Differential Dataset Cartography: Explainable Artificial Intelligence in Comparative Personalized Sentiment Analysis[⋆]

Jan Kocoń[0000−0002−7665−6896], Joanna Baran[0000−0001−6792−7028], Kamil Kanclerz[0000−0002−7375−7544], Michał Kajstura, and Przemysław Kazienko[0000−0001−5868−356X]

Department of Artificial Intelligence
Wrocław University of Science and Technology, Wrocław, Poland
jan.kocon@pwr.edu.pl

Abstract. *Data Maps* is an interesting method of graphical representation of datasets, which allows observing the model's behaviour for individual instances in the learning process (training dynamics). The method groups elements of a dataset into *easy-to-learn*, *ambiguous*, and *hard-to-learn*. In this article, we present an extension of this method, *Differential Data Maps*, which allows you to visually compare different models trained on the same dataset or analyse the effect of selected features on model behaviour. We show an example application of this visualization method to explain the differences between the three personalized deep neural model architectures from the literature and the HumAnn model we developed. The advantage of the proposed HumAnn is that there is no need for further learning for a new user in the system, in contrast to known personalized methods relying on user embedding. All models were tested on the sentiment analysis task. Three datasets that differ in the type of human context were used: user-annotator, user-author, and user-author-annotator. Our results show that with the new explainable AI method, it is possible to pose new hypotheses explaining differences in the quality of model performance, both at the level of features in the datasets and differences in model architectures.

Keywords: differential data maps · data cartography · personalization · sentiment analysis · explainable artificial intelligence.

[⋆] This work was financed by (1) the National Science Centre, Poland, 2021/41/B/ST6/04471 (PK); (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, POIR.01.01.01-00-0288/22 (JK-S140), POIR.01.01.01-00-0615/21 (JK-MHS), POIR.04.02.00-00C002/19 (JB, KK); (4) the statutory funds of the Department of Artificial Intelligence, Wroclaw University of Science and Technology; (5) the European Union under the Horizon Europe, grant no. 101086321 (OMINO).

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-35995-8_11

1 Introduction

Popular approaches in text classification in natural language processing (NLP) assume the development of a general classifier that returns a label based on text-only input. Meanwhile, language and its perception are influenced by many factors: mood, emotions, world view, and sociodemographic conditions. This is a challenge, especially for subjective tasks such as identifying hate speech, aggression, or sentiment. In these tasks, getting a single true label for the text is already difficult at the annotation level and the inter-annotator agreement is low for the same text annotated by several people. Thus, creating a general model that always returns an answer consistent with the user's expectations is difficult.

An example of a subjective task in NLP, among others, is sentiment classification, where the polarity of a text depends on a person's experience and character, both from the perspective of the author of the text and the recipient. It happens that the same text can evoke drastically different reactions. Moreover, it can be argued that a single correct label often simply does not exist, and attempting to enforce it could lead to a model biased against a particular culture or world view, which could lead, e.g., to discrimination against minorities. Obtaining a gold standard from multiple annotations is usually done through aggregation based on instances, such as a mean or majority vote. Still, the more controversial the text, the more difficult it is to find a consensus that satisfies all users.

The aforementioned problems have led to the rapid development of personalized models in NLP, which are trained on datasets annotated by multiple people and using non-aggregated labels. The human context is considered in both the training and inference process, so one gets personalized predictions [9]. In most cases, it improves the prediction quality under the condition of providing human context. Most of the work on this topic is limited to presenting the magnitude of the performance improvement achieved. However, no one focuses on explaining what characterizes the cases for which improvement or deterioration is obtained?

In this work, we propose a new Differential Data Maps (DDM) method from the field of explainable artificial intelligence (XAI), which is an extension of the Data Maps method from the Dataset Cartography [33]. DDM visualizations can help formulate new hypotheses that better explain differences between models, including the impact of individual features and architectures. The capabilities of DDM are demonstrated with an example of personalized sentiment analysis for three datasets in which we have different human contexts.

Our contributions are as follows: (1) we developed a new DDM method that graphically presents the differences in training dynamics on the same datasets for different personalized models; (2) we developed a new HumAnn method that does not require a user ID in the personalized learning/inference process and does not need additional training for new users in the system; (3) we analysed with DDM four personalized methods, including two methods that use the user ID in the model fine-tuning process; (4) we analysed with DDM three datasets with the following human contexts: user-annotator, user-author, and user-author-annotator. (5) through DDM visualizations, we show which situa-

DOI: 10.1007/978-3-031-35995-8_11

tions personalized methods work best and where improvements and deteriorations come from, depending on how the human context is considered.

2 Background

This section briefly describes related works in the area of personalized NLP and provides motivation for the use of XAI methods within this field.

