Abstract—Designing predictive models for subjective problems in natural language processing (NLP) remains challenging. This is mainly due to its non-deterministic nature and different perceptions of the content by different humans. It may be solved by Personalized Natural Language Processing (PNLP), where the model exploits additional information about the reader to make more accurate predictions. However, current approaches require complete information about the recipients to be straight embedded. Besides, the recent methods focus on deterministic inference or simple frequency-based estimations of the probabilities. In this work, we overcome this limitation by proposing a novel approach to capture the uncertainty of the forecast using conditional Normalizing Flows. This allows us to model complex multimodal distributions and to compare various models using negative log-likelihood (NLL). In addition, the new solution allows for various interpretations of possible reader perception thanks to the available sampling function. We validated our method on three challenging, subjective NLP tasks, including emotion recognition and hate speech. The comparative analysis of generalized and personalized approaches revealed that our personalized solutions significantly outperform the baseline and provide more precise uncertainty estimates. The impact on the text interpretability and uncertainty studies are presented as well. The information brought by the developed methods makes it possible to build hybrid models whose effectiveness surpasses classic solutions. In addition, an analysis and visualization of the probabilities of the given decisions for texts with high entropy of annotations and annotators with mixed views were carried out.

Index Terms—artificial neural networks, natural language processing, human profile modelling, probabilistic technique

I. INTRODUCTION

Human affective states, including emotions, strongly depend on the individual, the stimulant eliciting them, and the associated context [1]. Therefore, the reasoning of a person’s perception based on machine learning bears a significant degree of uncertainty. It refers to the reaction to any content, including text reading. We can say that disagreements in human textual inferences are inherent [2]. Most solutions to subjective problems in natural language processing (NLP), like recognition of emotions, hate speech, sarcasm, sense of humor, sentiment, and many others, rely on generalized perspectives. They consider only text and its single generalized interpretation. Then, the commonly used solution is to simplify multiple distinct views, i.e., annotations provided by many annotators using majority voting or other methods to achieve a sole perception. Overall, we can identify two sources of uncertainty: (1) humans, who are unsure and imprecise in their annotations (this is a hidden factor), and (2) a community of annotators. The latter refers to discrepancies between people in understanding the problem, and perception of a given text [3]–[5]. The standard measures for inter-rater agreement are Krippendorff’s alpha [6] or Fleiss’ kappa [7]. However, they provide only a single value characterizing the set of all annotations for all texts. Yet another (3) source of uncertainty: the trained model itself. It means that the model is not capable of precisely learning about concepts (what is joy or hate speech?) and relations from the available learning samples. This leads to errors and proximate reasoning. Simultaneously, emotions can be considered multidimensional objects, which requires multi-task learning [8] and further complicates the problem of uncertainty modeling. Most of the proposed approaches for subjective modeling in the NLP domain focus on deterministic predictions. In this work, we propose to enrich the family of emotional methods by introducing Emotional Normalizing Flow – an entirely probabilistic framework that utilizes conditional Normalizing Flows to model uncertainty. We postulate to represent the considered tasks as multivariate regression problems and represent the distribution of the outputs with conditional flows. This approach allows us to model complex multimodal distributions of multidimensional outputs. The experiments and validation were carried out on emotion detection (ten tasks) and hate speech (two tasks). We examine various choices of flow models and compare their performance with the mixture of Gaussians, showing the superiority of Emotional Normalizing Flow compared to the selected baseline. Moreover, we show that incorporating personalization into our model leads to better distribution adjustment measured with negative log-likelihood (NLL) value.

To summarize, the contributions of this work are as follows:

- We introduce a novel approach for probabilistic modeling in subjective NLP-based problems;
- We examine the impact of personalization on the quality of the model and show that in most of the considered experimental cases, additional information about the reader leads to better probability adjustment;
- We show that our approach outperforms the standard baseline that utilizes a mixture of Gaussians;

1These authors contributed equally to this work.
• We propose a hybrid approach utilizing Normalizing Flows and personalization that outperform previous models.

