2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII) | 978-1-6654-0019-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/ACII52823.2021.9597451

Multi-dimensional Affect in Poetry (POCA) Dataset: Acquisition, Annotation and Baseline Results

Akbir Khan*
Tractable AI
London,United Kingdom
akbir.khan@tractable.ai

Jack Hopkins
Point 72 Ventures
London,United Kingdom
jack.hopkins@p72.com

Hatice Gunes
Dep. of Computer Science & Technology
University of Cambridge, United Kingdom
hatice.gunes@cl.cam.ac.uk

***Abstract*—Detecting emotions and affect in text has received enormous attention in recent years, and yet majority of the works in this area reduce the nuanced emotional responses into 'positive', 'negative' and 'neutral'. In this paper, we introduce a novel multi-dimensional affect in poetry (POCA) dataset for sentiment analysis annotated using the Geneva Emotion Wheel (GEW), to capture and analyse the multi-dimensional affect evoked in listeners. The POCA dataset is based on poems and their corresponding recitals from an online poetry database where recitals are curated by the website, and performed by the poet or an approved artist. The POCA dataset contains 330 poems (text and audio), from the English language, each of which is annotated across 20 different emotion classes, by 5 listeners. A subset of the dataset (50 poems) have also been annotated by an inlab study by 3 listeners each while their Electrodermal activity (EDA) was being recorded. As a proof of concept, we (i) introduce representative problem formulations to be addressed by machine learning approaches using the POCA dataset, from single emotion recognition (e.g., does this poem evoke joy?) to continuous affect prediction (e.g., what level arousal and valence does this poem evoke?), (ii) provide baseline results for text-based affect recognition using several classification and regression models, and (iii) provide baseline results for EDA-based affect prediction. Our results show that (i) for text-based affect recognition, classical approaches can provide as accurate results as their fine-tuned neural network counterparts, and (ii) in EDA-based affect prediction, in general there is a strong relation between the EDA signals and the self-reported valence and arousal quadrants, while predictions are better for arousal than valence.**

***Index Terms*—poetry, natural language processing, multi-dimensional affect, sentiment analysis, affective computing**

I. Introduction

Poetry is an advanced form of linguistic communication that has been largely omitted from prior sentiment analysis work. Poetry is more challenging than prose because in addition to conveying connotative and denotative information (what is said), it also contains phonological information (how it is said). Furthermore, it is a medium specifically designed to evoke aesthetic emotions and affectual responses in the reader and/or listener. However, most existing sentiment analysis is limited to binary valence classification tasks on prose datasets.

In this work, we introduce a novel multi-dimensional affect in poetry (POCA) dataset for sentiment analysis annotated using the Geneva Emotion Wheel (GEW) and the affect dimensions of valence and arousal. The POCA dataset comprises 330 annotated poems (written in English), and was developed to address several limitations in previous contributions, namely, the lack of affectual expressivity and the lack of phonological information in sentiment analysis and affectual analysis tasks. This dataset was developed to capture and analyse the multiple emotions evoked in a listener and to evaluate the ability to fine-tune [4] and few-shot learn [14] large scale models that were pre-trained, for instance, by self-supervised learning. Our hope is that the POCA dataset will aid the development of novel machine learning and NLP methodologies towards understanding human emotions felt during poetry recitals.

To address the issue of affectual expressivity, we adopt the Circumplex Model of core affect [33], in which a textual utterance can invoke any affect existing within a continuous plane defined by valence and arousal axes, which can be mapped to an emotional class [35]. To ensure that phonological information - critical in poetry - was included, we presented our listeners with audible recitals of the underlying texts rather than the poetic texts themselves. In addition to providing the category of the invoked emotion, we include fine-grained real value scores that provide information on the intensity of the emotion felt. We provide mean and variances across all emotion classes and measure inter-annotator agreement utilising these. We also present baseline results on this new dataset, in which classical methods such as Support Vector Machines and Naive Bayes are able to outperform deep learning models such as fine-tuned BERT model for regression.

II. Related Work

A. Affect and Poetry

The link between affect and art has long been an area of study within psychology [37]. This link has been explored in multiple modes, including film [2, 8], music [5, 11], and poetry [38].

*A. Khan finalised this study as part of his MPhil in ACS Degree at the University of Cambridge.

978-1-6654-0019-0/21/$31.00 ©2021 IEEE

Meditation on Statistical Method

Plato, despair!
We prove by norms
How numbers bear
Empiric forms,
How random wrong
Will average right
If time be long
And error slight,
But in our hearts
Hyperbole
Curves and departs
To infinity.
Error is boundless.
Nor hope nor doubt,
Though both be groundless,
Will average out.

