In the training process, we used Adam optimizer (Kingma and Ba, 2015) and set the cross-entropy (Zhang and Sabuncu, 2018) as our loss function. The hyperparameter values including the learning rate, the number of epochs, and the size of the training batch were optimized separately for each dataset. *HuBi-Medium* model contains additional hyperparameters related to user representation. The size of the user embedding is set to 50. We initialized the weights of the embedding layer with the values we acquired from the uniform distribution within the range $(-0.01, 0.01)$.

In the case of classification tasks performed on the WikiDetox Aggression dataset, we measured the macro

f1-score (F1). For the regression tasks performed on the MHS dataset, we used the R^2 measure. To measure the significance of the difference between different experiment configurations, we performed statistical tests. After ensuring that the test assumptions are met, we applied the independent samples *t*-test with the Bonferroni correction. If the assumptions could not be fulfilled, we used the Mann-Whitney *U* test.

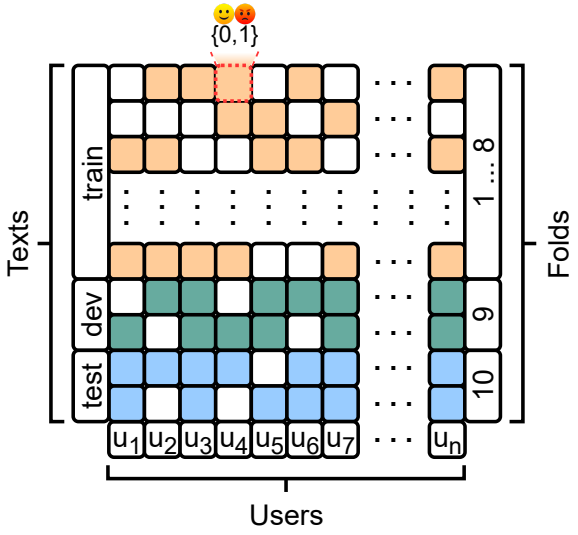

Figure 3: Data split strategy used for each dataset using the example of the WikiDetox Aggression dataset. White blocks are texts, which are not annoated by a specific user.

6. Results

Experiments were carried out for each set presented in Tab. 1. For the Measuring Hate Speech dataset, the results of the model that predict the exact value of each dimension are presented in Tab. 2. For 9 out of 10 dimensions, we see a strong predominance for the UserId model. Thus, it also occurred in the average score for the entire model, 48.67% vs 40.95% (the second-best model, HuBi-Medium). HuBi-Formula and HuBu-Medium models compared to the TXT-Baseline perform significantly better in 5 out of 10 dimensions. They were also superior to UserId on the *dehumanize* dimension. When comparing the score of TXT-Baseline (generalized approach) and HuBi-Medium (personalized approach), we see an analogous jump in the average score as between HuBi-Medium and UserId. The average scores for these two models are 35.45% (TXT-Baseline) and 40.95% (HuBi-Medium), respectively. HuBi-Formula (37.90%) compared to TXT-Baseline (35.45%) also performs slightly better. The most problematic dimension for all models was the *status* dimension. The results for each architecture were at least one third worse than for the other dimensions.

For the second dataset, WikiDetox: Aggresion, the results are shown in Tab. 3. In this case, we are looking at a classification task. The UserId model proved to be the best for the positive class and the macro scale. For the cases where we had no aggression, all 4 models achieved similar results. For a simple binary determination of the content of an utterance type in a text, the differences between the models were no longer as apparent as for the first dataset. The most visible and significant differences for the positive class are around the same values. These are 52.72% (TXT-Baseline), 60.54% (HuBi-Formula), 65.46% (HuBi-Medium), and 69.99% (UserId), respectively. The macro difference between the generalized approach (TXT-Baseline, 72.60%) and the best personalized approach (UserId, 81.91%) is 9.31%. However, between the second best personalized model (HuBi-Medium, 79.49%) and the best personalized model (UserId, with a score of 81.91%), the difference, although significant, is already marginal with respect to the computational complexity of the model and is 2.42%.

The bivariate histogram showing the difference between the regression results obtained for the HuBi-Medium and the TXT-Baseline models is presented in Fig. 4. The points in the upper left half of the diagram (above the red line) are users for whom the personalized HuBi-Medium architecture achieved better results than the generalized baseline. However, the points located in the lower right half of the histogram (under the red line) are the users whose annotations were better predicted by the TXT-Baseline model. It can be seen that the personalized model (HuBi-Medium) achieves the best results in all tasks. The use of personalization improved the performance of the model in the tasks: *humiliate*, *dehumanize*, *violence*, and *genocide*.

