

Multi-Modal Personalized Hate Speech Analysis using Differential Dataset Cartography

Jan Kocoń, Joanna Baran and Kamil Kanclerz

Department of Artificial Intelligence, Wrocław University of Science and Technology, Poland

Abstract

In recognizing hate speech in text, a frequently overlooked aspect is the specific recipient of the content. Information about the user can be considered as another potential modality in addition to the textual representation. In this work, we present the multi-modal hate speech detection problem as a task of personalized prediction based on text and human representation learned from historical user decisions against offensive content, also as the subjective perception of humiliation, insult, sentiment, and violence. In addition, we present our Differential Data Maps method for visually comparing models for hate speech detection. Our results show that personalized models significantly better predict hate speech against a given individual, and the proposed explainable artificial intelligence method allows us to formulate new hypotheses about the impact of personalization on model performance.

Keywords

hate speech, natural language processing, personalization models, differential dataset cartography

1. Introduction

In the classical approach to text classification in natural language processing (NLP), the goal of a task is to assign one or more labels to a text based on its content [1, 2, 3, 4]. For example, it might be identifying fake news [5, 6], emotions [7, 3, 8, 9, 10, 11, 12], or hate speech [13, 14, 15]. However, it is difficult to define these tasks unambiguously, and one can find various definitions in the literature that are not consistent [16, 17]. Similarly, it is natural to react differently to the same content. These differences may be due to where we were born, how old we are, what kind of education we have, and what cultural background we belong to. A growing body of research shows that even with this information, we are not necessarily in a position to significantly better predict how a particular person would react to the content of a text [18, 17].

In recent years, personalized models have become increasingly popular in prediction tasks [18, 19, 20, 21]. These models get information about a person and the content as input. It turns out that the most important from the perspective of a subjective task (e.g., recognizing the offensiveness of a text) is at least minimal knowledge about a person in the form of his or her decisions concerning a few examples of content [22]. Moreover, this type of information

De-Factify 2: 2nd Workshop on Multimodal Fact Checking and Hate Speech Detection, co-located with AAAI 2023. 2023 Washington, DC, USA


✉ jan.kocoon@pwr.edu.pl (J. Kocoń); joanna.baran@pwr.edu.pl (J. Baran); kamil.kanclerz@pwr.edu.pl (K. Kanclerz)

🌐 <https://ai.pwr.edu.pl/author/jan-kocoon/> (J. Kocoń); <https://ai.pwr.edu.pl/author/joanna-baran/> (J. Baran);

<https://ai.pwr.edu.pl/author/kamil-kanclerz/> (K. Kanclerz)

🆔 0000-0002-7665-6896 (J. Kocoń); 0000-0001-6792-7028 (J. Baran); 0000-0002-7375-7544 (K. Kanclerz)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

significantly improves the prediction quality, and the value of this improvement is much higher than if only demographic information is provided [23]. However, this approach tends to be more costly, requiring manual annotations from hundreds and sometimes even thousands of users for model development. Additionally, it is desirable (though not necessary) that multiple users evaluate the same text in such a collection [24]. Additionally, inference on a production model requires a minimum of information about the user for whom the model predicts potential responses [18, 19, 23].

There are many different ways to represent a human in such personalized architecture. In state-of-the-art approaches, this is usually trainable user embedding. It can be part of the transformer model, such as the UserId model [25]. Still, it can also be a separate component supporting the transformer model and conceptually similar to neural collaborative filtering, such as HuBi-Medium [19]. Analysis of results of such models from the literature for related tasks, e.g., emotion recognition in text [26, 23], shows that state-of-the-art models produce very similar results when using traditional measures such as F1-Score or R-Squared. However, detailed case analysis often shows that different personalized models improve different cases. It is essential to develop more sophisticated explainable artificial intelligence (XAI) methods that show differences that are not visible using standard classification or regression quality measures.

In this work, we present an adaptation of personalized models to analyze multi-modal hate speech in text. We consider modalities such as the content of the text and the identifier of the user who rated the text. In addition, we present our new XAI method called Differential Data Maps (DDM), which allows us to analyze differences in models that have similar classification results. We studied the Measuring Hate Speech dataset [27] for selected offensive categories. The results show that personalized models significantly improve hate speech prediction quality for known users. In addition, we present examples of hypotheses that can be drawn from analyses of differences between baseline and personalized models using DDM.

2. Related Work

The lack of a precise definition of hate speech makes it a rather complex phenomenon that requires additional expertise to conduct a proper analysis. One can take advantage of the user context to tackle the complex nature of detecting it. The use of user perspective appears to significantly improve the performance of many hate speech-related domains, including sarcasm detection [28, 29], sentiment analysis [30], self-deprecating humor recognition [31], offensive content detection [32], and general hate speech analysis [33, 16]. The assumption regarding the influence of user preferences on the final label is contrary to the concept of the gold standard, which is commonly used in many natural language processing tasks. Some authors [34] believe that the truth has a purely relative nature and is strongly related to agreement and consensus. However, there are many approaches focused on addressing the users' various points of view. The most common is the generalized approach, which assumes that the majority's perspective is the gold standard [35]. Another approach is to generate user clusters according to their beliefs and then represent each group's point of view as a separate true label [36]. Experiments indicate that providing knowledge on diverse user perspectives outperforms a model trained on fully

aggregated data [37, 19].

