Towards Model-Based Data Acquisition for Subjective Multi-Task NLP Problems

Kamil Kanclerz, Julita Bielaniewicz, Marcin Gruza,
Jan Kocoń, Stanisław Woźniak, and Przemysław Kazienko
Department of Artificial Intelligence, Wrocław University of Science and Technology, Poland
{kamil.kanclerz, julita.bielaniewicz, marcin.gruza,jan.kocon,
stanislaw.wozniak, kazienko}@pwr.edu.pl

***Abstract*—Data annotated by humans is a source of knowledge by describing the peculiarities of the problem and therefore fueling the decision process of the trained model. Unfortunately, the annotation process for subjective natural language processing (NLP) problems like offensiveness or emotion detection is often very expensive and time-consuming. One of the inevitable risks is to spend some of the funds and annotator effort on annotations that do not provide any additional knowledge about the specific task. To minimize these costs, we propose a new model-based approach that allows the selection of tasks annotated individually for each text in a multi-task scenario. The experiments carried out on three datasets, dozens of NLP tasks, and thousands of annotations show that our method allows up to 40% reduction in the number of annotations with negligible loss of knowledge. The results also emphasize the need to collect a diverse amount of data required to efficiently train a model, depending on the subjectivity of the annotation task. We also focused on measuring the relation between subjective tasks by evaluating the model in single-task and multi-task scenarios. Moreover, for some datasets, training only on the labels predicted by our model improved the efficiency of task selection as a self-supervised learning regularization technique.**

***Index Terms*—natural language processing, personalization, self-supervised learning, data acquisition, model-based annotation optimization**

I. Introduction

One of the most crucial parts of developing any machine learning solution is the data acquisition process. A well-prepared dataset will significantly increase the amount of knowledge obtained by the model during training. However, in most cases, obtaining a high-quality dataset includes a large annotation process and further data post-processing, including filtering out a considerable part of the dataset. In this way, the financial expenses and time required to collect the filtered data are wasted. The most common approaches to tackle this problem focus on maximizing the inter-annotator agreement or selection of texts, which should further improve the model performance. However, these methods assumed the existence of only one true label for a text. Additionally, data annotation optimization techniques allowed for including or fully omitting the specific text during the annotation process. This led to the loss of valuable knowledge that could be extracted from the subset of labels for the omitted text. In addition, some labels of texts selected for the annotation procedure may not provide additional knowledge to the model and be just a waste of time and money.

To the best of our knowledge, currently, there are no methods designed for subjective multi-task NLP tasks, which focus on the above issues. Therefore, we present our novel model-based data acquisition strategy, which operates on the level of individual labels and allows the user to annotate only a subset of labels for a specific text while providing automatic annotations for the rest of the labels. To evaluate our method, we performed a complex evaluation that included several experimental scenarios. We also developed our own measures and used them alongside commonly used ones to better verify the effectiveness and reliability of the proposed technique.

The main contributions of this work are as follows: (1) we proposed a novel model-based data acquisition optimization strategy focused on reducing the annotation effort by predicting the valuable labels for each text resulting in up to 25% benefit and up to 40% reduction of the annotation effort (Fig. 1); (2) we applied our method on three datasets regarding personalized multi-task NLP problems; (3) we developed new evaluation metrics appropriate for the problem and leveraged them along with the standard ones like macro F1-score; (4) our evaluation included the self-supervised scenario, where the model was trained only on labels previously predicted by itself to measure the amount of knowledge not learned during the training procedure; (5) we also tested the impact of training dataset size on the model performance; (6) we analyzed the knowledge transfer between tasks in single-task and multi-task scenario; (7) we also analyzed the relation between the number of annotations per single text, number of unique texts in the dataset, and the model performance; (8) we conducted the evaluation of the personalized architecture for multi-task subjective problems.

II. Related Work

Ever since the very beginning of research related to artificial intelligence, there has been a consistent series of issues regarding data acquisition. A crucial part of the said possible hardships include estimating the number of information needed for a reasonable analysis and, of course, the costs needed for obtaining it. There is no doubt that without adequate quality and quantity of data, we will certainly omit important information regarding relations between data. Moreover, such miscalculation may lead to false conclusions that will distort

2375-9259/23/$31.00 ©2023 IEEE
DOI 10.1109/ICDMW60847.2023.00099

the message of a number of studies and, consequently, the entire area of study. Such oversight must be prevented by all means, and thus researchers must find the balance between quality and expense.

A. Data Acquisition

There are a considerable number of studies that specialize in the area of data acquisition, most of which revolve around the generalization approach to data analysis. The work [1] foresees an accurate image of the current state of data acquisition. Managing large amounts of data comes down to counting losses we are ready to sacrifice, and finding balance is an approach where the authors reduced the size of a training test with the cost of decreasing variance. Cantrell [2] suggests that online data collection methods provide an advantage in the possibility of using online tools, but in his works it is not reconsidered whether the costs of the quantity approach may have been avoided to some extent.

B. Dataset Distillation

An interesting measure of quality was introduced in the work [3] that signals the level of quality of a chosen dataset. This creates an opportunity to calculate the metric each time we reduce the amount of data; however, when faced with ad hoc analysis, there is a strong bias to the quantity that overshadows the data quality. The article [4] presents an empirical view on the matter, implying that certain fields in which data are collected are burdened with the natural impossibility of acquiring quality datasets. Thus, since data quality tools are as good as the data collected, it prevents certain areas of study from having a proper analysis. This approach requires the use of a set of metrics that can guarantee high-quality, considerable quantity data, regardless of the field of study. In the book [5] there is a carefully conducted analysis of possible approaches to dataset management and methods for extracting quality data from available sources. An article [6] is especially interesting because it provides insight into the data acquisition process used in large corporations. Although efficient, they seem to be too strict in avoiding quantity bias, discarding a lot of useful data in the process. An approach presented in the article [7] aims to receive a distilled version of datasets through the use of descending gradient and different initialization techniques. Although promising, it only works on very simple datasets, struggling when faced with multidimensional scenarios. D. Barrett [8] argues that improving data collection techniques is the key to having a distilled data set from the beginning. This study aims to focus on certain methods for obtaining quality data, which is clearly an engaging process, but we disagree with omitting acquisition beyond the data collection process. In the work [9] the authors perform a dataset distillation using tailored algorithms applied to convolutional architectures, which results in interesting enhancements of the distilled data, but additional analyses and explorations would be needed to provide insight into the full potential of the presented methods. The article [10] examines the task of data acquisition by conducting a systematic review of publicly available datasets for the detection of abusive content, focusing on improvements in training datasets when distilling data. The authors of article [11] focus on the multimodal nature of the distillation process and try to find the right balance for techniques that specialize in multidomain problems. This approach performs well in the multi-task approach, but it comes with the cost of lower performance values per modality. The data acquisition researched in the article [12] aimed to receive a distilled dataset with reduced memory size and an improved training time using feature regression. The intriguing research proves to generalize well on different types of image datasets, but would unfortunately not work on the textual data. The work [13] introduces a data distillation technique that utilizes factorization in order to separate the dataset into two, analyzing groups of hallucination values and the base values. One of the key advantages of this approach is the small number of hyperparameters needed for good results, but the method does not perform well if the time factor is crucial. The article [14] provides a summary of the dataset distillation-based solutions to deep learning tasks with a focus on quality measurement after distillation. The authors do come to the conclusion that massive image datasets for image classification are vastly optimized for acquisition purposes, but the same assumption cannot be applied to the textual datasets. When optimizing the initial and target network parameters for large-scale datasets, the authors of the work [15] compute and store training trajectories of expert networks. Although it outperforms many available methods, it comes with the cost of additional computational costs that may not be possible to achieve for many researchers. The authors of the work [16] present a formal framework for data distillation, along with providing a detailed taxonomy of existing approaches with respect to multi-task data. Although many advanced methods for annotation process optimization have been developed in the field of computer vision, no similar advances have been made in the field of natural language processing, which implies a promising future for the area of textual data acquisition. The article [17] proposes a student-teacher network that participated in the data acquisition process in long-tailed scenarios. The introduced framework benefits from sample diversity and learns generalized representation, which may indicate the possible area of personalization tasks to be enhanced in the future. The authors of article [18] propose a straightforward filtering strategy that significantly reduces the size of the dataset and achieves improved performance across zero-shot vision-language tasks. The prominent disadvantage can be noticed in noisy datasets, as there is a certain loss of performance when faced with preprocessed datasets.