J. V. Cunningham

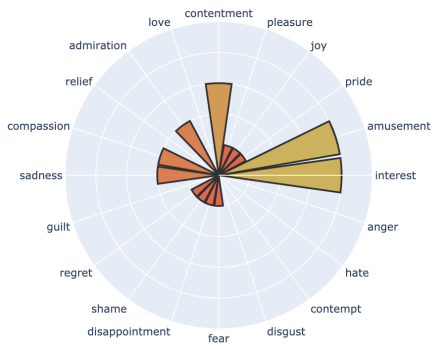

Fig. 1. A representative poem from the POCA dataset together with the *rose plot* of the emotions reported by five listeners.

Poetry is an important form of literature in which the use of phonaesthetics, symbolism and prosody is designed to evoke an affectual response in the reader. As such, it forms the ideal avenue to evaluate the efficacy of NLP systems specifically geared toward capturing affect. There has been some work centred around poetry within the NLP community. These works have mainly focussed on poem generation [17] in specific contexts, be it poetic style, theme or form [20]. The work in [26] uses emotions to condition lexicon, however this fails to fully appreciate how affect can be conveyed within the aesthetic, rhyme and form.

B. Affect and Natural Language Processing

NLP has historically analysed affect through textual opinion mining tasks, such as Sentiment Analysis [9, 10, 24, 30]. These tasks differ by the type of text (e.g tweets, poems, reviews) and the granularity of their annotations (good/bad, happy/sad/neutral, anger/sadness/happiness/etc).

Many sentiment analysis tasks focus on which words or sentences convey positive or negative affective information. Annotations are generally binary (positive/negative) and thus only use valence to express connotation [10]. Valence is commonly used in these tasks because it is a simple measure with direct real-world applications, e.g predicting whether a stock price will increase using the valence of relevant tweets [30]. This simplicity leads to greater inter-annotator agreement - an unsolved problem when dealing with aesthetic emotional responses, which tend to be complex, subjective and non-deterministic [9]. For this reason, affect datasets which have been annotated with labels other than valence are sparse.

Several of the existing datasets adopt the model based on Ekman's [15] six basic emotions (Neutral, Joy, Sadness, Fear, Anger, Surprise, and Disgust). For example, Emotionlines [13] and its multimodal successor MELD [32] provide annotations to the script of the famous sitcom Friends. However each segment of dialogue is specifically labelled as one of these emotions - restricting the expressivity of the model [18, 28]. Further critics of Ekman's model observe that these 6 basic emotion categories are not applicable to various domains including aesthetic experience, and have been reported to be not universal across different cultures [1, 22].

Recently diverse annotations have been applied to a corpus of tweets [27]. Here we find three main distinctions compared to our work. Firstly, it relies on a corpus of tweets that may lack sufficient context to invoke nuanced affectual responses such as 'trust' and 'anticipation'. Secondly, annotators were asked 'What was the tweeter feeling?' as opposed to our work which relies on individuals self-reporting on how the poem made them feel. Finally, whilst they are motivated by the PAD (Pleasure, Arousal and Dominance) model of affect [25], they do not provide the values of the annotations along these dimensions.

Affect-based analysis of poetry has been considered, but the choice of annotation varies. Binary sentiment has been mined in Chinese poems [21], and Ekman-motivated fine-grained annotation ('Joy', 'Anger', 'Fear' and 'Sadness') has been evaluated in Spanish [3]. Within English, PO-EMO consists of a corpus of 64 poems, with the annotations motivated by the aesthetics of poetry ('Beauty', 'Sadness', 'Uneasiness', 'Vitality', 'Suspense', 'Sublime', 'Humor', 'Annoyance', and 'Nostalgia') [7]. To the best of our knowledge, our work is the first one to introduce a poem dataset for affect recognition annotated with the Geneva Emotion Wheel.

III. Data Collection

We first extracted poems and their corresponding recitals from an online poetry database [31]. We collected 13,250 published poems, covering a range of subjects including, Love, Nature, Social Commentaries, Religion, Living, Relationships, Mythology, and Folklore. The collection also included forms such as, Free Verse, Blank Verse, Syllabic and Common Measure. Recitals are curated by the website and performed by the poet or an approved artist, suggesting these recitals are