Modern artificial intelligence (AI) methods are characterized by complex nature. A large number of parameters allows them to learn intricate data patterns. However, there is a risk that the model has memorized specific examples from the training set but does not have general knowledge of the phenomenon it should learn about. To prevent this, explainable artificial intelligence methods should be used to understand the model behavior[38, 39]. Moreover, identifying a missing part can significantly improve the effectiveness of a model [40].

On the other hand, apart from scientists, there is a growing need for everyday users to understand AI solutions thoroughly. AI’s ethics, trust, and bias are difficult to pinpoint when the algorithm is treated as a black box [41]. Explanations must make the AI algorithm expressive, improving human understanding and confidence that the model makes just and impartial decisions [42]. Furthermore, to guarantee the personalized model’s trust, transparency, and fairness, it is necessary to provide an advanced evaluation procedure focused on explaining the impact of user context on model behavior.

3. Dataset

To evaluate various personalized architectures, we leveraged the Measuring Hate Speech (MHS) dataset [27]. It contains 39,565 samples representing comments obtained from social media services, including YouTube, Twitter, and Reddit. The texts are annotated by 7,912 people from the Amazon Mechanical Turk platform. The users focused on intensity levels for five types of offensiveness: (1) hate speech, (2) humiliation, (3) insult, (4) sentiment, and (5) violence. We treated each type as another NLP task – a distinct output of the model. The distribution of labels for each task is presented in Figure 1. Most MHS dimensions are heavily unbalanced, like

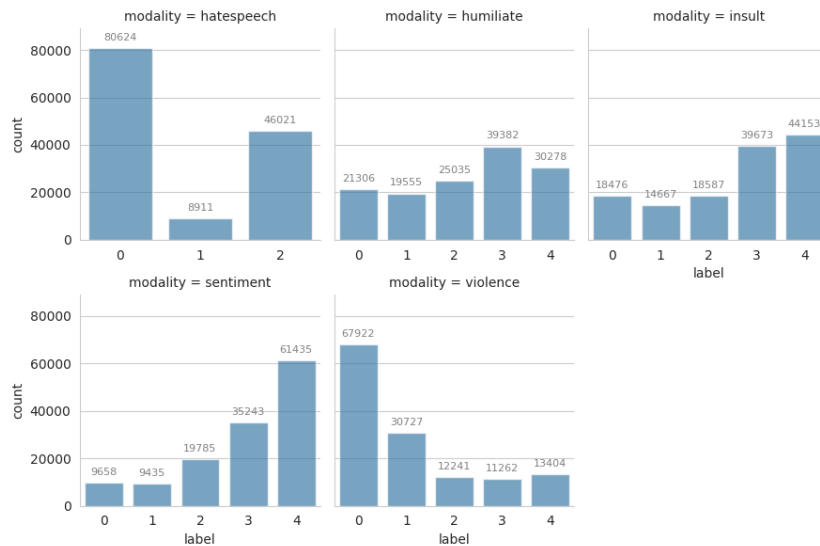


Figure 1: MHS dataset label distribution across different modalities. The diversity of values between similar dimensions demonstrates the subjective nature of the offensiveness detection.

hate speech, sentiment, and violence. An inverse correlation has been observed between the last two.

4. Personalized Models

To explore the impact of providing knowledge about the user, we selected four different neural-based architectures: one non-personalized (Baseline) and three personalized (HuBi-Medium, UserId, and UserIdentifier). We used the backpropagation algorithm during the training procedure for all described models.

1. **HuBi-Medium** [19] – leveraging the idea of collaborative filtering [43], this architecture learns a personal vector representation that encapsulates personal preferences about the selected task. Similar to the original collaborative filtering, the user vector is multiplied using Hadamard product operation with the textual vector. The final vector is then fed to linear layers.
2. **UserId** [17] – this model represents the information about the user by appending their unique ID token to the beginning of the text. The vector representation is obtained via the transformer model by encoding the concatenation of the text and user ID. We manually added the user ID tokens to the model’s special token set to avoid splitting them during the tokenization procedure.
3. **UserIdentifier** [18] takes into account the identity of the text’s author. A data augmentation method involves adding tokens that identify the user. The string is generated from the username or sampled uniformly from the tokenizer vocabulary and then appended to the beginning of a text. UserIdentifier uses the same set of parameters to embed both sample content and user identifiers, which is more straightforward than relying on user-specific embeddings and has been shown to achieve high performance.

5. Differential Data Maps

The idea was inspired by work [44]. The authors present a Data Maps method using a machine learning model to visualize a dataset. It allows seeing how specific elements of the training set are characterized during the learning process. The intuition behind training dynamics is that the model learns to recognize some elements immediately. For other elements, the model needs more learning epochs, during which it can interchangeably make good or bad decisions relative to the ground truth. Finally, the model cannot learn the ground truth for the last group of elements. Three major training dynamics measures for the i th sample in the dataset were introduced:

1. **Confidence**, $\hat{\mu}_i$ – captures how confidently the model assigned a *true* label to the sample, calculated as a mean probability across epochs;
2. **Variability**, $\hat{\sigma}_i$ – measures how the model was indecisive about sample label during training using standard deviation (low value means the stable prediction of one label, and high value - often change of assigned label);

3. **Correctness**, \hat{c}_i – a fraction of correctly predicted labels for the sample across training epochs.