C. Label Distillation

Another promising work [19] implies that it is the labels that should be distilled instead of the data. The authors of the article focus on the crafting of synthetic labels for arbitrarily chosen standard data, which works in analyzed research, but can fail to perform in a different area of study than the examined one. The work [20] investigates a new crowd counting task in an incremental domain training setting using a single model updated by the incremental domains. The method is interesting and performs well when not dealing with missing annotations, as this scenario heavily burdens the model, resulting in a possible decline in performance. The authors of the survey [21] study the research scope for data collection and data quality primarily for deep learning applications with a special focus on bias and fairness of data distillation. It is especially emphasized that noisy or missing labels cause poor generalization of the test data and this implies the potential for personalization research.

D. Model-based Techniques

The authors of the article [22] propose model-based self-supervised self-distillation methods that extract representations of the target dataset and generate pseudo labels through clustering. The downside of this technique is mostly the lack of adaptability for textual data, otherwise very promising in future directions of data acquisition.

E. Research Gap

When it comes to textual data, there is a prominent lack of research when facing the data acquisition process, especially in the domain of personalization [23]–[36]. In this work, we have focused on the matter of developing a set of measures that help evaluate the overall quality of textual data after the distillation process, which also performs well in the personalization scenario.

III. Model-Based Data Acquisition for Subjective NLP Problems

A. Subjective NLP Tasks

The variety of problems in the field of natural language processing is generously wide, so much so that in the perspective of just the group of subjective problems, there is an area of many possibilities for an individual understanding of the many of said perspectives are still yet to be discovered, as each person can interpret a single information very differently. The usual approach in the dominant number of studies across all NLP fields focuses mainly on the majority of annotators. This not only discriminates against people who do not tend to follow trends, but also excludes a significant amount of useful data that could otherwise broaden the perspective of certain peculiarities and tendencies, as presented in Fig. 2. In our case, we simultaneously consider multiple subjective problems (labels) such as emotions, offensiveness, irony, and humor for each text, up to 23 labels (Sec. IV-A). This approach not only implies a wider range of analysis per user, but also hints at a much broader grasp of relations between each dimension.

B. Problem Description

In almost all subjective tasks in NLP, there are many texts being collected and found by all or almost all people irrelevant to the problem, e.g. *not funny* in humor detection. This problem is even more crucial if we want to annotate texts simultaneously for many not related problems (multi-task), e.g. *sadness* and *funniness*. If a given text d is annotated by many humans as *sad*, it is unlikely to be annotated by a significant fraction of the same group of people as *funny*. It means that asking for annotation of d with *funny* would provide only or almost only neutral labels (zeros). Then, we should avoid such useless annotations. To identify such cases, we have developed a model-based procedure for data acquisition, Sec. III-D. Having the labeled data acquisition model, we can estimate its quality using appropriate measures (Sec. III-E - III-H). We also exploited the received labeled data in the real models solving subjective problems in a personalized and multi-task setup, i.e. prediction of all labels $l \in L$ for a given text d and individual reader (Sec. V-B).

C. Valuable Text Label (VTL)

To estimate how valuable the label l (problem, task) should be considered for text d, we developed a new measure called the Valuable Text Label (VTL). It divides the number of non-zero annotations ($a \notin \{0\}$) by the number of all annotations $|A_{d,l}|$ with respect to the specified label l in the context of a certain text d. The calculated value is then compared with the threshold t, which takes values in the range $[0, 1]$. The $VTL(d, l, t)$ measure returns 1, if the computed value is greater than or equal to the threshold t, and 0 otherwise. VTL quantifies whether the text d has the potential to be suitable for at least t percentage of the population to find it relevant to the task l. For example, text d that no one or very few people find *funny* should not be annotated with the $l = funny$ label. We assume that collecting labels for text d with only or almost only $l = 0$ does not provide any useful knowledge. The measure value is calculated as follows:

$$VTL(d,l,t) = \begin{cases} 1, & \text{if } \frac{\sum_{a \in A_{d,l}} \mathbb{1}_{\{a \notin \{0\}\}}}{|A_{d,l}|} \geq t \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

D. Model-Based Data Acquisition

To leverage the possibilities of deep neural architectures in the data acquisition process, we developed model-based data acquisition that is used to annotate texts and identify those of them that are valuable for further human annotations, Fig. 1.

We start with the preprocessing of the relatively small initial set of multiple annotations previously collected from humans. It is used to train our model and quantify how valuable each label l is in the context of the specific text d. For that case, we use the $VTL(d, l, t)$ measure described in Sec. III-C. The output $\hat{y}$ of our model is the predicted value of $= VTL(d, l, t)$ for a given text d and task l that is directly used for model-based annotations or pre-selection for human annotations.