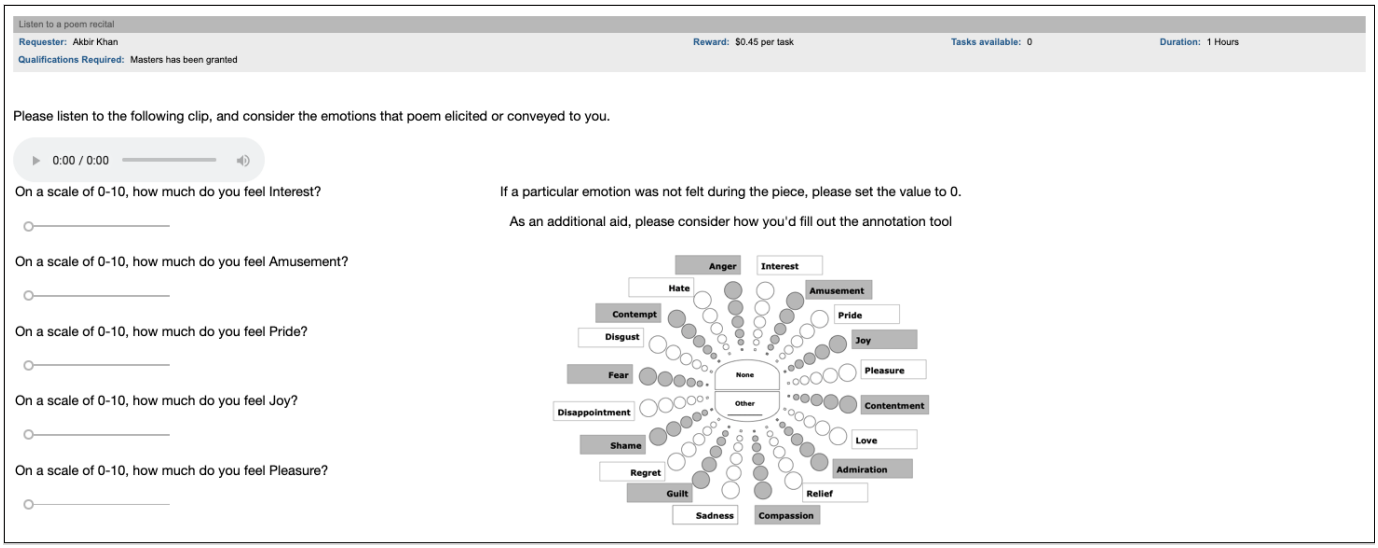

Fig. 2. Online tool for recording poem annotations.

representative of the poet's true feelings. The distribution of poem lengths is shown in Fig 3.

We then performed an in-lab study to ensure that the chosen method of poetry recitation actually invoked affectual responses in listeners [38]. Following this, we conducted a larger online study to increase the scope of the dataset. In both studies we adopted the Geneva Emotional Wheel (GEW) [36] as the annotation tool. This was motivated by its direct mapping to the Circumplex Model [33] and its extensive use in other emotion related studies [12] as well as its recent adoption for non-linguistic musical prosody listening tasks [34]. The GEW is a theoretically derived and empirically tested instrument to measure emotional reactions to objects, events, and situations [35]. Like the Circumplex Model it was primarily developed as a tool for self-evaluation, and its utility as a tool to measure the perceived affect has already been empirically validated [35].

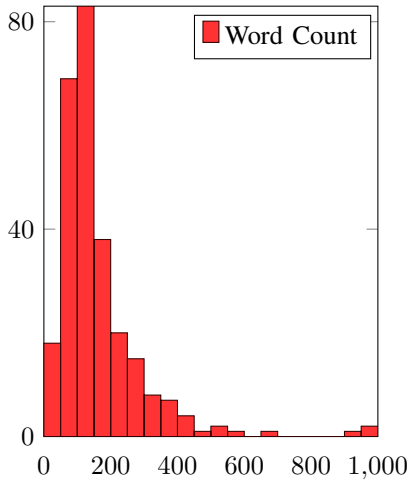

Fig. 3. Histogram showing the distribution of word counts in each poem across the dataset.

A. The In-Lab Study

We initially conducted a controlled in-lab study to ensure that the poem recitations indeed evoked emotions in listeners (participants) in a measurable manner. We conducted this study with 30 participants, each of whom listened to 5 poems ensuring that each poem received emotion ratings / annotations from 3 listeners.

Participants were first introduced to the study and were asked to wear an Affectiva Q[1] sensor on their non-dominant wrist. Each session began with listening to 2 minutes of relaxing music to establish baselines. After this, the participants were presented with the web-based annotation tool illustrated in Figure 2. They were allowed to ask questions at this point to clarify the annotation task. After this was completed, they were asked to listen to a pre-recorded audio recital of a single poem twice, with a 30 second break between the two recitals. After listening to the second recital, the participants were asked to mark 'the emotions that poem made them feel'. To further validate the participants' affective experiences, Electrodermal Activity (EDA) was recorded by the Affectiva Q sensor worn on the participant's wrist.

All participants were native English speakers from the University of Cambridge, UK with 14 out of 30 identifying as female, and ages ranging from 19 to 34. Participants' poetry enthusiasm was evenly distributed, ranging from 3 individuals that had listened to/read a single poem in the last year, to 10 individuals having listened/read over 30 poems in the last year.