For the last dataset, Unhealthy Conversations, the results are presented in Tab. 4. As a consequence of the unbalanced dataset (almost 80% are cases of healthy statements), this is the most difficult dataset presented from a prediction quality perspective. In this case, the model based on the fine-tuned transformer showed tremendous gains. The differences between the other architectures here were as much as tens of percent (e.g. 74.25% vs 46.10% for the *antagonize* dimension in the case of TXT-Baseline). The HuBi-Formula model showed almost no gains relative to the TXT-Baseline model. For the HuBi-Medium architecture for 2 of the 8 classes, we had statistically significant improvements over TXT-Baseline. These were 49.65% vs 46.10% for the *antagonize* dimension and 52.85% vs 44.11% for the *healthy* dimension.

7. Discussion

The architectures evaluated during the experiments are characterized not only by different structures, but also at the level of information extraction. The *HuBi-Formula* model focuses on single-valued human bias (*HB*). It measures how much a user distinguishes themselves from other users based on their decisions. It can be calculated before the training procedure. The *HuBi-*

	respect	insult	humiliate	sentiment	attack-defend	dehumanize	status	hatespeech	violence	genocide	mean
TXT-Baseline	48.03±7.01	43.53±6.96	38.74±6.38	50.17±6.00	34.67±6.19	27.18±5.97	22.79±6.95	32.74±5.81	30.23±6.78	17.98±8.93	34.54±4.39
HuBi-Formula	47.38±6.44	43.20±5.95	41.69±5.44	49.23±5.88	35.19±4.82	**38.77**±3.51	26.58±4.87	35.48±4.32	36.91±6.62	25.19±10.84	37.90±2.85
HuBi-Medium	48.53±6.07	44.45±5.77	**43.52**±4.83	49.80±6.30	36.66±4.58	**42.29**±4.11	**29.73**±6.01	39.10±4.48	**42.57**±9.51	**33.42**±14.36	**40.95**±3.48
UserId	**60.73**±5.24	**55.44**±5.44	**48.86**±5.52	**62.76**±5.05	**48.00**±4.56	37.27±5.36	**34.21**±5.10	**46.78**±6.22	**48.80**±11.53	**43.81**±15.39	**48.67**±8.70

Table 2: R^2 measure values for the Measuring Hate Speech dataset. The values in **bold** are significantly better than the values of other classifiers (rows). Underlined values are significantly better than in other tasks (columns).

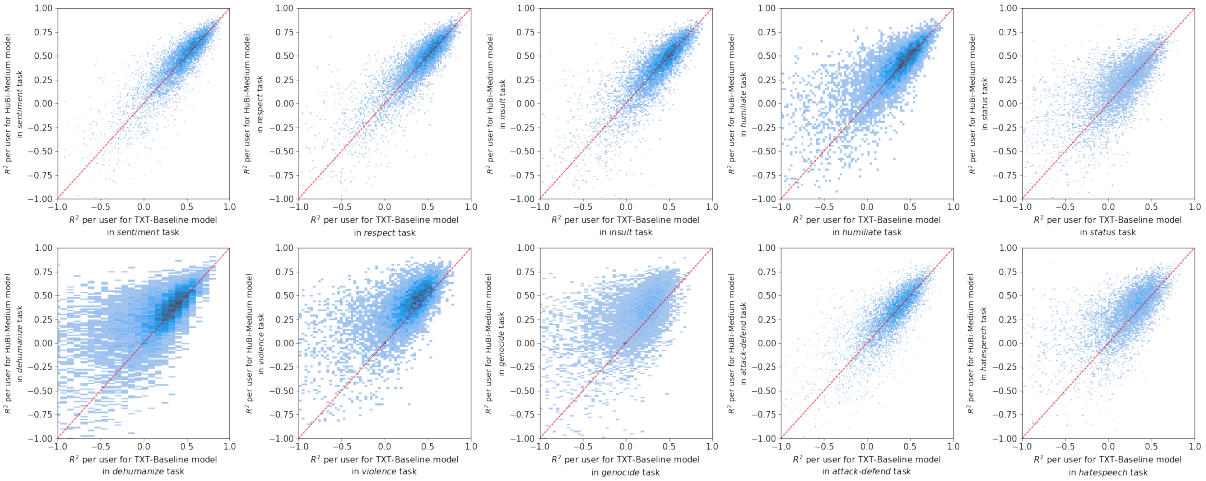

Figure 4: Bivariate histogram of the results R^2 obtained by the HuBi-Medium and TXT-Baseline models for individual users for each of the tasks in the Measuring Hate Speech Dataset. The area was narrowed down to [-1, 1] because at least 95% of users obtained this result on each task.