II. RELATED WORK

Initial work on emotion recognition in the text was based mainly on frequency analysis of words defined in lexicons of emotions [9], [10]. These lexicons contained words with assigned categories of basic emotions, e.g., joy, anger, sadness [11], [12]. Emotions occurring most frequently at the lexical level were then assigned to the entire text. With the development of text classification methods based on machine learning, datasets containing texts manually annotated with emotions began to emerge [13]–[20]. Due to annotators’ subjective perception of emotions, and thus low inter-annotator agreement, it was common to assign emotion labels to text based on majority voting [13], [18]. Based on such prepared data, text classification models were trained. Initially, such models as SVM [21], BiLSTM, and GRU [22] were used. Currently, transformer-based models such as BERT perform best in the task of emotion recognition [18], [23]. The aforementioned approaches require data for which the inter-annotator agreement is high. However, there are some data sets such as Wiki-Detox [24], Sentimenti [16] or Measuring Hate Speech [25], which contain an annotator identifier linked to their affective annotation. They also include multiple annotations for a given text from multiple annotators. For such data, new personalized approaches have recently been developed, in which the context of the annotator is taken into account in the model learning process [26]–[36]. This makes it possible, for example, to answer the question of what emotions a particular text evokes for a particular user. Recent method proposals also focus on neuro-symbolic approaches to explain decisions made [37], usage of large-scale pre-trained language model (PLM) for prompt-based classification tasks such as sentiment analysis and emotion detection [38], using recently popular large language models (LLMs) [39], or methods of complex persona attribute extraction [40]. However, the methods mentioned above do not model the uncertainty associated with the community’s subjective perception of emotions and the degree of indecision of the annotators themselves.

In this paper, to model uncertainty described in the Introduction we adapt the concept of Normalizing Flows. The best-known Normalizing Flow models such as NICE [41], RealNVP [42], MAF [43], and CNF [44] were originally used for density estimation and image generation tasks. These models were further extended and used as components for more sophisticated tasks or even for other domains of applications. In Computer Vision, there were proposed models such as RegFlow [45] for probabilistic future location prediction, Flow Plugin Network [46], PluGen [47], and StyleFlow [48] models for conditional image generation. For the tabular data, recently, TreeFlow [49] was proposed that utilizes a combination of tree-based models with conditional Normalizing Flows to estimate uncertainty for univariate regression problems. In terms of Natural Language Processing and Normalizing Flows, only Discrete Flow [50] was proposed to model character-level datasets using Normalizing Flows dedicated to the discrete data. To the best of our knowledge, no probabilistic approach has been proposed to model distributions of uncertainty in personalized natural language processing, and our Emotional Normalizing Flow is the first probabilistic model proposed for multi-task prediction of personalized emotions.

III. BACKGROUND

a) Generalized and Personalized Approach to Subjective NLP Problems.: In the classic approach to the task of text classification or regression, we assume a training set of the form $D = \{(t_i, y_i)\}_{i=1}^N$, where $t_i \in T$ is the $i$-th text document and $y_i$ is its annotation. However, many NLP tasks, such as recognizing emotions in a text or detecting hate speech, can be subjective because each person perceives these phenomena. This leads to a situation when we can have more than one annotation $y$ for the same text $t$, as different people may annotate the same texts differently. Therefore, a training set is in the form of $D = \{(t_i, p_i, y_i)\}_{i=1}^N$, where $y_i$ is the annotation given by person $p_i \in P$ for text $t_i \in T$.

One approach to subjective tasks in NLP is the so-called generalized. It assumes that the model predicts the result based solely on the text and returns the same prediction for every user. Generalized models usually consist of two parts: text encoder (language model), which creates text representation $e_t$ and classifier or regressor (usually fully-connected layer) that gives prediction $\hat{y}$. However, recent studies [28], [51], [52] show that this approach should not be considered correct, as adding information about the annotator significantly improves model quality and yields better results. The approach that combines information about the text and the human is so-called personalized. Compared to the generalized, personalized model adds another component called profile extractor, that creates human representation $e_p$. The comparison of generalized and personalized approaches is shown in Fig 1.