2.1 Personalization in NLP

The existing approaches to human-based NLP can be divided into two groups – based on users' metadata and on using past digital traces such as likes, ratings, or posted texts. Conceptually easier, attempts to adapt personalization in NLP tasks are based on users' metadata and their individual or social-group features. [36] use demographic variables (e.g., gender, age) directly as input into the traditional rule-based model aiming to learn gender differences between users. The demographic adaption is also introduced in the work of [13] and proves that models aware of that features outperform their agnostic counterparts. [37] design a residualized control approach by training a language model over the model's prediction errors using the sociodemographic variables only. Later, the results are combined with factor analysis.

More exploited personalization methods in the literature make advantage of digital traces left by the user. It could be published opinions or ratings. An interesting approach to group-wise sentiment classification is presented by [11]. Taking shared opinions between different people, the authors' solution introduces a non-parametric Dirichlet Process over the individualized models - one for each cluster of users. This lies at the heart of the social comparison theory that humans tend to form groups with others of similar minds and abilities. Inspired by the recommendation systems, the latent factor model can also be used to capture users' specific individuality due to different language habit [31].

The recent works mostly focus on using deep neural networks instead of classical machine learning techniques, especially on the SOTA transformer-based architectures. Those approaches often include global, shared model pre-training, and local, personalized fine-tuning. In the first phase, the model is trained on aggregated, non-personalized data, resulting in a global model unable to incorporate person-level information. After that, the shared model is fine-tuned for each user using their data. There are multiple ways of performing this step. The most basic one is to optimize the whole model, which results in a separate set of weights for each user [30], causing a significant computational and storage overhead. However, there are methods to share a single model between users and learn a unique representation for each person. This representation is combined with the text representation to produce a user-informed prediction [38, 22]. Even though these methods mitigate most of the memory-related issues, they continue to require user embedding optimization, which is easier than training the entire model. Still, they can be difficult if the number of users is large or they change

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-35995-8_11

frequently. Methods based on an aggregation of user labels [14] do not require training of user embeddings and thus can be easily applied in big-data scenarios. However, in this approach user embedding is fixed and depends only on past texts written or annotated by the user. This results in poor performance if the evaluated text sample introduces a new topic.

2.2 Explainable AI

Modern artificial intelligence (AI) methods are complex. Numerous parameters allow 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 behaviour [35, 8]. Moreover, identifying a missing part can greatly improve the effectiveness of a model [26, 20, 18, 3, 23, 34]. On the other hand, apart from scientists, there is a growing need for common 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 [1]. Explanations must make the AI algorithm expressive, improving human understanding and confidence that the model makes just and impartial decisions [6]. In addition to curiosity, the need to facilitate better results is growing, especially when the end user is the government [2].

Moreover, it is much more difficult to maintain the transparency, trust, and fairness of the personalized architecture inference process. This requires considering the impact of user context on model behaviour. To the best of our knowledge, no work on methods for analysing the performance of personalized models has been published so far.

3 Datasets

To explore the differences between baselines and personalized models, we used three datasets: (1) **Sentiment140, S140** [10] is a heuristically annotated dataset of 56,557 tweets collected using a predefined set of queries. 1,100 users do binary annotations regarding a positive or negative sentiment based on emoticons contained in the texts. In this dataset, the known user is only the author of the text (user-author). (2) **Internet Movie Database, IMDB** [7] contains 348,415 movie reviews from the IMDB database done by 54,671 users. The labels describe their sentiment in the $[1, 10]$ range. The authors of the corpus randomly selected 50, 000 movies and crawl all their reviews. During the data cleaning procedure, the creators of the dataset filtered out the users whose reviews did not contain numerical ratings and those with less than two reviews. In this dataset, the known user is the author of the text and the author of the evaluation, as it is the same person (user-author-annotator). (3) **Measuring Hate Speech, MHS** [17] consists of 135,556 annotations regarding 39,565 comments retrieved from YouTube, Twitter, and Reddit. They were annotated by 7,912 United States-based Amazon Mechanical Turk workers in various hate speech-related tasks: sentiment, disrespect, insult, humiliation, inferior status, and others. Here, we focused on sentiment analysis only. This dimension also proved to be difficult for non-personalized methods. In this dataset, the known user is a text annotator (user-annotator).

4 Personalized Architectures

The task of personalized sentiment analysis is approached from many diverse perspectives. We chose four existing methods for comparison: (1) **Baseline** is a conventional fine-tuning of pretrained RoBERTa, without including any user-specific information. (2) **UserIdentifier** [28] takes into account the identity of the text's author. A data augmentation method involves adding a sequence of 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 simpler than relying on user-specific embeddings and has been shown to achieve high performance. (3) **UserId** [21, 29] provides the person's identity by appending a user ID to the beginning of the annotated text as a special token before the training procedure. The vector representation of the text with the user ID is obtained via transformer encoding. In this model, all transformer weights are trained to learn the dependencies between the user and the text. (4) **HuBi-Medium** [19, 15, 4, 16] takes inspiration from the collaborative filtering methods. This model learns a personal latent vector for each user during the training procedure. The vector aims to model personal beliefs about the trained task. Similar to the neural collaborative filtering model [12], the personal latent vector is multiplied element-wise with the vector of the text. The resulting vector is fed to a fully connected classification layer.