	F1 negative	F1 positive	Macro F1
TXT-Baseline	92.21 ± 0.36	52.72 ± 1.64	72.60 ± 0.94
HuBi-Formula	92.91 ± 0.34	60.54 ± 1.05	76.82 ± 0.56
HuBi-Medium	93.38 ± 0.31	65.46 ± 0.96	79.49 ± 0.39
UserId	**93.83** ± 0.16	**69.99** ± 0.94	**81.91** ± 0.43

Table 3: Classification results for WikiDetox: Aggression dataset. Values in **bold** are significantly better than other classifiers (rows). Underlined values are significantly better than the performance of the given model in other tasks (columns). Metrics: F1 negative – F1 score for the nonaggressive class (0); F1 positive – F1 score for the aggressive class (1); Macro F1 – the macro average of the F1 scores for each class.

Medium model involves the user representation obtained during the training procedure through the backpropagation procedure. On the other hand, the *UserId* model takes advantage of the transformer-based architecture with masked language modeling and self-attention. Those two are different ways of information extraction, including the user representation generation procedure.

The *UserId* model achieved the best result on the vast majority of tasks in each of the evaluated datasets. This may be related to its much more complex structure compared to the other classifiers. The fine-tuned transformer architecture combined with self-attention mechanism allowed for a better understanding of the text and improved the ability to extract additional knowledge about the user preferences.

The greatest gains in the case of WikiDetox: Aggression dataset were observed for the *aggressive* class (1). This may be due to the much more subjective nature of this label.

Applying the 10-fold cross-validation allowed conducting statistical tests, and measuring the standard deviation between each model performance on specific folds provided information about its stability. Moreover, fine-tuning the language model in this setup would be much more expensive.

In addition to individual user annotations, metadata such as the context of texts, comments, and information about the author may allow the extraction of additional knowledge.

8. Conclusions and Future Work

The experiments carried out on three datasets allowed us to observe some interesting phenomena. The task of detecting hate speech is difficult due to its complex context. The first significant issue is the lack of the possibility of application of simple dictionary analysis because wordplay really matters in hate interpretation. For this reason, we have shown that using appropriate architectures and state-of-the-art solutions extracts representations containing complete knowledge from text.

The second problem is that each user may have very different perception of offensiveness. The personalized approach allowed us to substantially increase the prediction quality compared to the generalized approach.

This leads us to the general conclusion presented in Fig. 5: *the ground truth* is *subjective*. Therefore, we

	antagonize	condescending	dismissive	generalisation	unfair generalisation	healthy	hostile	sarcastic
TXT-Baseline	46.10 ± 0.21	45.80 ± 0.14	46.74 ± 0.20	47.99 ± 0.23	48.23 ± 0.20	44.11 ± 0.32	47.10 ± 0.15	46.31 ± 0.22
HuBi-Formula	46.15 ± 0.19	45.85 ± 0.17	46.76 ± 0.20	47.99 ± 0.23	48.23 ± 0.20	44.30 ± 0.34	47.11 ± 0.15	46.32 ± 0.22
HuBi-Medium	49.65 ± 2.49	48.03 ± 1.54	47.25 ± 0.43	47.99 ± 0.23	48.23 ± 0.20	52.85 ± 4.69	47.17 ± 0.19	46.37 ± 0.23
UserId	**74.25** ± 1.77	**71.88** ± 3.14	**67.87** ± 4.18	**68.72** ± 4.20	**67.78** ± 4.37	**66.68** ± 1.86	**70.40** ± 2.62	**65.96** ± 2.99

Table 4: Classification results for Unhealthy Conversations dataset. The values in **bold** are significantly better than other classifiers (rows). Underlined values are significantly better than the performance of the given model in other tasks (columns). Metrics: Macro F1 – the macro average of the F1 scores for each class.

Figure 5: Meme representing the moment of sudden realization that the ground truth we were all looking for is subjective and we cannot use approaches based on generalization.

should gather and incorporate knowledge about annotators into the reasoning models.

Our validation of three personalized architectures on three distinct datasets revealed that the UserId model usually performs best even though it requires the user to be precisely identified before the training process.

The code for all methods and experiments is publicly available on GitHub[2] under the MIT license.

Overall, we strongly believe that architectures capable of representing the user beliefs in the comprehensive way appear to be the future of inference for subjective NLP tasks including hate speech detection.

Based on our experiments on the Unhealthy Conversations dataset, we want to address the problem of dimensional imbalance in our future work. Only 20% of this dataset corresponds to instances with unhealthy speech. Thus, seven dimensions are massively underrepresented in relation to the healthy speech cases.