[1]https://affect.media.mit.edu/projectpages/iCalm/iCalm-2-Q.html

B. The Online Study

We next expanded the study to 1,445 participants via the Amazon Mechanical Turk (AMT) crowd-sourcing platform. In this experiment, again participants' emotional response was recorded via the GEW, however EDA reporting was unfeasible. To minimise random annotations and to ensure that high quality labels were provided, we restricted the pool to Masters certified workers and those who have already completed at least 500 Human Intelligence Tasks (HITs) while retaining a greater than 90% approval rate.

Each poem was presented to 5 listeners (workers) using the interface shown in Figure 2. Participants spent roughly 4 minutes per poem, and were paid \$0.45 per poem [2]. Over a two week period, we manually inspected and checked all poetry responses. We rejected 2% of those which were outright wrong, that had either the maximum or minimum affectual response reported. Unlike the in-lab study, participants were able to listen to a poem as many times as they wished - however a single session which lasted over 10 minutes was rejected as part of the annotation quality inspection.

1) Annotations Provided: POCA consists of 330 poems with their content, audio recital, 20 emotion labels on the GEW with intensities marked for each emotion. We also provide an additional set of 12,811 unlabelled poem texts for self-supervised learning tasks [31]. For each poem we aggregated listener annotations and calculated the mean and variance of each label (over the 5 listeners), as well as the inter-rater agreement values. To enable research work on the variability of affectual response, POCA dataset also contains the original listeners responses / annotations.

Emotion labels provided are intensities labelled along [0,5] on a discrete scale covering the 20 categories provided in the GEW: Interest, Amusement, Pride, Joy, Pleasure, Content, Love, Admiration, Relief, Compassion, Sadness, Guilt, Regret, shame, Disappointment, Fear, Disgust, Contempt, Hate and Anger. The range of emotion intensities provided by the listeners is presented in Figure III-B in terms of rose plot[3] for both the online and the in-lab study, demonstrating the variety of affect with POCA.

Similarly to [19], we additionally calculated valence and arousal values as the projection of each emotion onto the Valence-Arousal space. Each emotion is given a specific angle (θ_i) and a radial component depending on its position on the GEW and its reported intensity.

2) Inter-Rater Agreement: To measure the inter-rater agreement and to further differentiate it from poor quality annotations, we used two different metrics for inter-rater agreement, Krippendorff's alpha and rater variance. Krippendorff's alpha [23] is a generalization for Fleiss Kappa to ordinal labels and experiments with missing data (in our situation, where no single annotator labels all poems). Values are reported on a scale of -1 to 1 (between no agreement to full agreement).

Responses were assessed across 100 ordinal labels (i.e., 20 affect categories along 5 scales), thus a low agreement is expected. Over these 100 ordinal labels, for the in-lab study we obtained an average agreement level of 0.19 and for the online study, an agreement level of 0.09 (with a maximum agreement level of 0.65).The rater variance was taken over the set of normalised scores from each listener per poem. Using the norm of the variance vector for each poem, for the in-lab study and online-study we obtained an average variance of 0.11 and 0.42, respectively. Reducing the label space from 100 to 20 affect categories, by removing intensities, provided us a Krippendorff's alpha measure of 0.524, and a Pearson Correlation Coefficient of 0.295 when using only the intensity scales.

One might argue that we need to increase the amount of labelled data in the POCA dataset to obtain a more balanced label distribution. However, sentiment tasks are known to have a low inter-annotator agreement [9], a problem which is exacerbated when listeners use more expressive labelling systems. In addition, poetry is an intrinsically aesthetic and personal experience, balance is difficult to guarantee in poetry listeners due to this process being inherently subjective, and as such an affectual diversity of responses is indeed expected.

IV. Proof-of-Concepts Experiments and Results

A. Baseline Machine Learning Tasks

POCA is a rich dataset, and can be used for solving a range of research problems at different levels of abstraction, from single emotion detection ('does this poem evoke joy?') to predicting continuous affect values ('what level arousal and valence this poem evokes?'). As a proof of concept we present a representative set of tasks as follows.

Emotion Classification: 'How accurately we can determine whether a poem elicits a certain emotion - e.g. sadness?' - Any poem with a sadness[4] score below a threshold could be considered to not be a sad poem.

Emotion Intensity Prediction: This could also be formulated as the more difficult task of 'How accurately we can predict how much sadness this poem elicits?' by directly mapping the poem to intensity scores.

Multi-label Emotion Prediction: 'How accurately we can predict the distribution of emotional responses to this poem?'. Utilising the listener reported mean and variance of emotions per poem, with a specific prior, can we calculate the likelihood that a poem will elicit a specific emotional response?