After label preprocessing, we train the deep neural network on the texts annotated with the VTL value for each label. The obtained model is used to predict the values of our metric for each new text (candidates). The VTL values received from the model are used to decide whether the label should be annotated by humans ($\hat{y} \in \{1\}$) or should it be done automatically by the model ($\hat{y} \in \{0\}$). If the model recommends human annotations, i.e., its predicted value of $VTL(d, l, t) = 1$,

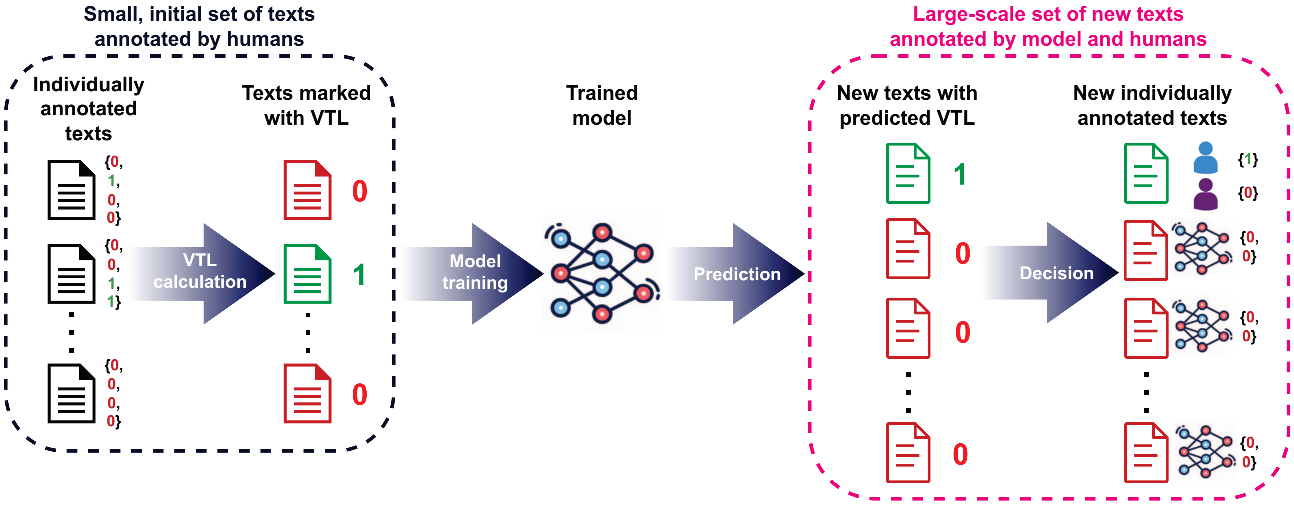
Small, initial set of texts annotated by humans
Individually annotated texts
{0, 1, 0, 0}
{0, 0, 1, 1}
{0, 0, 0, 0}
VTL calculation
Texts marked with VTL
0
1
0
Model training
Trained model
Prediction
Large-scale set of new texts annotated by model and humans
New texts with predicted VTL
1
0
0
0
Decision
New individually annotated texts
{1}
{0}
{0, 0}
{0, 0}
{0, 0}

Fig. 1. Our model-based labeled data acquisition schema with the use of the Valuable Text Label (VTL) metric, presented for a single label (one subjective NLP problem/task).

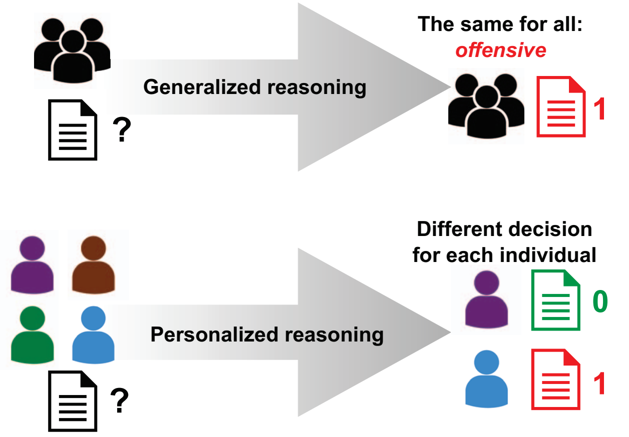

Fig. 2. Difference between the generalized and personalized approach in offensiveness detection.

then a given text d is labeled by all humans for the task l to capture individual peculiarities in human text perception (subjectivity of the task), i.e., expected controversy of the text and conformity of users [37].

In other words, if the model predicts $VTL(d, l_1, t) = 0$ and $VTL(d, l_2, t) = 1$, then the text d is annotated by the model with $l_1 = 0$ for all users, and the annotators manually annotate d with the task l_2. The lack of human annotation of d with l_1 is our gain, as $l_1 = 0$ has been achieved without any human involvement. In this way, we are able to select texts that are more relevant to individual tasks, which is very important in multi-task scenarios[1].

E. Annotation Effort Reduction (AER)

Our novel model-based method assumes that only valuable text labels are worth being annotated by annotators. This means that labels l, which are invaluable for a specific text d according to the threshold t, should be automatically marked with the $\{0\}$ class and skipped during the annotation process with real users. To efficiently measure the effort reduction caused by our approach, we developed the Annotation Effort Reduction (AER) metric. It counts invaluable labels ($VTL(d, l, t) \in \{0\}$), which were correctly predicted by our model ($VTL(d, l, t) = \hat{y}_{d,l}$) across all labels $l \in L$ and all texts $d \in D$. The calculated value is then divided by the number of all possible annotations, which is equal to the number of texts in the dataset ($|D|$) multiplied by the number of all possible labels ($|L|$). The result of the division is the percentage amount of the reduced annotation effort. The exact formula used for computing $AER(D, \hat{y}, t)$ is presented in Eq. 2.

$$AER(D, \hat{y}, t) = \frac{\sum_{d \in D} \sum_{l \in L} \mathbb{1}_{\{VTL(d,l,t) \in \{0\} \wedge VTL(d,l,t) = \hat{y}_{d,l}\}}}{|D| * |L|} \tag{2}$$

F. Absolute Annotation Loss (AAL)

Relying on the entire process on whether a specific label l is valuable for text d only on the model predictions $\hat{y}$ carries the risk of skipping labels that may in fact turn out to be important. To measure the possible loss of useful labels, we developed the Absolute Annotation Loss (AAL) metric. It calculates the number of valuable labels ($VTL(d, l, t) \in \{1\}$), for which the predictions of our model were wrong ($VTL(d, l, t) \neq \hat{y}_{d,l}$). This operation is conducted across all labels $l \in L$ and on all texts $d \in D$. Then, the computed value is divided by the

[1] See Sec. IV-A with our Doccano dataset that contains 23 simultaneously acquired tasks.

number of all important labels in the dataset ($VTL(d,l,t) \in \{1\}$). The formula for calculating $AAL(D,\hat{y},t)$ is following:

$$AAL(D,\hat{y},t) = \frac{\sum_{d\in D}\sum_{l\in L}\mathbb{1}_{\{VTL(d,l,t)\in\{1\}\wedge VTL(d,l,t)\neq\hat{y}_{d,l}\}}}{\sum_{d\in D}\sum_{l\in L}\mathbb{1}_{\{VTL(d,l,t)\in\{1\}\}}} \tag{3}$$