[2] https://github.com/CLARIN-PL/personalized-nlp/releases/tag/2022-lrec-nlperspectives

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

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

Amir, S., Wallace, B. C., Lyu, H., and Silva, P. C. M. J. (2016). Modelling context with user embeddings for sarcasm detection in social media. *arXiv preprint arXiv:1607.00976*.

Aroyo, L. and Welty, C. (2015). Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596.

Basile, V., Fell, M., Fornaciari, T., Hovy, D., Paun, S., Plank, B., Poesio, M., Uma, A., et al. (2021). We need to consider disagreement in evaluation. In *1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21. Association for Computational Linguistics.

Basile, V. (2020a). It's the end of the gold standard as we know it. In *International Conference of the Italian Association for Artificial Intelligence*, pages 441–453. Springer.

Basile, V. (2020b). It's the end of the gold standard as we know it. on the impact of pre-aggregation on the evaluation of highly subjective tasks. In *2020 AIxIA Discussion Papers Workshop, AIxIA 2020 DP*, volume 2776, pages 31–40. CEUR-WS.

Chetty, N. and Alathur, S. (2018). Hate speech review in the context of online social networks. *Aggression and violent behavior*, 40:108–118.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M.,

Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July. Association for Computational Linguistics.

Fortuna, P. and Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

Gong, L., Haines, B., and Wang, H. (2017). Clustered model adaption for personalized sentiment analysis. In *Proceedings of the 26th International Conference on World Wide Web*, pages 937–946.

Gultchin, L., Patterson, G., Baym, N., Swinger, N., and Kalai, A. (2019). Humor in word embeddings: Cockamamie gobbledegook for nincompoops. In *International Conference on Machine Learning*, pages 2474–2483. PMLR.

He, X., Liao, L., Zhang, H., Nie, L., Hu, X., and Chua, T.-S. (2017). Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*, pages 173–182.

Hovy, D., Berg-Kirkpatrick, T., Vaswani, A., and Hovy, E. (2013). Learning whom to trust with mace. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130.

Huang, S.-J., Chen, J.-L., Mu, X., and Zhou, Z.-H. (2017). Cost-effective active learning from diverse labelers. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 1879–1885. AAAI Press.

Kamal, A. and Abulaish, M. (2019). Self-deprecating humor detection: A machine learning approach. In *International Conference of the Pacific Association for Computational Linguistics*, pages 483–494. Springer.

Kennedy, C. J., Bacon, G., Sahn, A., and von Vacano, C. (2020). Constructing interval variables via faceted Rasch measurement and multitask deep learning: a hate speech application. *arXiv e-prints*, page arXiv:2009.10277, September.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *ICLR (Poster)*.

Kocoń, J., Gruza, M., Bielaniewicz, J., Grimling, D., Kanclerz, K., Miłkowski, P., and Kazienko, P. (2021). 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. IEEE.

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

Kolhatkar, V., Wu, H., Cavasso, L., Francis, E., Shukla, K., and Taboada, M. (2020). The sfu opinion and comments corpus: A corpus for the analysis of online news comments. *Corpus Pragmatics*, 4(2):155–190.

Leonardelli, E., Menini, S., Palmero Aprosio, A., Guerini, M., and Tonelli, S. (2021). Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Liu, T., Venkatachalam, A., Sanjay Bongale, P., and Homan, C. (2019). Learning to predict population-level label distributions. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 1111–1120.

Mondal, A. and Sharma, R. (2021). Team_KGP at SemEval-2021 task 7: A deep neural system to detect humor and offense with their ratings in the text data. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1169–1174. Association for Computational Linguistics, August.

Plank, B., Hovy, D., and Søgaard, A. (2014). Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511.

Powers, D. M. W. (2012). The problem with kappa. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 345–355.

Price, I., Gifford-Moore, J., Fleming, J., Musker, S., Roichman, M., Sylvain, G., Thain, N., Dixon, L., and Sorensen, J. (2020). Six attributes of unhealthy conversation. *arXiv preprint arXiv:2010.07410*.

Rajadesingan, A., Zafarani, R., and Liu, H. (2015). Sarcasm detection on twitter: A behavioral modeling approach. In *Proceedings of the eighth ACM international conference on web search and data mining*, pages 97–106.

Weerasooriya, T. C., Liu, T., and Homan, C. M. (2020). Neighborhood-based pooling for population-level label distribution learning. *arXiv preprint arXiv:2003.07406*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.

Zhang, Z. and Sabuncu, M. (2018). Generalized cross entropy loss for training deep neural networks with

noisy labels. *Advances in neural information processing systems*, 31.

Zhang, M., Zhang, Y., and Fu, G. (2016). Tweet sarcasm detection using deep neural network. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: technical papers*, pages 2449–2460.