Sentiment Classification: 'How accurately we can determine the overall positivity or negativity a poem evokes in listeners?' - Using the sign of the valence label for a poem, we can infer if it has positive or negative sentiment.

Dimensional Affect Prediction: 'How accurately we can predict the arousal/valence values a poem evokes?' - Using the continuous affect labels for a poem, we can create a mapping from its textual content to these values, evaluating a model's predictive ability.

[2]The workers were paid \$6.75 per hour, the living wage in countries they were sourced from.

[3]https://www.itl.nist.gov/div898/software/dataplot/refman1/auxillar/roseplot.htm

[4]This can be applied to any of the 20 distinct emotional classes.

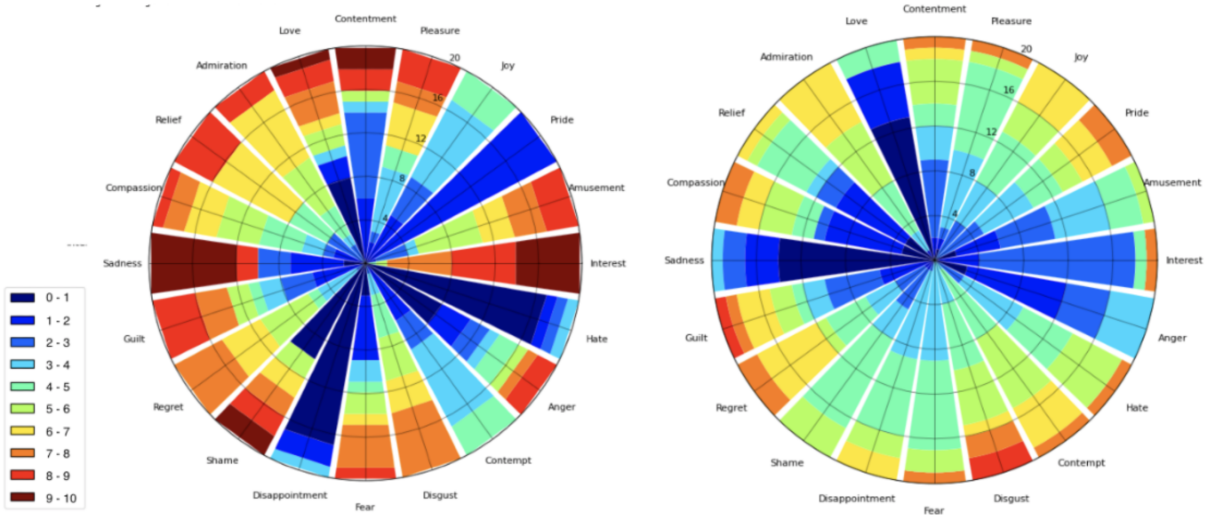

Fig. 4. Density of the annotations, in terms of rose plot for the online (left) and the in-lab study (right) for *all* emotion classes self-reported as *felt* for the poem recitals the participants listened to.

TABLE I
SUMMARY OF DATASETS UTILISED, THEIR SAMPLE SIZE AND CLASS IMBALANCE RATIO.

Name	Task	Dataset Size	Class Imbalance
IMDB	Classification	49582	1:1
SST2	Classification	215154	4:3
EmotionLines	Classification	29245	3:100
POCA		330	
Sadness	Classification	-	3:5
Joy	Classification	-	2:3
Sentiment	Classification	-	1:4
Dimensional Affect	Regression	-	n/a
Label Uncertainty	Regression	-	n/a

Prediction of Label Uncertainty: 'Can we predict the inter-rater agreement of a poem?' - Inter-rater agreement is described by a continuous number from [-1, 1] for a poem. Currently inter-rater agreement is used to evaluate the reliability of listener labels.

EDA-based Emotion Classification/Prediction: "How accurately we can classify/predict the emotion(s) and arousal/valence a poem evokes in a listener by analysing their EDA signal?" - Using the EDA signal recorded for a listener together with the continuous emotion labels provided by the listener for a poem, we can create a mapping from EDA signals to the valence/arousal values, evaluating a model's predictive capability. Our EDA analysis can be seen as a triangulation technique to analyse and understand both subjective (label) and objective (signal) responses to poetry, and to provide another baseline for comparison.

B. Text-Based Affect Recognition

We performed a set of experiments to help contextualise the dataset and its contributions as compared to existing datasets and works in the NLP literature.

1) Tasks: We train models for a number of tasks to demonstrate the richness of the POCA dataset. To provide a comparison to existing sentiment analysis tasks, we conduct experiments using the IMDb and the Stanford Sentiment TreeBank (SST2) datasets. To provide a comparison to newer affect analysis tasks, we conduct experiments using the Emotionlines dataset and provide experiments for four novel tasks that are formulated using the POCA dataset. In all experiments, we use 80% of the data for training, 10% for validation, and the remaining 10% for testing, unless stated otherwise. We summarise task information in Table I.