In this work, we extend the idea of Data Maps by proposing visualizing the differences between models in the listed training dynamics measures. Our new method, Differential Data Maps, allows us to interpret differences in the performance of different model architectures and analyze the effect of selected characteristics describing the data on the difference in training dynamics on the same dataset. We define three new metrics based on those presented for Data Maps. Let M1 and M2 be different models trained on the same dataset. Then for i th sample in this dataset, we define new measures:

1. **Confidence change**: $\hat{\mu}_i^C = \hat{\mu}_i^{M2} - \hat{\mu}_i^{M1}$
2. **Variability change**: $\hat{\sigma}_i^C = \hat{\sigma}_i^{M2} - \hat{\sigma}_i^{M1}$
3. **Correctness change**: $\hat{c}_i^C = \hat{c}_i^{M2} - \hat{c}_i^{M1}$

6. Experiments

The experimental part was performed on the previously described MHS dataset, which was divided into three sets to provide sufficient prior knowledge in training set about users’ profiles. For this purpose, the data were grouped by *annotator_id* and filtered out those who gave less than 20 reviews. The entries rated by the same annotator were then divided between the splits in a ratio of 6:2:2. The final statistics of the training and evaluation data are shown in Table 1.

Table 1
MHS dataset split setup.

Split	#Samples	Avg text count per user
Train	29,170	12.6 ± 0.8
Validation	9,817	4.2 ± 0.4
Test	10,888	4.7 ± 0.5

RoBERTa-base pretrained language model [45] was used as a baseline and in personalized approaches. Models were fine-tuned using AdamW optimizer with learning rate 1e-5 and batch size equal to 32. Additionally, we used a linear warm-up schedule for 1000 training steps. The maximum sequence length was set to 512. The best model was selected according to the validation F-score across 30 epochs. For UserIdentifier, we took ten tokens drawn from the tokenizer vocabulary as an identifier. This has enabled better differentiation between users than relying on usernames or strings of numbers. On the other hand, the UserId model leveraged the embedding of a special ID token, which is a concatenation of the word *user*, an underscore character (`_`), and a unique index number for a specific user, i.e., *user_1*.

All experiments were repeated ten times with the same data order but different weights (the model seed was changing). To plot DDM, training dynamics were also logged after each epoch. Two classification metrics were reported to compare model performance. In the last step, we assessed the statistical significance of achieved results differences between each method. After checking the assumptions of the t-student test, its score was calculated accordingly. The Mann-Whitney U test was applied if t-test conditions were not met for independent samples.

Table 2

Evaluation results for specific dimensions in the Measuring Hate Speech dataset. Metrics: F1 – macro F1-score, Acc – Accuracy. Values in **bold** indicate significantly better performance than the Baseline model, and the underlined values are the best among all others.

Metric	Model	Hate Speech	Insult	Violence	Humiliate	Sentiment
F1	Baseline	52.86±0.6	42.64±0.1	41.60±0.6	40.89±1.2	58.1±1.2
	UserIdentifier	57.78±0.7	44.73±0.5	49.56±0.8	43.95±0.6	58.8±0.6
	UserId	57.24±0.6	44.32±0.5	49.08±0.8	43.70±0.6	59.2±0.6
	HuBi-Medium	<u>59.89±0.6</u>	<u>46.91±0.4</u>	<u>51.14±0.7</u>	<u>44.97±0.5</u>	<u>61.3±0.5</u>
Acc	Baseline	70.81±1.8	49.29±1.3	50.87±0.9	42.53±0.9	48.4±0.7
	UserIdentifier	73.33±0.1	50.39±0.4	61.03±0.9	45.60±0.6	48.5±0.7
	UserId	73.21±0.2	50.27±0.4	60.85±0.9	45.24±0.6	48.6±0.7
	HuBi-Medium	73.42±0.2	50.44±0.3	60.83±0.8	44.97±0.5	48.3±0.6

7. Results

Table 2 shows the results of the Baseline model and the other three personalized models. The models were evaluated on an MHS dataset for five dimensions of offensiveness. We measured the performance quality using the macro F1-score and accuracy. Analysis of the results shows that for the F1-macro measure for all dimensions of offensiveness, a significant quality gain is observed relative to the baseline for personalized models. The best among them is the HuBi-Medium model, for which the gain is 7.03 pp for *hate speech*, 4.2 pp for *insult*, 9.54 pp for *violence*, 4.08 pp for *humiliate* and 3.2 pp for *sentiment*, respectively. For the Accuracy measure, statistically significant quality gains were observed for three of the five dimensions, i. e: *hate speech*, *violence*, and *humiliate*. Further analysis showed no significant differences between the personalized models for this measure.