There are few existing architectures [28], [51] utilizing this fact. Still, all of them are deterministic, meaning none model uncertainty as a direct optimization of negative log-likelihood.

b) Normalizing Flows.: Normalizing Flows [53] are a class of generative models that enables estimation of the uncertainty of prediction thanks to the access to log probability function and thus enable direct optimization of negative log-likelihood (NLL). The goal of the model is to transform base distribution $p_U(\mathbf{u})$ (usually Gaussians with independent components) to the complex distribution of the data $p_Y(y)$ using a series of $K$ invertible functions that can be written as $\mathbf{u} = f_K \circ \cdots \circ f_1(y)$. For that purpose, Normalizing Flows utilize the change-of-variable formula and then the NLL $y$ is given by

$$
\log p_Y(y) = \log p_U(\mathbf{u}) - \sum_{k=1}^K \log \left| \det \frac{\partial f_k}{\partial z_{k-1}} \right|.
$$

(1)

To specify the exact Normalizing Flow model, we need to define transformations $f_1, \ldots, f_K$. Here, multiple models were
Fig. 1: Comparison of (a) generalized and (b) personalized deterministic models. (a) consists of two parts: a text encoder (language model) that creates text embedding $e_t$ and a classifier or regressor (mostly fully-connected layer) that provides prediction $\hat{y}_t$. This approach is not considered suitable for subjective NLP tasks, like emotion recognition, because it does not respect the individual perception of the text. Model (b) fixes this problem by adding a profile extractor in the form of user representation $e_p$. It allows human individual characteristics to be included in the inference process. Both models are deterministic, giving us limited, spotlight information about subjective tasks.

IV. OUR APPROACH

In this section, we introduce Emotional Normalizing Flow - the probabilistic model for subjective uncertainty modeling in the NLP domain. The general schema of the proposed approach is provided in Fig. 2. The model is composed of Profile extractor that is responsible for creating the representation of the person, $e_p$, and Text encoder that creates embedding $e_t$ directly from the input text. Both components can be represented by various models (trainable and fixed), and we elaborate on this further in this section.

The extracted vectors $e_p$ and $e_t$ are further delivered to the conditional flow represented by the complex transformation function $f(\cdot)$. The role of the function is to transform multivariate regression outputs $y$ to $z$ that represents the variable in the base space, assuming given vectors, $e_p$ and $e_t$. Formally, we have $z = f(y, e_p, e_t)$, where $f$ is invertible with respect to $y$, $y = f^{-1}(z, e_p, e_t)$. Moreover, the complex transformation $f$ can be decomposed into a sequence of simple functions, $z = f_K \circ \cdots \circ f_1(y, e_p, e_t)$, where the $K$ is number discrete transformations. With such assumptions, the probability distribution for $y$ that represents the distribution over the regression outputs can be calculated using the formula:

$$p_Y(y | e_t, e_p) = p_Z(z) \cdot \prod_{k=1}^{K} \left| \det \frac{\partial f_k}{\partial z_{k-1}} \right|,$$

where $z_0, \ldots, z_K$ are intermediate steps after discrete transformations, assuming $z_0 := y$, and $z_K := z$. $p_Z(z)$ is the assumed base distribution for $z$ with the known density.
function, usually represented by Gaussian. Consequently, we have direct access to the density function for that conditional distribution. Therefore we can calculate the likelihood function for a set of input-output pairs to evaluate the quality of the model. We can sample an infinite number of output values assuming given inputs and interpret the results.

The proposed model can quickly adapt to the problems without personalization, simply skipping $e_p$ conditioning in the flow. Our approach is independent of the conditional Normalizing Flow type, and we experimentally compare the performances of the most popular models. We follow the methodology of incorporating conditional components described in [46].

a) Profile extractor: Vector $e_p$ contains information about the user specific to the personalization architecture used. This can include information such as the deviation of responses from the majority voice, metadata about the user, user identifier [28], the correlation of the text’s context with responses from the majority voice, metadata about the user, used. This can include information such as the deviation of $y$ regarding subjective annotation

$$L = -\sum_{n=1}^{N} \log p_{\mathcal{Y}}(y_n|e_{n,p}, e_{n,t}),$$

where $e_{n,p}$ is a vector, that represents profile of the person $p_n$, and $e_{n,t}$ is an embedding of the text $t_n$. The model can be trained in a two-stage mode or end-to-end paradigm depending on the form of Profile extractor and Text encoder. In the first case, the embeddings $e_p$ and $e_t$ are extracted in the first stage, and parameters of the flow are trained while optimizing $L$. Alternatively, suppose the Profile extractor or Text encoder are represented by differentiable architectures. In that case, the entire system can be optimized end-to-end, directly minimizing the negative log-likelihood function.