IMDb is a dataset of 348,415 reviews collected on movies [24]. The dataset is labelled by 2 classes, 'positive' and 'negative', both classes are balanced. **SST2** consists of 11,855 single sentences extracted from movie reviews [10]. Similar to the IMDb, it is labelled by 2 classes. **EmotionLines** is a dataset of 29,245 utterances from a famous TV sitcom. Each utterance is labelled to be one of 6 emotions. Here we look in particular at the ability to correctly predict an utterance was labelled 'sadness'.

POCA Sadness Classification uses 330 poems from the POCA dataset that are labelled either 0 or 1 depending on a the mean sadness intensity recorded. A threshold of 0.3 (mean of the dataset) was used to partition the dataset into those containing sadness and those that do not. **POCA Joy Classification** is a dataset created in a similar fashion, using the label of 'joy' instead. **POCA Sentiment Classification** uses 330 poems from the POCA dataset that are labelled as either 0 or 1. We calculate labels by inspecting the sign of the valence intensity for each poem. No poems in the dataset contained a valence value of 0. **POCA Dimensional Affect Prediction** uses 330 poems from the POCA dataset, each labelled by a pair of real numbers from the interval $[-0.3, 0.3]^2$. Values relate to the mean affect value (arousal/valence) reported by the participants listening to those poems. **POCA Prediction of Label Uncertainty** task utilises the fact that each poem is labelled by a single value from the interval $[-1, 1]$ related to the

TABLE II
PERFORMANCE OF DIFFERENT MACHINE LEARNING MODELS ON BASELINE TASKS AND POCA TASKS.

Task	metric	SVM/SVR	Naive-Bayes	Bi-LSTM	BERT	Dummy	Single-Annotator
IMDB	F1-Score	0.878	0.843	0.793	**0.910**	-	-
SST2	F1-Score	0.846	0.826	0.892	**0.912**	-	-
EmotionLines	F1-Score	0.514	0.483	0.539	**0.596**	-	-
POCA							
Sadness Classification	F1-Score	0.003	0.003	0.285	**0.385**	0.469	0.745
Joy Classification	F1-Score	0.526	0.526	0.794	**0.825**	0.539	0.521
Sentiment Classification	F1-Score	**0.751**	0.003	0.331	0.425	0.367	0.668
Continuous Affect Prediction	RMSE	**0.108**	1.13	0.202	0.152	0.131	0.162
Prediction of Label Uncertainty	RMSE	**0.191**	3.94	0.381	0.231	0.260	0.310

Krippendorff's alpha, a measure of its inter-rater agreement.

2) Machine Learning Models: We experimented with a number of different machine learning models and classifiers, using several popular sentiment analysis methods to provide fair comparisons. Our experiments evaluated the following five models: Support Vector Machines (SVMs), Naive Bayes Classifiers, Bidirectional Long Short-Term Networks (Bi-LSTM) and fine-tuned BERT [4]. The results obtained were further compared to an additional single human annotator (single-annotator).

Classical Methods: We train both SVMs, and Naive Bayes for sentiment analysis, on the TF-IDF bag of words feature representations. For continuous labels we train Support Vector Regression or a Gaussian Naive Bayes method, scaling where appropriate. For discrete labels we use either Support Vector Machines or multivariate Bernoulli distributions. We optimise parameters for both classifiers using grid search conducted over the validation set.

Bi-LSTM: Our model consists of a bidirectional LSTM (hidden size 50) with recurrent dropout (0.1) and global max pooling following the LSTM. We feed this (fixed-length) representation through a fully connected layer with ReLU activation (hidden size 50) and then a fully connected output layer with softmax. We use the WordPiece tokenizer [4] and set a max sequence length of 500. We train using an Adam optimiser, with a learning rate of $1e-3$ and batch size of 32.

BERT: We use an off-the-shelf uncased base BERT [5], fine-tuning for each task. We apply a max pool layer over the final hidden states, and through a fully connected layer. To account for BERT's sub-word tokenization, we set the maximum token length to 500. We fine-tune BERT up to 5 epochs with the same using the BERT Adam optimizer with a batch size of 16 (to fit on a Tesla V-100 GPU). We used a linear learning rate scheduler learning rates with initial rate of $5e-5$ and a burn in of 10 iterations. We used a dropout of 0.1.

Neural Nets: All neural network classifiers were implemented in PyTorch. For tasks involving discrete labelling (IMDb, SST2, EmotionLines, Sadness/Joy/Sentiment classification) we use Cross Entropy as the loss function and final layer is a softmax function. For classification tasks that utilise continuous labels, we use Mean Squared Error and a tanh layer.