It is much more interesting to analyze the differences between the baseline and personalized models using DDC. Figure 2 shows the original DC graph, generated using the method described in [44]. For virtually all dimensions, the graphs for the baseline model look very similar, and larger differences are observed between the graphs of the various personalized models. The original DC graph, however, is difficult to interpret, as all that can be said is how the distribution of confidence and variability values changes in a general way for all cases. There are some dimensions (e.g., sentiment) for which the DC charts are very similar, regardless of the model used. However, significant differences can be seen by analyzing the DDC graph for data samples; see Figure 3. For the *sentiment*, it can be seen that the UserIdentifier model improves the correctness of the model for the vast majority of samples. At the same time, for UserId, there appears to be a smaller but significantly large group of samples for which the correctness decreases. Paradoxically, for the best model, HuBi-Medium, we observe the largest groups of cases characterized by the greatest decrease in correctness. We hypothesize that the person component of the multimodal model causes the model to start getting wrong more often in the learning process. However, the result is ultimately better than the baseline model. The UserIdentifier model makes the least amount of these mistakes in the learning process, but this does not translate into better quality as measured by the F1-score. HuBi-Medium is also the

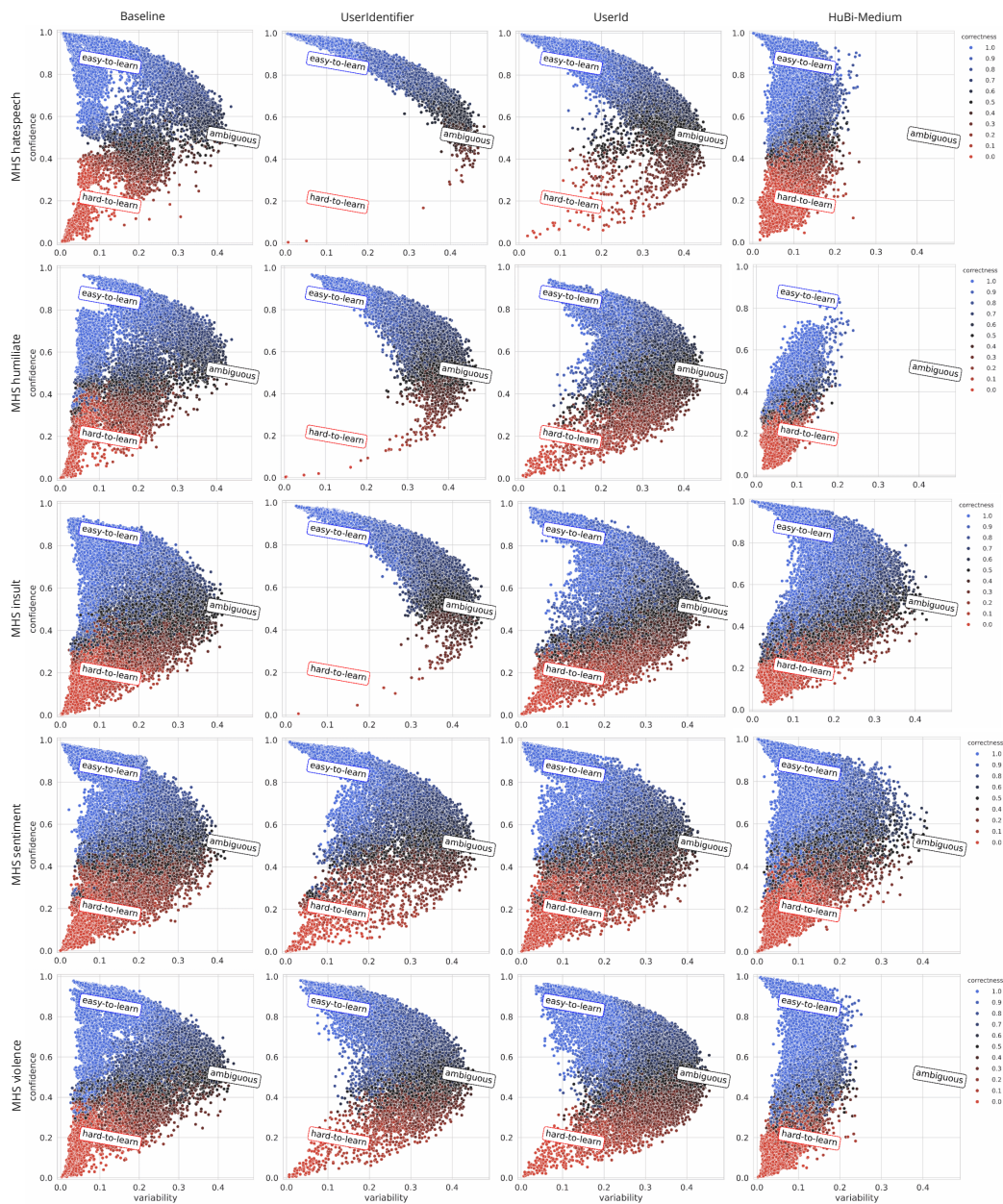


Figure 2: The original Data Maps method for data samples across different modalities. Depending on the model used, data division between hard-to-learn, ambivalent and easy-to-learn changed.

only model for which variability decreases relative to baseline in the vast majority of samples, indicating that the personalization component for this model significantly affects the rate at which the model converges to a local optimum, after which there is little variation in subsequent epochs.