G. Mean Label Rarity Annotation Loss (MLRAL)

To measure the possible loss of information caused by our model with respect to the distribution of each label, we propose the Mean Label Rarity Annotation Loss (MLRAL) metric. In the first step, it calculates the Label Annotation Loss (LAL) for each of the possible labels l separately. The $LAL(D,l,\hat{y},t)$ computes the percentage value of valuable labels ($VTL(d,l,t) \in \{1\}$), for which the model made the incorrect decision ($VTL(d,l,t) \neq \hat{y}_{d,l}$). The calculated value is further divided by the number of all samples $d \in D$, for which the specific label l is considered useful ($VTL(d,l,t) \in \{1\}$):

$$LAL(D,l,\hat{y},t) = \frac{\sum_{d\in D}\mathbb{1}_{\{VTL(d,l,t)\in\{1\}\wedge VTL(d,l,t)\neq\hat{y}_{d,l}\}}}{\sum_{d\in D}\mathbb{1}_{\{VTL(d,l,t)\in\{1\}\}}} \tag{4}$$

In the next step, the LAL metric values are averaged across all possible labels $l \in L$:

$$MLRAL(D,\hat{y},t) = \frac{\sum_{l\in L} LAL(D,l,\hat{y},t)}{|L|} \tag{5}$$

H. Model Benefit (MB)

To measure and better interpret the advantage of applying our model-based approach in the data acquisition process, we developed the Model Benefit (MB) metric. It is the difference between the gain defined by $AER(D,\hat{y},t)$ and the knowledge loss calculated by $AAL(D,\hat{y},t)$ for a specific dataset D, the predictions of the model $\hat{y}$, and the threshold t:

$$MB(D,\hat{y},t) = AER(D,\hat{y},t) - AAL(D,\hat{y},t) \tag{6}$$

The positive value of $MB(D,\hat{y},t)$ indicates that the reduction of the annotation effort was greater than the loss of knowledge caused by incorrect model predictions. On the contrary, the negative MB value means that the loss of knowledge affects the greater part of the dataset than the one acquired automatically through model predictions.

IV. Datasets

The great importance of data used during our experiments was a key element in obtaining genuine results. We needed to accumulate data that were sufficiently diverse so that each subjective problem was adequately represented. For this reason, we have launched a project named Doccano 1.0, where individuals annotated a diverse number of texts that corresponded to subjective problems. Furthermore, after a thorough analysis of many datasets, we have also chosen to expand our experimental set of data by adding two sources, Measuring Hate Speech and Unhealthy Conversations. Although the volume and quality of the data were undoubtedly crucial, the vast difference between the datasets is prominent to a degree that allows for an accurate display of comparison between traditional and our approach. Table I presents a brief summary of the data and statistics on the datasets used during our experiments.

A. Doccano 1.0

After a thorough analysis of the available sources regarding subjective NLP tasks, we have noticed a certain lack of datasets focused on a variety of dimensions. Although part of the subjectivity area in NLP features one-coded labels of sarcasm or offensiveness, it may as well use a group of emotions, an example being Pluchik's wheel of emotions. However, none of them include a wide range of emotions, opinions, and feelings of the annotators. For this reason, we have launched a project named Doccano 1.0, where users annotated a diverse number of texts that corresponded to subjective problems. We have recruited around 40 individual people that were tasked to annotate 880 texts in the scope of 23 different subjective NLP tasks each. Each person annotated around 702 texts, and each text contains around 32 different annotations. It means that in total we acquired over 700k individual annotations. Labels available for anotating were as follows: (1) *positive*, (2) *negative*, (3) *joy*, (4) *delight*, (5) *inspiration*, (6) *calm*, (7) *surprise*, (8) *compassion*, (9) *fear*, (10) *sadness*, (11) *repulsion*, (12) *anger*, (13) *ironic*, (14) *embarrassing*, (15) *vulgar*, (16) *political*, (17) *interesting*, (18) *understandable*, (19) *incomprehensible*, (20) *offensive to me*, (21) *offensive to someone*, (22) *funny to me* and (23) *funny to someone*. Each of the 23 available labels had to be graded from 0 to 10, where 0 equals disagreement, and the latter a strong agreement.

B. Unhealthy Conversations (UC)

The Unhealthy Conversations (UC) dataset [38] was published in October 2020 and consists of 44k comments. Each piece of data can contain up to 250 characters sourced from Globe and Mail opinion articles that were sampled from the Simon Fraser University Opinion and Corpus dataset [39]. The labels used to annotate these comments are as follows: (1) *antagonize*, (2) *condescending*, (3) *dismissive*, (4) *generalization*, (5) *generalization unfair*, (6) *healthy*, (7) *hostile*, and (8) *sarcastic*. There were at least three annotators per comment, and each text had to be described as at least one of the labels. Furthermore, to eliminate any possible bias, each comment was isolated from the context of the news articles and presented to the annotators as individual pieces of text.

C. Measuring Hate Speech (MHS)

The Measuring Hate Speech (MHS) dataset [40] is a collection of texts that were made available in 2020. It contains 39,565 comments from popular media platforms: Youtube, Twitter, and Reddit. The 7,912 Amazon Mechanical Turk workers specifically from the United States were involved in the annotation process. Since the data focus on offensiveness, it is possible to use labels that focus on different types of offensiveness, specifically: (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*. In our experiments, we treated each different type of offensiveness as a separate NLP task.

V. Experimental Setup

During our experiments, we used the 10-fold cross-validation (CV). In each iteration, we trained the model using 8 folds. One of the remaining folds was used as a validation set and another as a test set. Then, we calculated the mean and standard deviation of the evaluation metrics. In addition, to measure the significance of differences in the evaluation results between labels, we performed statistical tests. After verifying the test assumptions, we applied the *t*-test for independent samples with Bonferroni correction. If the assumptions were not met, we used the Mann-Whitney *U* test.

A. Deep Neural Architectures

We fine-tuned the two variants of pre-trained transformer models, depending on the language of the data. In the case of the Doccano 1.0 dataset, we used the HerBERT model [41]. For the UC and MHS datasets, we leveraged XLM-RoBERTa [42]. The implementation of both models was obtained from the HuggingFace library [43]. In both cases, we used the cross-entropy as a loss function with a learning rate set to 10^{-5}. We leveraged the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and the L2 regularization hiperparameter weight decay equals 10^{-3}.

B. Self-Supervised Model Evaluation

The first experiment was the self-supervised evaluation. We trained and evaluated the model on the original dataset. From the CV, we obtained the predicted labels for each text in the dataset. In the next step, we used them as target labels. In this way, we generated a second dataset in which text samples are exactly the same as in the original dataset, but target labels are predicted by the model. Subsequently, we trained the model on the predicted labels and evaluated it on the original set of labels. The self-supervised evaluation scenario is described in Fig. 3.