5 HumAnn

Here, we introduce another personalized architecture, based on **hum**an **ann**otation (HumAnn), which does not require training the whole model for new users. HumAnn is a method combining text representations obtained from a language model, like RoBERTa, and an aggregated user-level score computed by a retrieval module. Texts previously written or annotated by the user are retrieved from the database. Then the text similarity scores are computed and used to calculate a user score representing their preferences. There are various ways of aggregating multiple labels into a single score. In the experiments, we used a KNN-based aggregation that averages the labels of the K most similar samples, where K=3. Textual features are concatenated with a user score. This personalized representation is then passed to a linear classifier for a person-informed prediction. Fig. 1 shows components of the entire system.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-35995-8_11

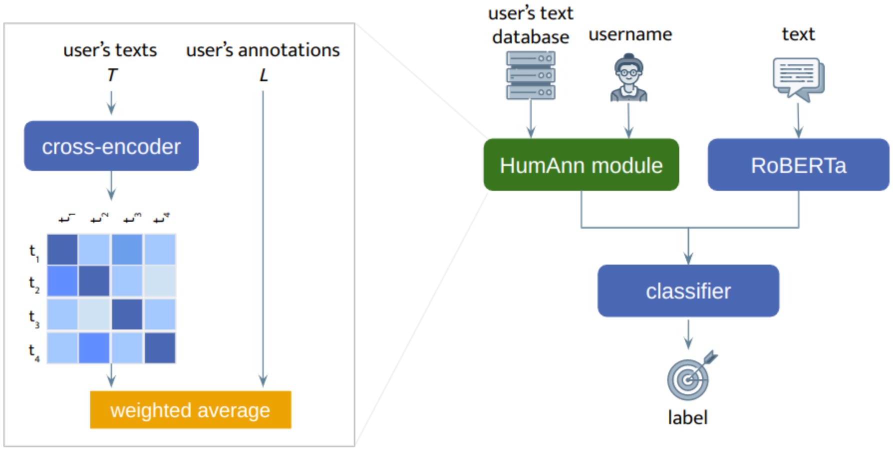

Fig. 1: HumAnn combines text representation from a language model with an aggregated user score. The average user label is weighted by a text similarity computed using a cross-encoder.

In HumAnn, text similarity scores influence the aggregation of previous users' text labels. The labels of samples most similar to the current text impact the final user score. The simplest method of aggregating multiple targets is a weighted arithmetic mean. Similarity score s between a pair of texts plays a role in weighting coefficients. Therefore, if another text is very similar to the sample being evaluated, it has a weight close to 1, and the weight of the dissimilar text is close to 0.

$$s(t_i, T, L) = \frac{1}{N-1} \sum_{n=1}^{N} \mathbb{1}_{n \neq i} \cdot l_n \cdot \text{similarity}(t_i,\, t_n) \tag{1}$$

Where s denotes similarity score, t_i is the text currently predicted, T is a sequence of all user's texts, and L is a sequence of all user's labels. The last two are retrieved from a database of past annotated texts, where N ones belong to the specific user. Similarly, for the KNN-based method, only the K most similar samples are considered during label aggregation.

During the training stage, all the user's texts and labels from the training set, apart from the currently used sample t_i, are utilized to compute an aggregated score. Also, for validation and testing, the method considers only the labels of training examples, preventing data leakage. The model is trained to minimize a standard cross-entropy loss for classification concerning a single, shared parameter set θ.

$$\mathcal{L}_{\text{CE}}(t_i, y_i, T, L; \theta) = -logPr(y_i | [t_i; s(t_i, T, L)]) \tag{2}$$

$$\theta = argmin_{\theta} \mathcal{L}_{\text{CE}}(t_i, y_i, T, L; \theta) \tag{3}$$

where y_i denotes the class of ith text example and θ - model parameters.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-35995-8_11

6 Differential Data Maps

We propose a novel XAI method for personalized models. The idea was inspired by work [33]. 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 ith 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 E:

$$corr = \frac{\sum_{e=1}^{E}(y_{pred} = y_{true})}{E} \tag{4}$$

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 analyse 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 ith 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**: $\hat{c}_i^C = \hat{c}_i^{M2} - \hat{c}_i^{M1}$; where $M1$ is the model whose measures we want to obtain compared to the base model $M2$.