V. EXPERIMENTS

In this subsection, we evaluate our approach on a set of challenging datasets, investigating the impact of adding contextual information about the person in the model. Moreover, we compare flow-based probabilistic models to a simple Gaussian Mixture Model. Then, we compare our solution to the deterministic models using sampling from flow and discretization. Finally, we mix deterministic and probabilistic approaches to create a hybrid model.

A. Datasets

a) Wikipedia-Detox: The Wikipedia Detox project has created a crowd-sourced dataset that contains one million annotations covering 100,000 discussions of page edits on Wikipedia [24]. These were often filled with toxic statements, verbal aggression, and even personal attacks. Each comment was annotated by about ten annotators provided by the Crowdflower service.

The collection containing toxic statements consists of 160,000 texts. It includes a binary determination of toxicity (where: 0 = non-toxic, 1 = toxic), as well as a rating from -2 to 2 (where: 2 = very healthy, 0 = neutral, and -2 = very toxic).

Sets for personal attacks and verbal aggression consist of 100,000 of the same comments. In addition to the binary marks for aggression (0 = neutral or friendly comment, 1 = aggressive or attacking), aggression is put on a scale analogous to toxicity from -2 to 2 (where: 2 = very friendly, 0 = neutral, and -2 = very aggressive). Personal attacks are divided into types: quoting, recipient, third party, or another type of attack. In addition to texts shared between these collections, the same applies to annotators. Thus, we can use knowledge from one collection to benefit from it in another or a collective approach. Those willing to participate in the study also completed questionnaires so that we have demographic information about them available.

b) Emotion Simple: This collection consists of 100 texts marked on 10 scales by 5,365 annotators [55]. Texts are opinions posted on websites. This gives 53.65 annotations per text and 1.69 markings from a single user. The texts were rated for eight basic emotions (sadness, anticipation, joy, fear, surprise, disgust, trust, and anger) and emotional arousal on a scale from 0 to 4 for each dimension. In addition, the tenth aspect rated is the valence expressed on a scale from -3 to 3 (where -3 = negative, 0 = neutral, and 3 = positive). In the set of individuals with two marks, those with three or more annotations also appear.

c) Emotion Meanings: In [56], a huge collection containing 6,000 assessed word collocations was prepared and published. It contains dimensions and scales analogous to the Emotion Simple collection – the basic emotions from Robert Plutchik’s Wheel of Emotions [57].

The scale of the collection makes it one of the most interesting and, simultaneously, the most difficult for personalizing emotion detection. It has 303,143 annotations from 16,101 people who participated in the study. Each collocation has been evaluated 50.67 times, and a single annotator has an average of 18.83 annotations.

The difficulty in working with these data is also because these are not full-fledged textual statements containing context but just two words. An example item from the collection: "colorful beads". Annotator data include information such as
gender, age, education, size of residence, relationship status, income, or political views.

B. Setups

The dataset was divided into training, validation, and testing splits. Users and texts were not mixed between sets to bring the evaluation as close to the real-world scenario. Each experiment consisted of 10-fold cross-validation, and obtained results were averaged. Statistical significance tests were performed: t-test with Bonferroni correction to address the problem of multiple comparisons. In the tables within the rows, comparisons were made between models without and with personalization. Bold indicates the best result, and underline indicates the absence of statistically significant differences for each dataset. Within the “Type” column, the best probabilistic model type or no significant difference between the two was similarly marked for each dataset separately.

a) Baselines.: We have three reference points. To check the impact of personalization, we compared personalized models with a baseline that uses only textual information (TXT-Baseline); it is a generalized approach. To investigate the impact of normalizing flows, we compared them with a Gaussian Mixture Model to have a reference point in the form of another, less complex probabilistic method. Finally, we used MAF (maf) and CNF (cnf). For multi-dimensional Emotions Meanings and Emotions Simple, we used two extra flows: RealNVP (real_nvp) and NICE (nice). We compared the results against the baseline that uses mixtures of Gaussians to model the probability (gmm).