C. Incremental Knowledge Evaluation

Another experiment focused on increasing the train set size in an incremental way. We started with a training set consisting of only 1 fold for each iteration of the CV. Then, we increased the train set size to 2 folds, by adding the next fold to the one used in the previous experiment step. We kept increasing the train set size by 1 fold during the next steps. In the final iteration, we used 8 folds as a train set for each of the CV iterations.

D. Threshold t Value Evaluation

We also focused on the evaluation regarding various values of the t threshold value used in the VTL measure described in Sec. III-C. To better understand the impact of the threshold t, we tested values in the range $[0.1, 0.25]$. In other words, we considered scenarios from the situation when we considered the label as valuable ($VTL(d, l, t) = 1$) if at least 10% of users annotated it with a non-zero value to the situation when the label needed to receive at least 25% of non-zero annotations to be considered as valuable.

E. Single-Task vs. Multi-Task

During the evaluation, we also analyzed the transfer of knowledge between tasks. Therefore, we measured the impact of inter-task knowledge by comparing the model performance in single-task and multi-task scenarios.

F. Diversity of People vs. Diversity of Texts

In this scenario, our goal was to investigate how many annotations from different people we should collect for one text to get the best quality of the model. In other words, we want to answer the question: *Given a budget for (N) annotations, how many texts should we annotate?* To compare models trained on different combinations of annotation and text numbers, we used the following training dataset undersampling procedure: to create a training dataset with N annotations and M texts, we sort the texts by their number of annotations, descending. Then, we take top M texts, and sample their annotations in a round-robin way, until we get N annotations. The validation and test dataset splits remain fixed. We compare the model performance on the test dataset split. We tested text numbers ranging from 100 to 500 in increments of 100, with annotation numbers ranging from 1,000 to 7,000 in increments of 1,000. An experiment was carried out for each combination on five test folds. We used a personalized UserID model with the HerBERT language model.

VI. Results

In the case of self-supervised evaluation, the training on the labels predicted by the model resulted in a significant decrease in the model performance only for 7 out of 23 labels (∼30%) in comparison to the training on the original labels for the Doccano dataset described in Fig. 4 and Fig. 5. The *Difference* marks the difference in performance between the model trained on the original labels and those predicted by the model. The shaded area marks the MB metric that describes the reduction in annotation effort with respect to the loss of knowledge.

For the MHS dataset, the model trained in a self-supervised way achieved better results in 9 out of 10 labels (90%), Fig. 6.

For incremental knowledge evaluation, a positive MB value was observed since training on 1 fold. The highest MB value (0.15) was noted for training on 8 folds, Fig. 7.

During the threshold t value evaluation, the highest AER and MB values were observed for $t = 0.25$ and were equal to 0.39 and 0.25, respectively, for the Doccano dataset, Fig. 8.

TABLE I
DATASET PROFILES AFTER PRE-PROCESSING. EACH DATASET CONTAINS A SET NUMBER OF LABELS, WHICH ARE EXPLAINED IN FULL DETAIL IN SECTION IV. THE FIELD *Number of annotated labels* DESCRIBES THE NUMBER OF ANNOTATIONS FOR ALL AVAILABLE LABELS IN A SPECIFIC DATASET.

Dataset Property	Doccano 1.0	Unhealthy Conversations	Measuring Hate Speech
Textual content profile	comments & discussions	comments & discussions	comments
Number of tasks	23, i.e., delight, offensive to me, funny to someone	8, i.e. hostile, sarcastic, unfair generalization	10, i.e. violence, attack-defend, dehumanize
Label values	{0, . . . , 10}	{0, 1}	{0, . . . , 4}
Output / ML task	23*regression	8*binary classification	10*regression
Number of texts	1,000	44,355	39,565
Number of annotations	31,338	244,468 (227,975 valid)	135,556
Number of annotated labels	720,774	1,823,800	1,355,560
Number of annotators	40	558	7,912
Avg. annotations per text	31.34	4.66	3.43
Avg. annotations per annotator	783.45	387.71	17.13
Language	Polish	English	English

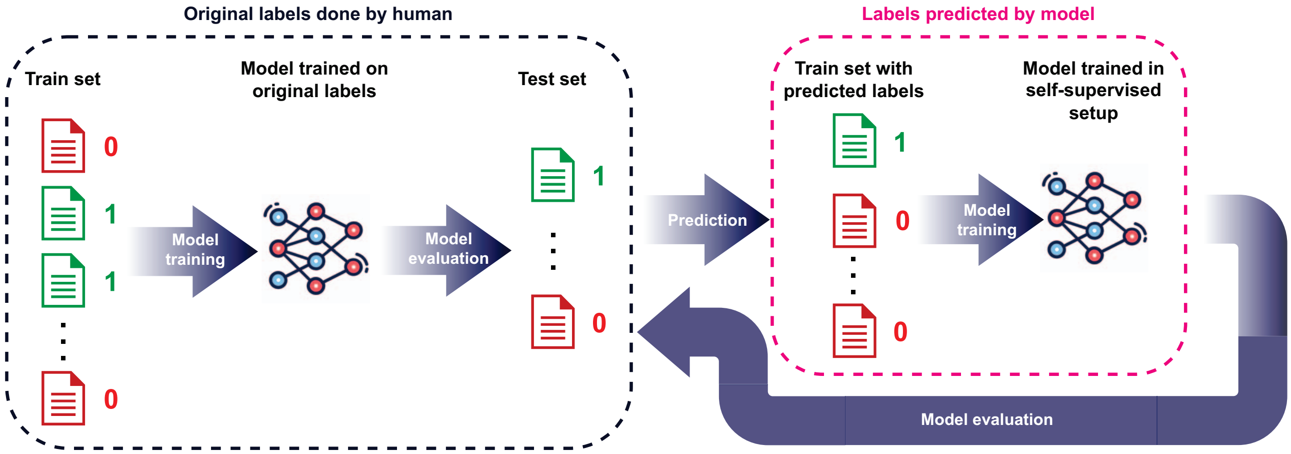
Original labels done by human
Train set
Model trained on original labels
Test set
0
1
1
0
Model training
Model evaluation
1
0
Prediction
Labels predicted by model
Train set with predicted labels
Model trained in self-supervised setup
1
0
0
Model training
Model evaluation

Fig. 3. The self-supervised evaluation scenario, where the model is trained on labels predicted by the instance trained on the original data, presented for a single label.