7 Experimental Setup

We evaluated the proposed personalized methods on three sentiment datasets presented in Section 3. We used the data split methodology described in the following papers: S140 [24], IMDB [38] and MHS [17]. The datasets contain metadata on the context in which the annotation process occurred. The type of information on users makes it possible to point out some differences between the collections. The MHS dataset provides the most detailed description of the texts' annotators, such as ID, gender, education, income, or severity, but there is no information on the authors of the texts. In contrast, S140 provides information on the texts' authors (namely, the nicknames of the users that tweeted), but there

DOI: 10.1007/978-3-031-35995-8_11

is no information about the annotators. In IMDB, the author and annotator of the text in IMDB are the same person, and we know his or her ID.

We used the RoBERTa-base language model [25] as a baseline and in personalized approaches. For text similarity calculations in HumAnn, we utilized MPNet-based Bi-Encoder, trained on various text-pair datasets [32]. The Cross-Encoder was based on a RoBERTa trained on Semantic Text Similarity Benchmark [5]. Models were fine-tuned using AdamW optimizer with learning rate 1e-5, linear warm-up schedule, batch size 16, and maximum sequence length 512 for 50000 training steps, and the best model was selected according to the validation F-score. For the KNN-based aggregation in HumAnn, K was set to 3. For UserIdentifier, we use 10 tokens drawn from the tokenizer vocabulary as an identifier. This has been shown to enable better differentiation between users than relying on usernames or strings of numbers. Experiments were repeated five times, and the mean F1 score was reported. We conducted statistical tests to measure the significance of differences between each method's performance. Firstly, we checked the assumptions of the t-test for independent samples and conducted it to determine if they were met. Otherwise, the Mann-Whitney U test was used.

8 Results

	S140		IMDB		MHS	
	F1	Acc	F1	Acc	F1	Acc
Baseline	85.9±0.3	85.1±0.2	43.8±0.8	41.1±1.2	58.1±1.2	48.4±0.7
UserIdentifier	87.4±0.4	86.7±0.4	**47.0**±2.0	**44.5**±1.9	58.8±0.6	48.5±0.7
UserId	85.2±0.4	86.2±0.2	45.2±1.4	41.8±0.7	59.2±0.6	48.6±0.7
HuBi-Medium	**88.9**±0.2	83.8±0.4	43.3±0.6	42.3±2.0	**61.3**±0.5	48.3±0.6
HumAnn	87.0±0.2	**88.2**±0.2	44.0±0.8	40.5±1.0	58.5±0.8	**51.2**±0.9

Table 1: **F1**-macro and **Acc**uracy reported for methods. **Bold** values indicate the best performance among all architectures.

A comparison of the methods on the three datasets is shown in Table 1. For each dataset, at least one personalized method achieves significantly better results than the baseline. Based on the results, it can be determined that there is no one-size-fits-all architecture. Similarly, depending on the measure, some methods may be better or worse within a particular dataset. In the case of the S140 and MHS datasets, for the F1-macro measure the best method is HuBi-Medium, and for the accuracy measure it is HumAnn. For the IMDB collection, for both quality measures, the best method is UserIdentifier. It is worth noting that the methods belong to two different groups: 1) those that require training

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-35995-8_11

the entire model (including finetuning the language model) for new users (UserIdentifier, UserID); 2) those that only require training the representation of the new user (HuBi-Medium, HumAnn). In addition, the datasets represent three different human contexts: 1) user-author (S140), 2) user-annotator (MHS), 3) user-author-annotator (IMDB). From this perspective, it can be assumed that for datasets with a context limited only to the annotator (MHS) or to the author (S140), methods in which the user representation is separated from the language model (HuBi-Medium, HumAnn) are preferable.

It is important to note that the sentiment analysis task is much less subjective than the hate speech detection or emotion recognition tasks [22, 27], for which personalized methods from the literature achieved much better quality gains when annotator context was added relative to the baseline. For some emotions, up to 40 pp of improvement was reported, while for sentiment analysis, the quality gains for F1 and Acc measures are, respectively: 3pp and 3.1pp on S140, 3.2pp and 3.4pp on IMDB, and 3.2pp and 2.8pp on MHS.

The differences in the results of personalized models may not be large for F1 and Acc measures. However, much more interesting conclusions come from analysing differences using Data Cartography. Fig. 2 shows the results of the original Data Maps (DM) method for the samples (text, annotation pair), Fig. 3 shows the results of the Differential Data Maps (DDM) method for the samples as well, while Fig. 4 presents the DDM for the users (results for the samples aggregated by user ID) and in both of these Figures the points in each quadrant of the graph are counted. Additionally, in Fig. 4, instead of correctness change, the entropy of user annotation in the set is presented as the colour of a point on the map. Each figure presents results for five architectures in the rows (i.e., baseline and four personalized) and three datasets in the columns. In the case of DM, these are data maps for the pair (model, dataset), and in the case of DDM, these are differential data maps for the pair (baseline-personalized model, dataset). The first interesting observation is that within each set there happen to be data maps for personalized models very similar to the baseline (Fig. 2), for which there are very different ones for differential data maps (Fig. 3 and 4). This means that calculating differences for training dynamic measures gives additional information about differences in the behaviour of models that are not visible at first glance by comparing only the data maps themselves.