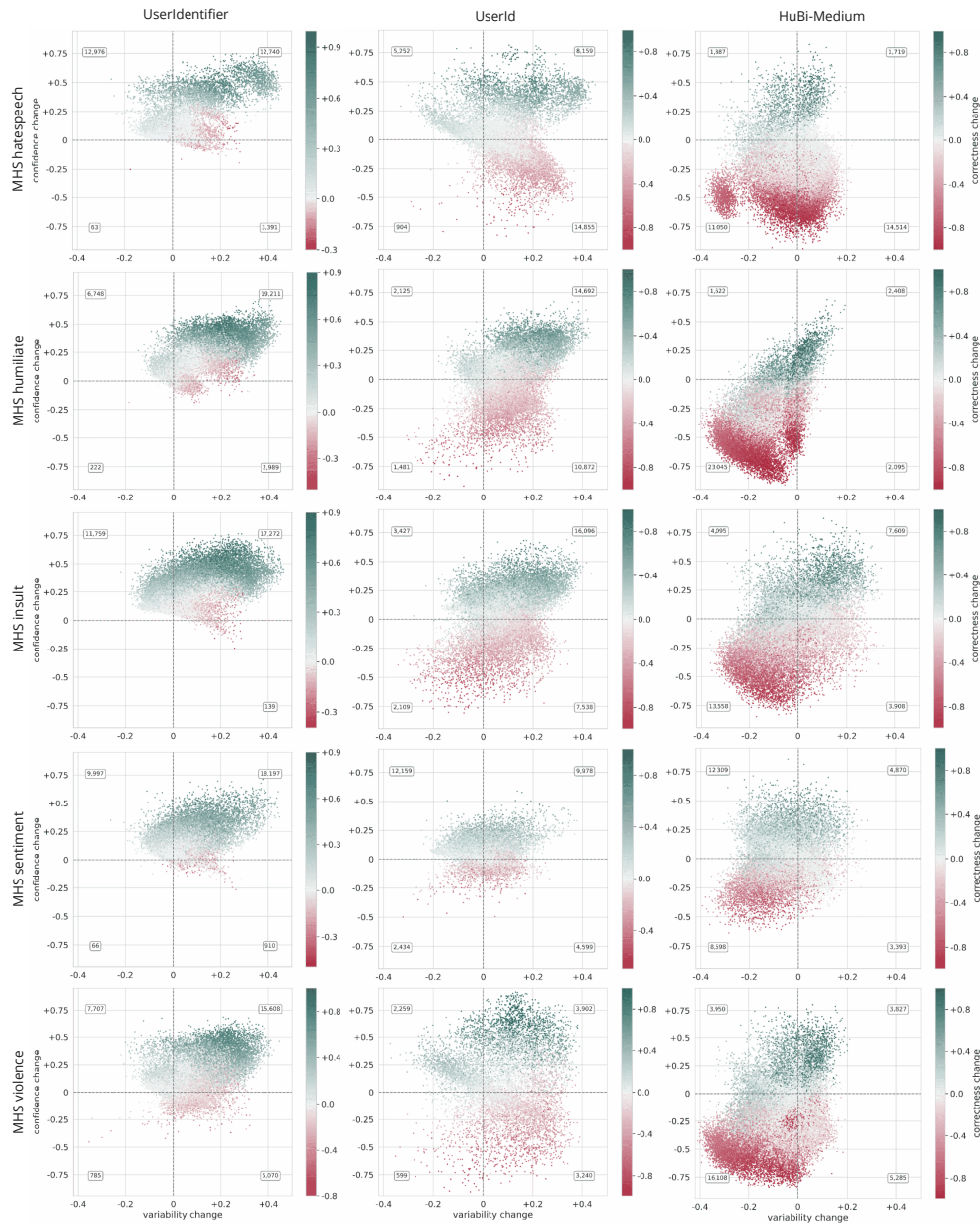


Figure 3: Results of the Differential Data Maps for data samples.

Furthermore, interesting results can be observed in the DDC variant, in which the training data are aggregated after users; see Figure 4. Each point on the graph represents cases annotated by a particular person, and the place on the graph indicates the shift of the personalized model results relative to the baseline. In addition, we added information indicating the entropy of the user's ratings in the set. The UserIdentifier model for all dimensions of offensiveness resulted