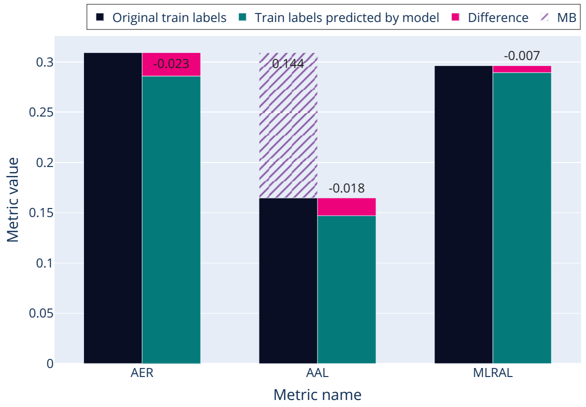

Fig. 4. The values of effort reduction AER, loss AAL, and MLRAL, as well as final model benefit MB for the self-supervised evaluation on the Doccano dataset.

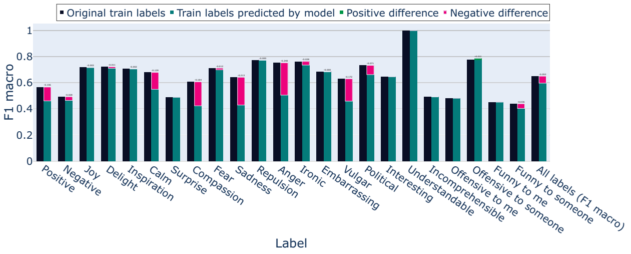

Fig. 5. Macro F1 values for the self-supervised evaluation on the Doccano dataset.

For the UC dataset, the highest values for these metrics were equal to 0.41 and 0.14, Fig. 9.

For the single-task vs. multi-task evaluation, a significant increase in model performance was observed for every label in the MHS dataset, Fig. 10. The highest increase caused by multi-task learning was 0.36 for *Sentiment* label.

The results for the experiments on the dependence of the model quality on the number of texts in the training dataset are

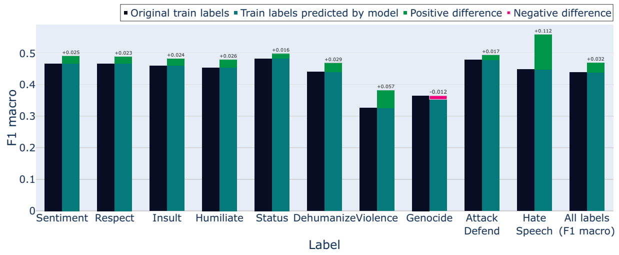

Fig. 6. Macro F1 values for the self-supervised evaluation on the MHS dataset.

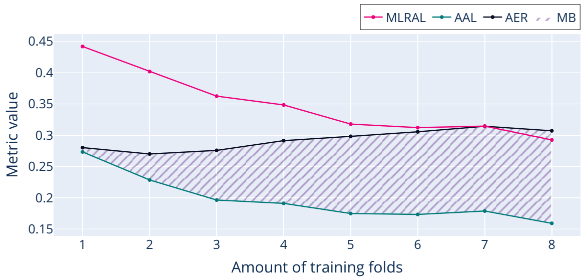

Fig. 7. The values of AER, AAL, and MLRAL metrics for the incremental knowledge evaluation on the Doccano dataset. The hatched area is Model Benefit (MB): the gain in effort reduction (AER) minus loss (AAL).

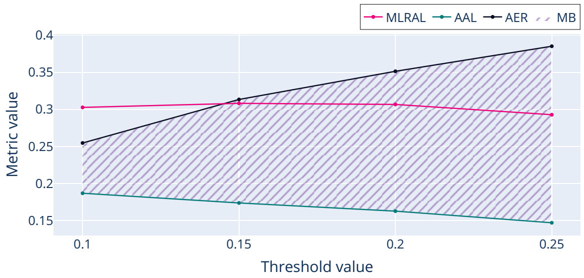

Fig. 8. The values of AER, AAL, and MLRAL metrics for the threshold evaluation on the Doccano dataset. The hatched area is Model Benefit (MB), i.e. the gain in effort reduction (AER) minus loss (AAL).

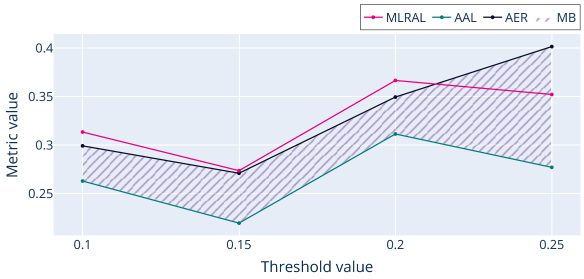

Fig. 9. The values of AER, AAL, and MLRAL for the threshold evaluation on the UC dataset. The hatched area is Model Benefit (MB): the gain in effort reduction (AER) minus loss (AAL).

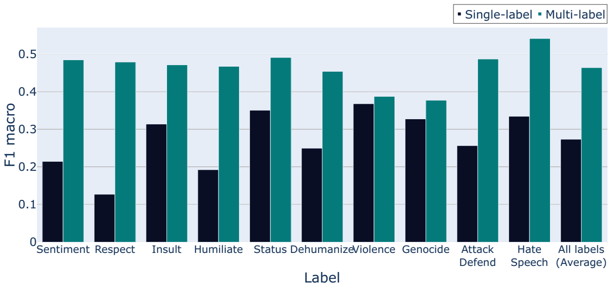

Fig. 10. Macro F1 values for the single-task vs. multi-task evaluation on the MHS dataset.

TABLE II
MODEL PERFORMANCE FOR DIFFERENT TEXT/ANNOTATIONS NUMBER FOR DOCCANO 1.0 DATASET. WE USED AVERAGED R^2 OVER ALL DATASET TASKS AS A METRIC.

Annotations number \ Texts number	100	200	300	400	500
1,000	08.34	09.59	11.16	12.77	10.73
2,000	19.24	17.46	21.34	22.86	22.75
3,000	–	23.73	24.96	27.31	26.10
4,000	–	25.00	27.75	28.65	28.05
5,000	–	–	28.04	29.07	29.59
6,000	–	–	29.05	30.13	29.72
7,000	–	–	–	30.22	29.96

presented in Tab. II. Due to the limited number of annotations per text for the Doccano dataset, we were unable to evaluate some combinations of the number of texts and annotations (e.g. 100 texts and 7,000 annotations). A slightly positive effect of more texts on the quality of the model is visible. This may be due to the fact that linguistic knowledge is more important than personalization in the case of small datasets.