Intuitively, it might seem that an increase in the quality of a personalized model relative to the baseline should be associated with a decrease in variability and an increase in confidence for most samples. However, in only three of the six cases among the best models relative to baseline is such a trend observed (S140/HuBi-Medium/F1, MHS/HuBi-Medium/F1, MHS/HumAnn/Acc). In other cases, there is a decrease in confidence and an increase in variability for most samples (S140/HumAnn/Acc, IMDB/UserIdentifier/F1, IMDB/UserIdentifier/Acc). Much more interesting is the observation for DDMs aggregated by user. For example, for the S140 set, for models based on user ID confidence increases significantly more for authors of similarly rated texts (with low entropy of ratings). For the second group of models, the opposite is true, i.e., these models do better

DOI: 10.1007/978-3-031-35995-8_11

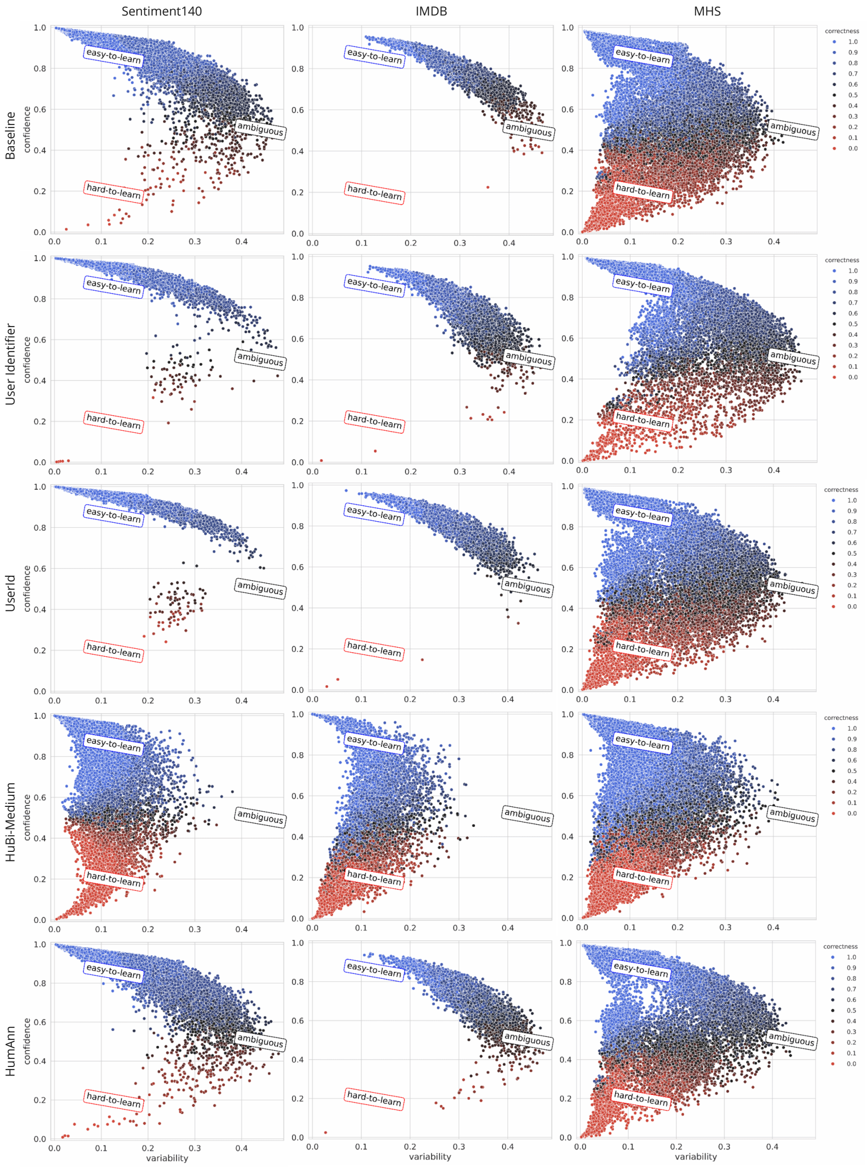

Fig. 2: Results of the original Data Maps method for data samples.