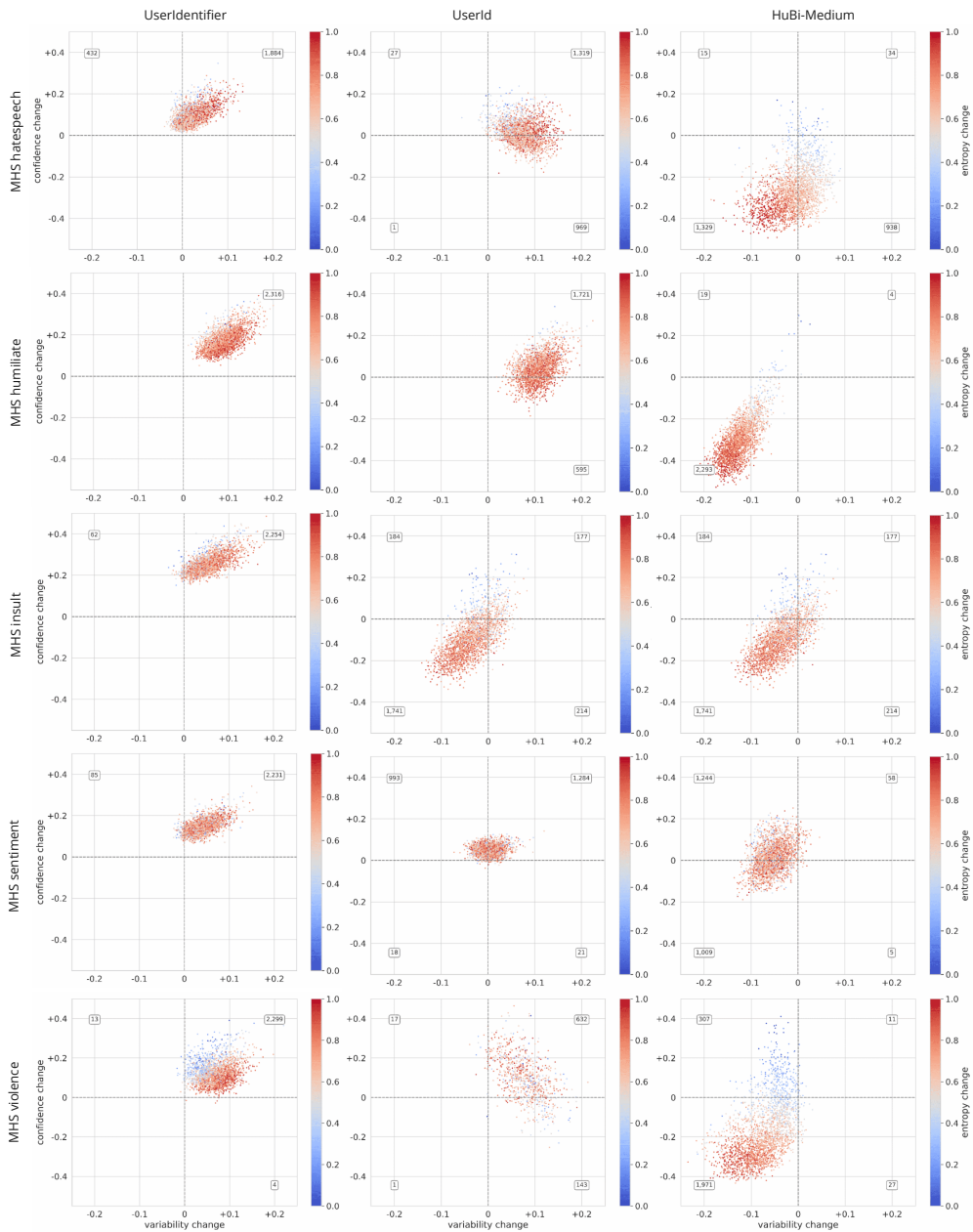


Figure 4: Results of the Differential Data Maps for data samples aggregated by users.

in increased correctness and variability for most users. The lower the entropy of user ratings, the higher the correctness. In the UserID model for dimensions in which a large group of users has variability lower than the baseline (insult, sentiment), we observe no significant differences between the baseline and personalized models. Finally, for HuBi-Medium, the lack of significant differences is strongly correlated with the largest increase in confidence. In addition, it can be

seen that for all dimensions, we observe a significant decrease in variability, with confidence decreasing most strongly for users with high entropy and mostly increasing for users with low entropy.

8. Conclusions and Future Work

In this article, we presented a novel evaluation method called Differential Dataset Cartography. It allows for pairwise visual comparison of model performance. During experiments, we have shown interesting findings provided by our new method. However, further analysis could help identify more insights that could not be obtained from the raw metric values.

The experiments show that including the user context results in significantly improved performance compared to the baseline model. The evaluation metrics show that the HuBi-Medium model outperformed other architectures in most tasks. However, using DDM provided an additional perspective for analyzing the model behavior. Our method emphasized the difference in user representations and the nature of training each of the personalized architectures. It also provided additional information on how each architecture gathers knowledge about users.

Moreover, by aggregating the DDM by users, we explored how much knowledge the model can extract from a specific user. This shows another aspect of human perception – we can discover how difficult it is to learn the preferences of a single person from the model point of view. This can be used as additional feedback relevant during the annotation process to estimate how much data we need about a particular user to learn their perspective effectively. Furthermore, a precise user learning difficulty estimate can be helpful during the architecture design process, which should consider awareness of the general difficulty level of the task.

In future work, we will conduct more experiments using other datasets to obtain more knowledge about the behavior of personalized models. In addition, we want to analyze in detail the samples that form the clusters that can be observed in the DDM charts. This will help to understand the impact of specific samples on the model efficiency. The code for all methods and experiments is publicly available in CLARIN-PL repository¹.