[5]https://huggingface.co/bert-base-uncased

Dummy: For completeness we include an sklearn dummy classifier. This is a model which predicts models using a trivially simple rule. In our case, we use a uniform distribution to generate label predictions uniformly at random, i.e., 'chance score'.

Single-Annotator Baseline: In order to compare classifier performances to a human baseline, we asked one additional listener to undertake the listening and annotation tasks with the POCA dataset - i.e., the same portion of the data that the ML models were tested on. They were also asked to provide a value [0,1] to the question 'How varied the emotions evoked (in listeners) by this poem will be?' . This refers to the listener / annotation agreement level for each poem they listened to - i.e., predicting the label uncertainty.

3) Results: We present the experimental results in Table IV-B. We report F1 scores for classification tasks and Root Mean Squared Error (RMSE) for regression tasks. For relevant models (Bi-LSTM/BERT), we conduct 10 repeat runs and report the mean score over these runs.

Classifier performances on the discrete classification tasks follow the expected trend in sentiment analysis, with BERT outperforming other methods. We posit that this is because these tasks are more similar to the training process and data (the existing knowledge) of the network. We also found that while BERT's pre-training approach is successful at more superficial tasks such as 'Sadnesss Classification', its performance decreases on more complex tasks, namely 'Dimensional Affect Prediction' and 'Prediction of Label Uncertainty'. These are aspects (dimensional affect and uncertainty) that are less likely to be represented in its training corpus or highlighted by its language model and sentence classification training tasks.

The Bi-LSTM performance is weak on the majority of the tasks, which we attribute to two factors: training data and document length. When comparing the Bi-LSTMs performance on the POCA dataset against traditional tasks it is clear that the small number of training samples significantly impacts the classification and prediction performance. Secondly we note that documents within the POCA dataset (see Fig 3) are much longer than the media on which Bi-LSTM performs well (single sentences and utterances).

In comparison, classical methods perform relatively well on the POCA prediction (regression) tasks. These methods are

TABLE III
LIST OF THE FEATURES EXTRACTED FROM THE EDA PHASIC AND TONIC COMPONENTS

Feature	Description
Max-Tonic/Phasic	Max value of signal
AUC-Tonic/Phasic	Area under the curve
Mean-Tonic/Phasic	Mean value of signal
STD-Tonic/Phasic	Standard deviation of signal
Deriv-Tonic/Phasic	Mean of derivative of signal

neither as sample intensive as Neural Networks or dependent on the length of the sequence input. Their failure to perform well on the classification tasks demonstrates that TF-IDF vectors are likely to be an ill-suited representation for poetry. We note that the Naive Bayes method fails in prediction (regression), which has previously been reported [16].

Finally, we note that on the prediction / regression tasks the performance of the single-annotator is also not in high agreement with the original annotations, and therefore appears to be somewhat poor. It is much easier for a listener to answer the question of 'on a scale of 0-5 how sad does this poem make you feel?' as opposed 'how negative does this poem make you feel?', as humans are more familiar with the emotion of sadness as compared to the more generic concept of valence (negativity in this case). Indeed, this is likely why annotation tools like the GEW are useful, particularly for assessing multi-dimensional affect experienced when listening to poetry. In general we noted that the single-annotator predicted much higher annotation agreement values (lower variance) than the original values collected during the online study, suggesting that it is difficult to predict other people's emotional reactions to poetry recitals.

We surmise that the low scores for SVM/R and Naive Bayes are a result of the complexity of poetry as a form of communication i.e., certain aesthetic features of the poem like rhythm, which contribute to the affect of a listener, arise from word order.

C. EDA based Affect Recognition

EDA was recorded during the inlab study with 30 participants, each of whom listened to 5 poems, resulting in 150 EDA readings and their corresponding emotion labels provided by the participants. In this section we present the feature extraction and affect recognition experiments we have conducted on this data. This relates to the **EDA-based Emotion Classification/Prediction** task we have formulated in subsection A above.

1) Feature Extraction: There are multiple techniques to extract signals from continuous signals. In particular, for EDA signal analysis, the most utilised approach is the Continuous Deconvolution Analysis (CDA) [6, 29] which is based on an explicit biophysical model, and optimisation of its parameters. Included is a routine preprocessing step where errorenous sections (artificats) are detected and removed (by human inspection). A cut of 2Hz is also imposed in order to limit the frequency bandwidth of the EDA signal.

TABLE IV
SPEARMAN CORRELATION COEFFICIENT (CC), SIGN AGREEMENT (SAGR) AND ROOT MEAN SQUARED ERROR (RMSE) BETWEEN THE PREDICTED AND THE GROUND TRUTH VALUES.