b) Models for conditional normalizing flows.: In our experimental evaluation, we consider Emotional Normalizing Flow with various types of conditional normalizing flows. For single-dimensional datasets: Wikipedia Detox: Toxicity, Wikipedia Detox: Aggression and Wikipedia Detox: Attack, we used MAF (maf) and CNF (cnf). For multi-dimensional Emotions Meanings and Emotions Simple, we used two extra flows: RealNVP (real_nvp) and NICE (nice). We compared the results against the baseline that uses mixtures of Gaussians to model the probability (gmm).

c) Models for personalization.: We investigate three approaches to respect the personalization context: OneHot, HuBi-Formula, and HuBi-Medium [51]. They are confronted with TXT-Baseline (generalized, non-personalized) that does not contain any information about the annotator. We exploit LaBSE [58] as a language model in every experiment.

C. Experimental scenarios

a) Experiment 1 - Comparison of generalized and personalized solutions in the probabilistic approach.: The first approach verifies the performance of Emotional Normalizing Flow with fixed hyperparameters on multiple data sets and tasks: Wikipedia Detox: Toxicity, Wikipedia Detox: Aggression, Wikipedia Detox: Attack, Emotion Meanings and Emotion Simple. We also verified the ability of the proposed Emotional Normalizing Flow to transfer knowledge between thematically similar multidimensional text labels. For this purpose, the Wikipedia Detox: Aggression and Wikipedia Detox: Attack datasets were joined, as they contain annotations for the same texts performed by the same annotators. As a result, we obtained a dataset with multi-dimensional labels. This experiment aimed to examine the effect of personalization models on the prediction of probability distributions, thus verifying whether the additional information provided to the model reduces its uncertainty and comparing Normalizing Flows to Gaussian Mixture Model.

b) Experiment 2 - Investigating the effect of hyperparameters selected for personalization and Emotional Normalizing Flow methods on the most difficult dataset.: The second approach was to verify the maximum possible reduction of model uncertainty by tuning the model hyperparameters to a given set and checking which normalizing-flow model obtained the best results. Due to limited resources, we decided to perform this experiment using only Emotion Meanings dataset. The parameters that we tuned were: the number of hidden layers, number of layers, number of blocks per layer, dropout probability, batch normalization within layers, batch normalization between layers, learning rate, the size of hidden layers used to prepare user embeddings, and the size of the output of these embeddings.

c) Experiment 3 - Comparison of probabilistic and deterministic approaches.: To compare with classical methods [8], which are deterministic, it was necessary to prepare conversions of the Emotional Normalizing Flow output to the form of exact values. Included in the body of the paper is the application of two best normalizing flows (RealNVP and CNF) for multidimensional datasets (Aggression & Attack [classification task] and Emotion Simple [regression task]).

For the first type of task, each text or text-user pair was sampled using an iterative method. In the preparation step, we increased the number of samples in the test part of the dataset so that the value from 0 to 1 with a step of 0.1 for the class was tested as a possible context. Iterations were done twice for values of 0 and 1 in the opposite class. Next, an exponential was applied to the 44 probabilities of the resulting sample (22 per class for each text). Within the values for the opposite sampling, (e.g., [0.5, 0] and [0.5, 1]) of a given dimension were summed, and then for each stopper (0.0, 0.1, ..., 1.0) divided by the sum of all values for the dimension. If the probability mass prevailed on the side from 0.0 to 0.5, it was considered that the class was not assigned and vice versa for the other part of the axis.

It was impossible for a 10-dimensional set for the regression task to sample each possible dimension in all values separately because of the number of possible combinations. Each item from the test subset was replicated 100 times containing random real values from 0 to 1 in each class. Majority voting was then conducted to determine the most likely response for the scale of each dimension. In the collection, each dimension had a value analogous to the slider setting during annotation. For this reason, the task was treated as an ordinary regression, and the resulting values were rounded to the nearest possible position. This assumption was used for both values from the deterministic and probabilistic approaches.

d) Experiment 4 - Hybrid approach (utilizing knowledge from the text and uncertainty modeling).: The combined approach, hereafter referred to as hybrid, was done in two
steps. In the first, the learned Emotional Normalizing Flow models were sampled in the same way as in Experiment 3, but for all the texts in the collection. Then, the network input was extended to the deterministic model with an additional feature. A vector containing the resulting probabilities for each text was entered along with its embeddings and, in the case of approaches with personalization, the user profile. This vector contained all the values from the sampling, and no additional mathematical operations were performed on it.