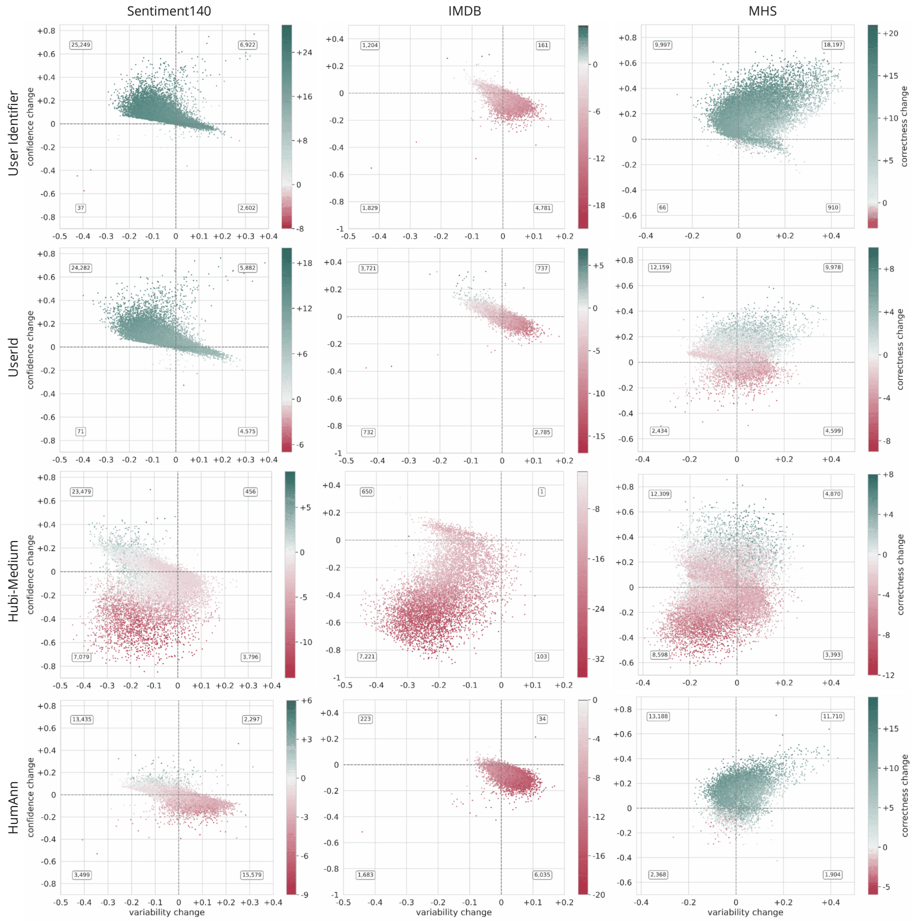

Fig. 3: Results of the Differential Data Maps for data samples.

DOI: 10.1007/978-3-031-35995-8_11

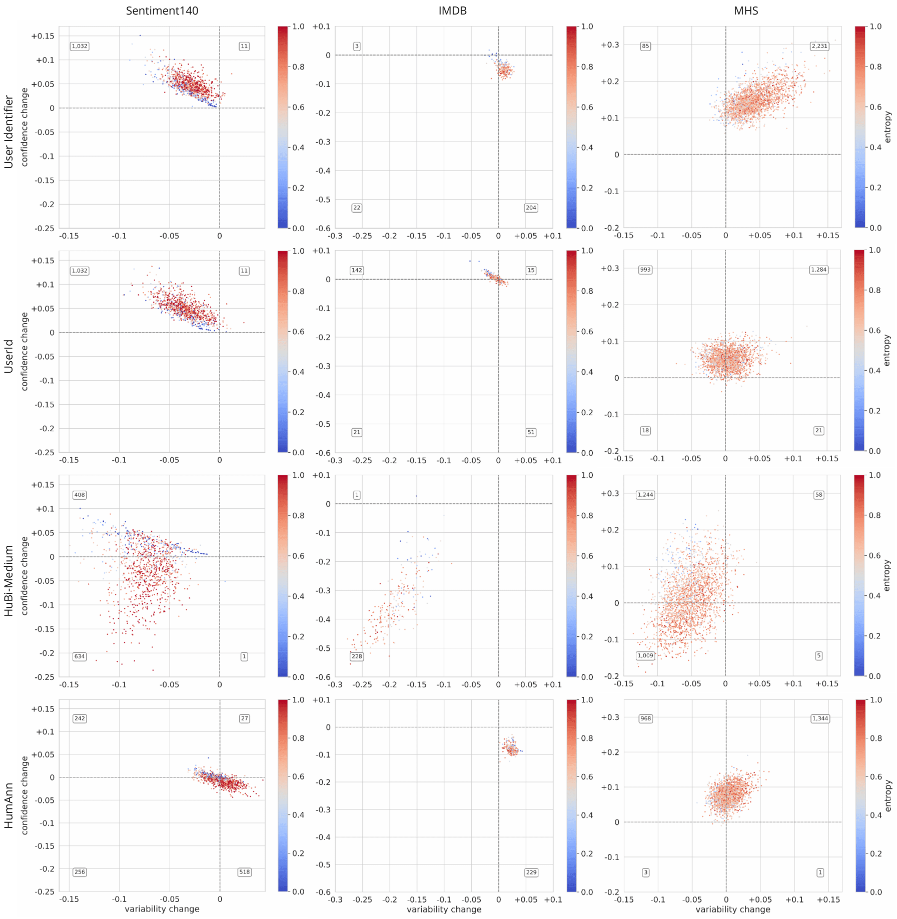

Fig. 4: Results of the Differential Data Maps for data samples aggregated by users.

with more diverse people and better model the complex combination of text and human contexts. Models based on user ID better reflect users who behave similarly.