Acknowledgments

This work was financed by (1) the European Regional Development Fund as a part of the 2014-2020 Smart Growth Operational Programme, project no. POIR.01.01.01-00-0288/22 (JK); POIR.04.02.00-00C002/19 (JB, KK), (2) the statutory funds of the Department of Artificial Intelligence, Wrocław University of Science and Technology.

References

- [1] S. Lai, L. Xu, K. Liu, J. Zhao, Recurrent convolutional neural networks for text classification, in: Twenty-ninth AAAI conference on artificial intelligence, 2015.

¹<https://github.com/CLARIN-PL/personalized-nlp/releases/tag/2023-aaai-defactify>

- [2] J. Kocoń, P. Miłkowski, M. Zaśko-Zielińska, 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), 2019, pp. 980–991.
- [3] J. Kocoń, A. Janz, P. Miłkowski, M. Riegel, M. Wierzba, A. Marchewka, A. Czoska, D. Grimling, B. Konat, K. Juszczak, et al., Recognition of emotions, valence and arousal in large-scale multi-domain text reviews, in: 9th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, 2019.
- [4] K. Kanclerz, P. Miłkowski, J. Kocoń, Cross-lingual deep neural transfer learning in sentiment analysis, *Procedia Computer Science* 176 (2020) 128–137.
- [5] R. Oshikawa, J. Qian, W. Y. Wang, A survey on natural language processing for fake news detection, in: Proceedings of the 12th LREC Conference, 2020, pp. 6086–6093.
- [6] N. Micallef, M. Sandoval-Castañeda, A. Cohen, M. Ahamad, S. Kumar, N. Memon, Cross-platform multimodal misinformation: Taxonomy, characteristics and detection for textual posts and videos, in: Proceedings of the International AAI Conference on Web and Social Media, volume 16, 2022, pp. 651–662.
- [7] J. Kocoń, A. Janz, Propagation of emotions, arousal and polarity in wordnet using heterogeneous structured synset embeddings, in: Proceedings of the 10th Global Wordnet Conference, 2019, pp. 336–341.
- [8] J. Kocoń, M. Maziarz, Mapping wordnet onto human brain connectome in emotion processing and semantic similarity recognition, *Information Processing & Management* 58 (2021) 102530.
- [9] J. Kocoń, J. Radom, E. Kaczmarz-Wawryk, K. Wabnic, A. Zajączkowska, M. Zaśko-Zielińska, Aspectemo: multi-domain corpus of consumer reviews for aspect-based sentiment analysis, in: 2021 International Conference on Data Mining Workshops, IEEE, 2021.
- [10] J. Kocoń, J. Baran, M. Gruza, A. Janz, M. Kajstura, P. Kazienko, W. Korczyński, P. Miłkowski, M. Piasecki, J. Szolomicka, Neuro-symbolic models for sentiment analysis, in: International Conference on Computational Science, Springer, 2022, pp. 667–681.
- [11] P. Miłkowski, M. Gruza, P. Kazienko, J. Szolomicka, S. Woźniak, J. Kocoń, Multiemo: language-agnostic sentiment analysis, in: International Conference on Computational Science, Springer, 2022, pp. 72–79.
- [12] M. Wierzba, M. Riegel, J. Kocoń, P. Miłkowski, A. Janz, K. Klessa, K. Juszczak, B. Konat, D. Grimling, M. Piasecki, et al., Emotion norms for 6000 polish word meanings with a direct mapping to the polish wordnet, *Behavior Research Methods* 54 (2022) 2146–2161.
- [13] A. Schmidt, M. Wiegand, A survey on hate speech detection using natural language processing, in: Proceedings of the fifth international workshop on natural language processing for social media, 2017, pp. 1–10.
- [14] A. Sheth, V. L. Shalin, U. Kursuncu, Defining and detecting toxicity on social media: context and knowledge are key, *Neurocomputing* 490 (2022) 312–318.
- [15] S. Ghosh, A. Ekbal, P. Bhattacharyya, T. Saha, A. Kumar, S. Srivastava, Sehc: A benchmark setup to identify online hate speech in english, *IEEE Transactions on Computational Social Systems* (2022).
- [16] P. Fortuna, S. Nunes, A survey on automatic detection of hate speech in text, *ACM Computing Surveys (CSUR)* 51 (2018) 1–30.
- [17] J. Kocoń, A. Figas, M. Gruza, D. Puchalska, T. Kajdanowicz, P. Kazienko, Offensive,