D. Results

a) Results of Experiment 1.: The first experiment proved that adding personalization reduces the uncertainty of probabilistic models, Tab. I. For Wikipedia Detox datasets (Aggression, Attack and Toxicity), all personalized models received significantly lower negative log-likelihood values compared to the non-personalized TXT-Baseline. For all three tasks, the best architecture was HuBi-Medium combined with CNF. For Aggression & Attack dataset, personalization improved most cases’ results. The best results were obtained by OneHot combined with RealNVP and HuBi-Formula combined with CNF. In the case of Emotion Simple and Emotion Meaning, personalization also reduced model uncertainty in most of the cases. For both datasets, the best results were obtained by the HuBi-Medium model combined with RealNVP. It is worth noting that compared to Gaussian Mixture Model, Normalizing Flows always obtain lower negative log-likelihood values. It suggests that target variables, i.e., emotions, have complex distributions, and using a simple probabilistic approach is not enough.

b) Results of Experiment 2.: In the second experiment, we carried out hyperparameter tuning on the most challenging dataset: Emotion Meanings, Tab. II. All possible combinations of hyperparameters were considered when performing the grid search. The results seem to confirm earlier speculations about MAF’s better ability to deal with multidimensional problems compared to other approaches. Moreover, none of the variants indicated the benefit of using text alone as input.

c) Results of Experiment 3.: In the third experiment, we compared results obtained by deterministic models and Emotional Normalizing Flow, Tab. III. It was carried out on two datasets: combined Wikipedia Detox: Aggression & Attack and Emotions Simple. We also decided to use only two Normalizing Flow Models that performed the best on both of these datasets: RealNVP and CNF.

In the case of Aggression & Attack dataset, the results obtained by deterministic models were better for every architecture, including the non-personalized one. In the case of Emotion Simple dataset, the results obtained by probabilistic models significantly outperformed deterministic models. The best model was a combination of HuBi-Medium and CNF. This result seems to confirm that the probabilistic approach is especially effective on complex multi-dimensional tasks.

d) Results of Experiment 4.: In the fourth experiment, we mixed the deterministic approach with the probabilistic, to create a hybrid model, Tab. IV. In both Aggression & Attack and Emotion Simple tasks, the results obtained by the hybrid approach outperformed previous methods by a large margin. For Aggression & Attack dataset, the best model turned out to be HuBi-Medium with CNF. For Emotion Simple dataset, HuBi-Medium with both RealNVP and CNF performed comparably well. The results of this experiment prove, that adding information about the model uncertainty makes big difference in the inference process, and helps to predict better for difficult and subjective tasks.

VI. Conclusions

In this paper, we proposed a novel Emotional Normalizing Flow approach to personalized NLP that opens up new perspectives on predicting reader behavior in a non-deterministic way. From the perspective of psychology and the variability of emotion sensation over time, the problem of emotion recognition is one of the most difficult and subjective tasks facing NLP. People do not perceive their emotions as zero-one, and most of the attempts so far classified their feelings in this way. The presented probabilistic approach based on normalizing flows provides more complex information about the uncertainty and diversity of possible emotional states. A comparative analysis of models for emotion recognition without and with personalization indicated that new methods are also effectively applicable in a non-deterministic setup. The generalized, non-personalized solution generates a completely different concentration of probability mass, directed toward a
work will focus on applications of our approach to some other new state-of-the-art on two very challenging tasks. Our future complex and subjective behaviors such as recognizing hate model uncertainty significantly improves the ability to predict single user. Finally, we showed that adding information about the problem in a contextual way by dedicating reasoning to a quantitative approach. Personalization can shift the view of the problem in a contextual way by dedicating reasoning to a single user. Finally, we showed that adding information about model uncertainty significantly improves the ability to predict complex and subjective behaviors such as recognizing hate speech or emotions in a text. The hybrid model we created significantly outperformed the previous methods, becoming a new state-of-the-art on two very challenging tasks. Our future work will focus on applications of our approach to some other tasks such as active or reinforcement learning.