Even the very similar models on the DM, i.e. UserID and UserIdentifier, behave differently when the DDM is analyzed. They show significant similarity only on the S140 set, while on the IMDB and on the MHS there are no analogous trends regarding the increase or decrease in confidence and variability. Much greater similarity is observed for users with a certain entropy, then, regardless of the dataset, the trends are very similar, i.e. the smallest increase in confidence for users with the lowest entropy on the S140 set, an inverse relationship for the IMDB and no relationship for the MHS.

9 Conclusions and Future Work

In the article, we presented a new Differential Data Maps method that can be used to draw more complex conclusions about either datasets or, more importantly, differences between models. We have presented a few examples of such findings, but further analysis could identify more interesting insights that are not apparent from DM analysis and F1/Acc values.

Our proposed HumAnn model proved to perform on par with other personalized SOTA approach which uses specially trained unique human representation. However, choosing text similarity between past users' written opinions has one major advantage over other methods – frequent retraining is unnecessary. This makes HumAnn easier to deploy in real-world applications. The only limitation of the model is the need to have a certain number of texts from a single person to predict the label in a subjective task correctly. In future, we plan to further train the cross-encoder part of the model to provide similarity scores of even higher quality. Some optimization techniques to reduce compute overhead should also be applied.

The next step in further work will also be to analyse such datasets on which the differences between baseline and personalized models are even greater. This will allow us to further understand for which types of samples the different models perform better and for which worse.

References

1. Adadi, A., Berrada, M.: Peeking inside the black-box: A survey on explainable artificial intelligence (xai). IEEE Access **6**, 52138–52160 (2018)
2. AI, H.: High-level expert group on artificial intelligence (2019)
3. Baran, J., Kocoń, J.: Linguistic knowledge application to neuro-symbolic transformers in sentiment analysis. In: 2022 IEEE International Conference on Data Mining Workshops (ICDMW). pp. 395–402. IEEE (2022)
4. Bielaniewicz, J., Kanclerz, K., Miłkowski, P., Gruza, M., Karanowski, K., Kazienko, P., Kocoń, J.: Deep-sheep: Sense of humor extraction from embeddings in the personalized context. In: 2022 IEEE International Conference on Data Mining Workshops (ICDMW). pp. 967–974. IEEE (2022)

5. Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., Specia, L.: SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In: Proc. of the 11th Workshop on Semantic Evaluation (SemEval-2017) (2017)
6. Das, A., Rad, P.: Opportunities and challenges in explainable artificial intelligence (xai): A survey. arXiv preprint arXiv:2006.11371 (2020)
7. Diao, Q., Qiu, M., Wu, C.Y., Smola, A.J., Jiang, J., Wang, C.: Jointly modeling aspects, ratings and sentiments for movie recommendation (JMARS). In: Proc. 20th ACM SIGKDD conference on Knowledge discovery and data mining (2014)
8. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. stat **1050**, 2 (2017)
9. Flek, L.: Returning the N to NLP: Towards contextually personalized classification models. In: Proceedings of the 58th Annual Meeting of ACL (2020)
10. Go, A.: Sentiment classification using distant supervision (2009)
11. Gong, L., Haines, B., Wang, H.: Clustered model adaption for personalized sentiment analysis. In: Proceedings of the 26th Conference on World Wide Web (2017)
12. He, X., Liao, L., Zhang, H., Nie, L., Hu, X., Chua, T.S.: Neural collaborative filtering. In: Proceedings of the 26th conference on world wide web (2017)
13. Hovy, D.: Demographic factors improve classification performance. In: Proceedings of the 53rd Annual Meeting of the ACL&IJCNLP (2015)
14. Kanclerz, K., Figas, A., Gruza, M., Kajdanowicz, T., Kocoń, J., Puchalska, D., Kazienko, P.: Controversy and conformity: from generalized to personalized aggressiveness detection. In: Proceedings of the 59th Annual Meeting of the ACL&IJCNLP (2021)
15. Kanclerz, K., Gruza, M., Karanowski, K., Bielaniewicz, J., Miłkowski, P., Kocoń, J., Kazienko, P.: What if ground truth is subjective? personalized deep neural hate speech detection. In: Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022. pp. 37–45 (2022)
16. Kazienko, P., Bielaniewicz, J., Gruza, M., Kanclerz, K., Karanowski, K., Miłkowski, P., Kocoń, J.: Human-centred neural reasoning for subjective content processing: Hate speech, emotions, and humor. Information Fusion (2023)
17. Kennedy, C.J., Bacon, G., Sahn, A., von Vacano, C.: Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application. arXiv preprint arXiv:2009.10277 (2020)
18. Kocoń, J., Baran, J., Gruza, M., Janz, A., Kajstura, M., Kazienko, P., Korczyński, W., Miłkowski, P., Piasecki, M., Szołomicka, J.: Neuro-symbolic models for sentiment analysis. In: Computational Science–ICCS 2022: 22nd International Conference, London, UK, June 21–23, 2022, Proceedings, Part II. pp. 667–681. Springer (2022)
19. Kocoń, J., Gruza, M., Bielaniewicz, J., Grimling, D., Kanclerz, K., Miłkowski, P., Kazienko, P.: Learning personal human biases and representations for subjective tasks in natural language processing. In: 2021 IEEE International Conference on Data Mining (ICDM). pp. 1168–1173. IEEE (2021)
20. Kocoń, J., Maziarz, M.: Mapping wordnet onto human brain connectome in emotion processing and semantic similarity recognition. Information Processing & Management **58**(3), 102530 (2021)
21. Kocoń, J., Figas, A., Gruza, M., Puchalska, D., Kajdanowicz, T., Kazienko, P.: Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach. Information Processing & Management **58**(5) (2021)
22. Kocoń, J., Gruza, M., Bielaniewicz, J., Grimling, D., Kanclerz, K., Miłkowski, P., Kazienko, P.: Learning personal human biases and representations