VII. Discussion

Behind every performance of NLP methods, there is the data and the significant impact of its quality. Even the slightest noise may hinder the model performance, and thus it is important to focus on the characteristics of our data. The number of positive and negative values of a single label in our experiments did not have balanced proportions, since the negative values were in the dominant part of each class. As a result, there is a certain bias towards negative values, which may aggravate the model predictions. This is a factor that adversely affects our methods, as the loss of a small number of minority class annotations could significantly reduce the quality of the model performance. Therefore, the high MB values and resulting loss-adjusted benefits of our methods confirm their versatility and robustness to unfavorable annotation distributions. On the other hand, we think that a re-evaluation of the scores with a technique that includes other features and values of each label (i.e., confusion matrix) would be applicable to further improve the stability of our methods.

The very similar trend observed for MLRAL and AAL for the Doccano dataset indicates that the loss of valuable (non-zero) annotations is independent of the diversity of annotation distributions for individual tasks. This reveals the prospect of using our methods effectively on datasets annotated with multiple tasks, regardless of their distributions.

Just as the value of the loss-adjusted MB measure is relevant for each collection used, we are fully aware that the objective of annotation may vary depending on the characteristics of the phenomenon being annotated. This means that in special cases, the loss of valuable annotation may prevent the model from extracting the correct signal from the data. That is when MLRAL minimization may prove more important than MB maximization. On the other hand, for a phenomenon that occurs quite frequently, maximization of AER may prove more efficient.

The low performance differences in the self-supervised scenario on the Doccano dataset indicate that the model stores much of the knowledge necessary to predict on unknown texts in a way that does not significantly reduce the quality of the labels it provides. This makes it possible to incrementally enlarge the dataset to, for example, support more domains or adapt the model to another language.

The improvement in performance in the MHS dataset in the self-supervised scenario as shown in Fig. 6 is due to the fact that training the model on the predicted labels can be treated as a regularization method.

The use of inter-task knowledge has significantly improved the model performance on the MHS set in the single-task vs. multi-task scenario. By simultaneously predicting VTL values for each task, the model learns inter-task relationships, which allows it to discover the hidden semantics of the tasks it learns.

VIII. Conclusions and Future Work

In this work, we proposed a new model-based approach to data acquisition for subjective multi-task NLP problems. We develop novel metrics to calculate the benefit and loss of our method. We also performed a complex evaluation to verify the efficiency of the proposed solution from multiple perspectives. One of the most important results of our experiments is that the self-supervised approach can be used as a regularization technique for subjective multi-task problems. With this setup, the model is able to remove outliers and subsequently improve the overall performance quality. The use of our approach can lead to a reduction in the overall annotation effort by 40%. Assuming the pricing in annotation services such as Amazon Mechanical Turk[2], the cost to annotate one label is \$0.012. On that basis, we can calculate that the cost of annotating the data set with a size comparable to the UC dataset would be \$219k, but the 40% reduction caused by our method would allow saving up to \$87.6k.

In future work, we will focus on further improvements of our model-based approach by using different model architectures, developing new measures tailored for specific business cases, and applying our method to more datasets regarding subjective multi-task problems. The code for all methods and experiments is publicly available [3] under the MIT license.

Acknowledgements

This work was financed by (1) the National Science Centre, Poland, project no. 2021/41/B/ST6/04471; (2) Contribution to the European Research Infrastructure 'CLARIN ERIC - European Research Infrastructure Consortium: Common Language Resources and Technology Infrastructure', 2022-23 (CLARIN Q); (3) the Polish Ministry of Education and Science, CLARIN-PL; (4) the European Regional Development Fund as a part of the 2014-2020 Smart Growth Operational Programme, projects no. POIR.04.02.00-00C002/19, POIR.01.01.01-00-0923/20, POIR.01.01.01-00-0615/21, and POIR.01.01.01-00-0288/22; (5) the statutory funds of the Department of Artificial Intelligence, Wroclaw University of Science and Technology; (6) the Polish Ministry of Education and Science within the programme "International Projects Co-Funded"; (7) the European Union under the Horizon Europe, grant no. 101086321 (OMINO). However, the views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor European Research Executive Agency can be held responsible for them.

References

[1] D. Brain and G. I. Webb, "On the effect of data set size on bias and variance in classification learning," in *Proceedings of the Fourth Australian Knowledge Acquisition Workshop, University of New South Wales*, 1999, pp. 117–128.

[2] M. A. Cantrell and P. Lupinacci, "Methodological issues in online data collection," *Journal of advanced nursing*, vol. 60, no. 5, pp. 544–549, 2007.

[3] A. M. Hutchinson, D. L. Milke, S. Maisey, C. Johnson, J. E. Squires, G. Teare, and C. A. Estabrooks, "The resident assessment instrument-minimum data set 2.0 quality indicators: a systematic review," *BMC health services research*, vol. 10, no. 1, pp. 1–14, 2010.

[4] L. T. Rose and K. W. Fischer, "Garbage in, garbage out: Having useful data is everything," *Measurement: Interdisciplinary Research & Perspective*, vol. 9, no. 4, pp. 222–226, 2011.

[5] J. M. Cavanillas, E. Curry, and W. Wahlster, *New horizons for a data-driven economy: a roadmap for usage and exploitation of big data in Europe*. Springer Nature, 2016.

[6] K. Lyko, M. Nitzschke, and A.-C. Ngonga Ngomo, "Big data acquisition," *New Horizons for a Data-Driven Economy: A Roadmap for Usage and Exploitation of Big Data in Europe*, pp. 39–61, 2016.

[7] T. Wang, J.-Y. Zhu, A. Torralba, and A. A. Efros, "Dataset distillation," *arXiv preprint arXiv:1811.10959*, 2018.

[8] D. Barrett and A. Twycross, "Data collection in qualitative research," *Evidence-based nursing*, vol. 21, no. 3, pp. 63–64, 2018.

[9] T. Nguyen, R. Novak, L. Xiao, and J. Lee, "Dataset distillation with infinitely wide convolutional networks," *Advances in Neural Information Processing Systems*, vol. 34, pp. 5186–5198, 2021.

[10] B. Vidgen and L. Derczynski, "Directions in abusive language training data, a systematic review: Garbage in, garbage out," *Plos one*, vol. 15, no. 12, p. e0243300, 2020.

[11] I. Sucholutsky and M. Schonlau, "Soft-label dataset distillation and text dataset distillation," in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021, pp. 1–8.

[12] Y. Zhou, E. Nezhadarya, and J. Ba, "Dataset distillation using neural feature regression," *Advances in Neural Information Processing Systems*, vol. 35, pp. 9813–9827, 2022.

[2] https://aws.amazon.com/sagemaker/data-labeling/pricing/

[3] https://github.com/CLARIN-PL/model-based-data-acquisition/releases/tag/2023-icdm-sentire

[13] S. Liu, K. Wang, X. Yang, J. Ye, and X. Wang, "Dataset distillation via factorization," *Advances in Neural Information Processing Systems*, vol. 35, pp. 1100–1113, 2022.