### LIMITATIONS

One important issue related to the nature of normalizing flows is their ability to convert probabilities to disambiguate uncertain answers. At the moment, there are no reference datasets available that contain text and annotator information simultaneously with multiple markings of the same text by the same person. This is due to cost constraints in preparing such data. However, we have conducted experiments on datasets with different characteristics in which (1) one person marked several hundred texts [Wikipedia Detox Datasets] and (2) one text was evaluated dozens of times by different people [Emotions Simple and Emotion Meanings datasets]. In order to address the problem mentioned in the introduction, one text should have $N$ annotations from the same person, e.g., a few days apart. If we gain access to or prepare such a dataset, we would be happy to conduct in-depth studies on it.

Due to the language of one of the datasets being different from English, a multilingual model was used to embed the text. This decision was made in order to allow for direct comparisons and cross-referencing. This would have added an unnecessary layer to the already relatively complex problem that was addressed. It is possible to experiment with other language models as well using the source codes provided\(^1\).

### ACKNOWLEDGMENTS

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. The work conducted by Maciej Zieba and Patryk Wielopolski was supported by the National Centre of Science (Poland) Grant No. 2021/43/B/ST6/02853.

### REFERENCES

\(^1\) [https://github.com/CLARIN-PL/personalized-emotion-prediction-with-normalizing-flows](https://github.com/CLARIN-PL/personalized-emotion-prediction-with-normalizing-flows)


APPENDIX

A. Personalization architectures

Architectures for personalization have been modified to return a probability distribution instead of the probability of a class. Input features are passed to the normalizing flow as context. We are dealing with four architectures:

- Baseline (Fig. 3) - the input is just an embedding of text
- OneHot (Fig. 4) - the user represented as a one-hot vector is concatenated to the text embedding
- HuBi-Formula (Fig. 5) - the deviation of the user’s response is taken as its representative feature
- HuBi-Medium (Fig. 6) - annotation-based learned user embedding combined with text embeddings provides the context

B. Implementation details

Experiments were performed using the code provided in the “Anonymous”.

For all experiment purposes, we used a machine with AMD Ryzen Threadripper PRO 3955WX 16-Core Processor CPU, 2 x NVIDIA GeForce RTX 3090 GPUs, and 256 GB RAM.

C. Visualization of probabilities

For the combined set of Aggression & Attack, visualizations of the waveform of the probability function were prepared as a result of Experiment 3. described in the publication, are presented in Fig. 7 and Fig. 8.
**Fig. 5:** HuBi-Formula architecture utilizing normalizing flows.

**Fig. 6:** HuBi-Medium architecture utilizing normalizing flows.

**Table V:** Hyperparameters for Normalizing Flows and their possible values during experiment 2. Note that for MAF, the dropout probability hyperparameter was not used at all.

<table>
<thead>
<tr>
<th>Hyperparameter</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>hidden features</td>
<td>[2, 4, 6, 8]</td>
</tr>
<tr>
<td>num layers</td>
<td>[1, 2, 3, 4, 5]</td>
</tr>
<tr>
<td>num. blocks per layer</td>
<td>[1, 2, 3, 4]</td>
</tr>
<tr>
<td>dropout probability</td>
<td>[0.0, 0.1, 0.2, 0.4]</td>
</tr>
<tr>
<td>batch norm within layers</td>
<td>[True, False]</td>
</tr>
<tr>
<td>batch norm between layers</td>
<td>[True, False]</td>
</tr>
</tbody>
</table>

**Table VI:** Hyperparameters for training and architectures, and their possible values during experiment 2.

<table>
<thead>
<tr>
<th>Hyperparameter</th>
<th>Values for TXT-Baseline &amp; OneHot &amp; HuBi-Formula</th>
<th>Values for HuBi-Medium</th>
</tr>
</thead>
<tbody>
<tr>
<td>embedding dim.</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>hidden dim.</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>output dim.</td>
<td>[128, 256, 512, 768]</td>
<td>[128, 256, 512, 768]</td>
</tr>
<tr>
<td>learning rate</td>
<td>[1e-5, 1e-4]</td>
<td>[1e-5, 1e-4]</td>
</tr>
</tbody>
</table>

**Fig. 7:** Visualizations of the waveform of the probability function for RealNVP with different architectures and for Attack and Aggression.

**Fig. 8:** Visualizations of the waveform of the probability function for CNF with different architectures and for Attack and Aggression.