- aggressive, and hate speech analysis: From data-centric to human-centered approach, *Information Processing & Management* 58 (2021) 102643.
- [18] F. Miresghallah, V. Shrivastava, M. Shokouhi, T. Berg-Kirkpatrick, R. Sim, D. Dimitriadis, Useridentifier: implicit user representations for simple and effective personalized sentiment analysis, *arXiv preprint arXiv:2110.00135* (2021).
- [19] J. Kocoń, M. Gruza, J. Bielaniewicz, D. Grimling, K. Kanclerz, P. Miłkowski, P. Kazienko, Learning personal human biases and representations for subjective tasks in natural language processing, in: *2021 IEEE International Conference on Data Mining (ICDM)*, IEEE, 2021, pp. 1168–1173.
- [20] P. Miłkowski, M. Gruza, K. Kanclerz, P. Kazienko, D. Grimling, J. Kocoń, 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*, 2021.
- [21] A. Ngo, A. Candri, T. Ferdinan, J. Kocoń, W. Korczynski, Studemo: A non-aggregated review dataset for personalized emotion recognition, in: *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022*, 2022, pp. 46–55.
- [22] K. Kanclerz, A. Figas, M. Gruza, T. Kajdanowicz, J. Kocoń, D. Puchalska, P. Kazienko, 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)*, 2021, pp. 5915–5926.
- [23] P. Miłkowski, S. Saganowski, M. Gruza, P. Kazienko, M. Piasecki, J. Kocoń, Multitask personalized recognition of emotions evoked by textual content, in: *2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, IEEE, 2022, pp. 347–352.
- [24] Y. Sang, J. Stanton, The origin and value of disagreement among data labelers: A case study of individual differences in hate speech annotation, in: *International Conference on Information*, Springer, 2022, pp. 425–444.
- [25] K. Kanclerz, M. Gruza, K. Karanowski, J. Bielaniewicz, P. Miłkowski, J. Kocoń, P. Kazienko, What if ground truth is subjective? personalized deep neural hate speech detection, in: *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022*, 2022.
- [26] A. Chatterjee, U. Gupta, M. K. Chinnakotla, R. Srikanth, M. Galley, P. Agrawal, Understanding emotions in text using deep learning and big data, *Computers in Human Behavior* 93 (2019) 309–317.
- [27] C. J. Kennedy, G. Bacon, A. Sahn, C. von Vacano, Constructing interval variables via faceted Rasch measurement and multitask deep learning: a hate speech application, *arXiv e-prints* (2020) arXiv:2009.10277. arXiv:2009.10277.
- [28] A. Rajadesingan, R. Zafarani, H. Liu, Sarcasm detection on twitter: A behavioral modeling approach, in: *Proceedings of the eighth ACM international conference on web search and data mining*, 2015, pp. 97–106.
- [29] S. Amir, B. C. Wallace, H. Lyu, P. C. M. J. Silva, Modelling context with user embeddings for sarcasm detection in social media, *arXiv preprint arXiv:1607.00976* (2016).
- [30] L. Gong, B. Haines, H. Wang, Clustered model adaption for personalized sentiment analysis, in: *Proceedings of the 26th International Conference on World Wide Web*, 2017.

- [31] A. Kamal, M. Abulaish, Self-deprecating humor detection: A machine learning approach, in: International Conference of the Pacific Association for Computational Linguistics, Springer, 2019, pp. 483–494.
- [32] A. Mondal, R. Sharma, 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), Association for Computational Linguistics, 2021, pp. 1169–1174.
- [33] N. Chetty, S. Alathur, Hate speech review in the context of online social networks, *Aggression and violent behavior* 40 (2018) 108–118.
- [34] L. Aroyo, C. Welty, Truth is a lie: Crowd truth and the seven myths of human annotation, *AI Magazine* 36 (2015) 15–24.
- [35] T. Liu, A. Venkatachalam, P. Sanjay Bongale, C. Homan, Learning to predict population-level label distributions, in: Companion Proceedings of The 2019 World Wide Web Conference, 2019, pp. 1111–1120.
- [36] S. Akhtar, V. Basile, V. Patti, Modeling annotator perspective and polarized opinions to improve hate speech detection, in: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, volume 8, 2020, pp. 151–154.
- [37] T. C. Weerasooriya, T. Liu, C. M. Homan, Neighborhood-based pooling for population-level label distribution learning, *arXiv preprint arXiv:2003.07406* (2020).
- [38] S. Tonekaboni, S. Joshi, M. D. McCradden, A. Goldenberg, What clinicians want: contextualizing explainable machine learning for clinical end use, in: Machine learning for healthcare conference, PMLR, 2019, pp. 359–380.
- [39] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, *stat* 1050 (2017) 2.
- [40] A. Lui, G. W. Lamb, Artificial intelligence and augmented intelligence collaboration: regaining trust and confidence in the financial sector, *Information & Communications Technology Law* 27 (2018) 267–283.
- [41] A. Adadi, M. Berrada, Peeking inside the black-box: A survey on explainable artificial intelligence (xai), *IEEE Access* 6 (2018) 52138–52160.
- [42] A. Das, P. Rad, Opportunities and challenges in explainable artificial intelligence (xai): A survey, *arXiv preprint arXiv:2006.11371* (2020).
- [43] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, T.-S. Chua, Neural collaborative filtering, in: Proceedings of the 26th international conference on world wide web, 2017, pp. 173–182.
- [44] S. Swayamdipta, R. Schwartz, N. Lourie, Y. Wang, H. Hajishirzi, N. A. Smith, Y. Choi, Dataset cartography: Mapping and diagnosing datasets with training dynamics, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 9275–9293.
- [45] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, *CoRR abs/1907.11692* (2019). URL: <http://arxiv.org/abs/1907.11692>. arXiv:1907.11692.