[14] H. T. Larasati, A. T. Prihatno, H. Kim *et al.*, "A review of dataset distillation for deep learning," in *2022 International Conference on Platform Technology and Service (PlatCon)*. IEEE, 2022, pp. 34–37.

[15] G. Cazenavette, T. Wang, A. Torralba, A. A. Efros, and J.-Y. Zhu, "Dataset distillation by matching training trajectories," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4750–4759.

[16] N. Sachdeva and J. McAuley, "Data distillation: A survey," *arXiv preprint arXiv:2301.04272*, 2023.

[17] S. Zhang, C. Chen, X. Hu, and S. Peng, "Balanced knowledge distillation for long-tailed learning," *Neurocomputing*, 2023.

[18] F. Radenovic, A. Dubey, A. Kadian, T. Mihaylov, S. Vandenhende, Y. Patel, Y. Wen, V. Ramanathan, and D. Mahajan, "Filtering, distillation, and hard negatives for vision-language pre-training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6967–6977.

[19] O. Bohdal, Y. Yang, and T. M. Hospedales, "Flexible dataset distillation: Learn labels instead of images," in *4th Workshop on Meta-Learning at NeurIPS 2020*, 2020.

[20] J. Gao, J. Li, H. Shan, Y. Qu, J. Z. Wang, F.-Y. Wang, and J. Zhang, "Forget less, count better: a domain-incremental self-distillation learning benchmark for lifelong crowd counting," *Frontiers of Information Technology & Electronic Engineering*, vol. 24, no. 2, pp. 187–202, 2023.

[21] S. E. Whang, Y. Roh, H. Song, and J.-G. Lee, "Data collection and quality challenges in deep learning: A data-centric ai perspective," *The VLDB Journal*, pp. 1–23, 2023.

[22] W.-C. Chen and W.-T. Chu, "Sssd: Self-supervised self distillation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 2770–2777.

[23] K. Kanclerz, P. Miłkowski, and J. Kocoń, "Cross-lingual deep neural transfer learning in sentiment analysis," *Procedia Computer Science*, vol. 176, pp. 128–137, 2020, knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 24th International Conference KES2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S187705092031838X

[24] P. Miłkowski, M. Gruza, K. Kanclerz, P. Kazienko, D. Grimling, and 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, pp. 248–259.

[25] J. Kocoń, M. Gruza, J. Bielaniewicz, D. Grimling, K. Kanclerz, P. Miłkowski, and 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.

[26] J. Kocoń, A. Figas, M. Gruza, D. Puchalska, T. Kajdanowicz, and P. Kazienko, "Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach," *Information Processing & Management*, vol. 58, no. 5, p. 102643, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306457321001333

[27] J. Kocoń, P. Miłkowski, and K. Kanclerz, "Multiemo: Multilingual, multilevel, multidomain sentiment analysis corpus of consumer reviews," in *Computational Science – ICCS 2021*, M. Paszynski, D. Kranzlmüller, V. V. Krzhizhanovskaya, J. J. Dongarra, and P. M. A. Sloot, Eds. Cham: Springer International Publishing, 2021, pp. 297–312.

[28] A. Ngo, A. Candri, T. Ferdinan, J. Kocoń, and 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.

[29] P. Miłkowski, S. Saganowski, M. Gruza, P. Kazienko, M. Piasecki, and 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.

[30] K. Kanclerz, M. Gruza, K. Karanowski, J. Bielaniewicz, P. Miłkowski, J. Kocoń, and 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, pp. 37–45.

[31] J. Bielaniewicz, K. Kanclerz, P. Miłkowski, M. Gruza, K. Karanowski, P. Kazienko, and J. Kocoń, "Deep-sheep: Sense of humor extraction from embeddings in the personalized context," in *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2022, pp. 967–974.

[32] T. Ferdinan and J. Kocoń, "Personalized models resistant to malicious attacks for human-centered trusted ai," in *The AAAI-23 Workshop on Artificial Intelligence Safety (SafeAI 2023)*. CEUR Workshop Proceedings, 2023.

[33] W. Mieleszczenko-Kowszewicz, K. Kanclerz, J. Bielaniewicz, M. Oleksy, M. Gruza, S. Woźniak, E. Dzięcioł, P. Kazienko, and J. Kocoń, "Capturing human perspectives in nlp: Questionnaires, annotations, and biases," in *The ECAI 2023 2nd Workshop on Perspectivist Approaches to NLP*. CEUR Workshop Proceedings, 2023.

[34] J. Kocoń, J. Baran, K. Kanclerz, M. Kajstura, and P. Kazienko, "Differential dataset cartography: Explainable artificial intelligence in comparative personalized sentiment analysis," in *International Conference on Computational Science*. Springer, 2023, pp. 148–162.

[35] P. Kazienko, J. Bielaniewicz, M. Gruza, K. Kanclerz, K. Karanowski, P. Miłkowski, and J. Kocoń, "Human-centered neural reasoning for subjective content processing: Hate speech, emotions, and humor," *Information Fusion*, vol. 94, pp. 43–65, 2023.

[36] J. Kocoń, I. Cichecki, O. Kaszyca, M. Kochanek, D. Szydło, J. Baran, J. Bielaniewicz, M. Gruza, A. Janz, K. Kanclerz *et al.*, "Chatgpt: Jack of all trades, master of none," *Information Fusion*, p. 101861, 2023.

[37] K. Kanclerz, A. Figas, M. Gruza, T. Kajdanowicz, J. Kocon, D. Puchalska, and 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)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 5915–5926. [Online]. Available: https://aclanthology.org/2021.acl-long.460

[38] I. Price, J. Gifford-Moore, J. Flemming, S. Musker, M. Roichman, G. Sylvain, N. Thain, L. Dixon, and J. Sorensen, "Six attributes of unhealthy conversations," in *Proceedings of the Fourth Workshop on Online Abuse and Harms*, 2020, pp. 114–124.

[39] V. Kolhatkar, H. Wu, L. Cavasso, E. Francis, K. Shukla, and M. Taboada, "The sfu opinion and comments corpus: A corpus for the analysis of online news comments," *Corpus Pragmatics*, vol. 4, no. 2, pp. 155–190, 2020.

[40] C. J. Kennedy, G. Bacon, A. Sahn, and C. von Vacano, "Constructing interval variables via faceted Rasch measurement and multitask deep learning: a hate speech application," *arXiv e-prints*, p. arXiv:2009.10277, Sep. 2020.

[41] R. Mroczkowski, P. Rybak, A. Wróblewska, and I. Gawlik, "HerBERT: Efficiently pretrained transformer-based language model for Polish," in *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*. Kiyv, Ukraine: Association for Computational Linguistics, Apr. 2021, pp. 1–10. [Online]. Available: https://aclanthology.org/2021.bsnlp-1.1

[42] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, É. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 8440–8451.

[43] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://aclanthology.org/2020.emnlp-demos.6