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-35995-8_11

for subjective tasks in natural language processing. In: 2021 IEEE International Conference on Data Mining (ICDM). pp. 1168–1173 (2021). https://doi.org/10.1109/ICDM51629.2021.00140
23. Korczyński, W., Kocoń, J.: Compression methods for transformers in multidomain sentiment analysis. In: 2022 IEEE International Conference on Data Mining Workshops (ICDMW). pp. 419–426. IEEE (2022)
24. Li, T., Sanjabi, M., Smith, V.: Fair resource allocation in federated learning. CoRR **abs/1905.10497** (2019)
25. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized BERT pretraining approach. CoRR **abs/1907.11692** (2019)
26. Lui, A., Lamb, G.W.: Artificial intelligence and augmented intelligence collaboration: regaining trust and confidence in the financial sector. Information & Communications Technology Law **27**(3), 267–283 (2018)
27. Miłkowski, P., Saganowski, S., Gruza, M., Kazienko, P., Piasecki, M., Kocoń, J.: Multitask personalized recognition of emotions evoked by textual content. In: Pervasive Computing and Communications Workshops (2022)
28. Mireshghallah, F., Shrivastava, V., Shokouhi, M., Berg-Kirkpatrick, T., Sim, R., Dimitriadis, D.: Useridentifier: Implicit user representations for simple and effective personalized sentiment analysis. CoRR **abs/2110.00135** (2021)
29. Ngo, A., Candri, A., Ferdinan, T., Kocoń, J., Korczynski, W.: Studemo: A non-aggregated review dataset for personalized emotion recognition. In: Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022. pp. 46–55 (2022)
30. Schneider, J., Vlachos, M.: Mass personalization of deep learning. CoRR **abs/1909.02803** (2019)
31. Song, K., Feng, S., Gao, W., Wang, D., Yu, G., Wong, K.F.: Personalized sentiment classification based on latent individuality of microblog users (07 2015)
32. Song, K., Tan, X., Qin, T., Lu, J., Liu, T.: Mpnet: Masked and permuted pre-training for language understanding. CoRR **abs/2004.09297** (2020)
33. Swayamdipta, S., Schwartz, R., Lourie, N., Wang, Y., Hajishirzi, H., Smith, N.A., Choi, Y.: Dataset cartography: Mapping and diagnosing datasets with training dynamics. In: Proc. of the EMNLP2020 (2020)
34. Szołomicka, J., Kocon, J.: Multiaspectemo: Multilingual and language-agnostic aspect-based sentiment analysis. In: 2022 IEEE International Conference on Data Mining Workshops (ICDMW). pp. 443–450. IEEE (2022)
35. Tonekaboni, S., Joshi, S., McCradden, M.D., Goldenberg, A.: What clinicians want: contextualizing explainable machine learning for clinical end use. In: Machine learning for healthcare conference. pp. 359–380. PMLR (2019)
36. Volkova, S., Wilson, T., Yarowsky, D.: Exploring demographic language variations to improve multilingual sentiment analysis in social media. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (2013)
37. Zamani, M., Schwartz, H.A., Lynn, V.E., Giorgi, S., Balasubramanian, N.: Residualized factor adaptation for community social media prediction tasks (2018)
38. Zhong, W., Tang, D., Wang, J., Yin, J., Duan, N.: Useradapter: Few-shot user learning in sentiment analysis. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. pp. 1484–1488 (2021)