Prediction	CC	SAGR	RMSE
Arousal	0.54	0.72	0.92
Valence	0.08	0.73	0.32

CDA decomposes the signal into its *tonic* and *phasic* phases, which is followed by extracting a set of features from both. We choose to extract features similar to prior tasks in emotion classification using EDA listed in Table III. Furthermore, we calculated additional two features, the mean and standard deviation of the normalised signal. This resulted in extracting a 12 -dimensional vector for each signal.

2) Experiments: We utilised a Multilayer Perceptron as the regression model for predicting affect from the extracted EDA features. We incorporated a single dense layer which maps directly from the features to the label. We used a 5-fold cross-validation protocol for evaluation and estimated the Spearman Correlation Coefficient (CC), Root-Mean-Squared (RMSE) and Sign Agreement (SAGR) between the model predictions and the ground truth. To accommodate for random model initialisations, we conducted 100 repeats per task, reporting mean results in Table IV.

We observe a large discrepancy in performance between the SAGR and the CC. We attribute this to the metrics evaluating the model at different scales. The results show that whilst the model is unable to predict the precise dimensional affect values well, in general there is a strong relation between the EDA signals and the self-reported valence - arousal quadrants. Overall, the model predictions are better for arousal than valence.

V. CONCLUSION

In this paper we presented the Multi-dimensional Affect in Poetry (POCA) Dataset, the first poem dataset annotated with multi-dimensional affect using the Geneva Emotion Wheel (GEW) that is publicly available for research purposes[6]. The POCA dataset contains two subsets. The *inlab subset* contains 50 poems each annotated in a lab setting by 3 listeners across 20 different emotion classes while their Electrodermal activity (EDA) was being recorded. The *online subset* contains 330 poems, each of which is annotated across 20 different emotion classes, by 5 listeners using crowdsourcing. In order to facilitate future research using this dataset, we introduced representative problem formulations and provided baseline results for text-based affect recognition and EDA-based affect prediction. We demonstrated that with its poems, annotations and EDA recordings, the POCA dataset enables the investigation of a variety of research problems at the crossroad of Natural Language Processing and Affective Computing fields which has not been possible to date due to the lack of the availability of such rich datasets.

[6]https://github.com/akbir/POCA/

References

[1] Jack et al. “Facial expressions of emotion are not culturally universal”. In: *Proc. of the National Academy of Sciences* (2012).

[2] Kimura et al. “Emotional State of Being Moved Elicited by Films: A Comparison With Several Positive Emotions”. In: *Frontiers in Psychology* (2019).

[3] Barros et al. “Automatic Classification of Literature Pieces by Emotion Detection: A Study on Quevedo’s Poetry”. In: 2013.

[4] Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *NAACL-HLT 2019*.

[5] Gold et al. “Musical reward prediction errors engage the nucleus accumbens and motivate learning”. In: *Proc. of the National Academy of Sciences* (2019).

[6] Greco et al. “Advances in Electrodermal Activity Processing with Applications for Mental Health”. In: ().

[7] Haider et al. “PO-EMO: Conceptualization, Annotation, and Modeling of Aesthetic Emotions in German and English Poetry”. In: (2020).

[8] Kaltwasser et al. “Sharing the filmic experience - The physiology of socio-emotional processes in the cinema”. In: *PLOS ONE* (2019).

[9] Rogers et al. “Rusentiment: An enriched sentiment analysis dataset for social media in russian”. In: *ICCL*. 2018.

[10] Socher et al. “Recursive deep models for semantic compositionality over a sentiment treebank”. In: *EMNLP*. 2013.

[11] M Ardizzi et al. “Audience spontaneous entrainment during the collective enjoyment of live performances: physiological and behavioral measurements”. In: *Scientific Reports* (2020).

[12] J. Bardzell et al. “Emotion, engagement and internet video”. In: *Emotion* (2008).

[13] SY. Chen et al. “Emotionlines: An emotion corpus of multi-party conversations”. In: *arXiv preprint:1802.08379* (2018).

[14] Y Cheng et al. “Few-shot learning with meta metric learners”. In: *arXiv preprint arXiv:1901.09890* (2019).

[15] P. Ekman et al. “Universals and cultural differences in the judgments of facial expressions of emotion.” In: *J. of personality and social psychology* 53.4 (1987).

[16] E. Frank et al. “Naive Bayes for regression”. In: *Machine Learning* (2000).

[17] P. Gervás. “Wasp: Evaluation of different strategies for the automatic generation of spanish verse”. In: *Proc. of AISB Symposium on creative & cultural aspects of AI*. 2000.

[18] H. Gunes et al. “Emotion representation, analysis and synthesis in continuous space: A survey”. In: *Proc. of IEEE Face and Gesture 2011*.

[19] A Al-Hamadi et al. “Emotional Trace: Mapping of Facial Expression to Valence-arousal Space”. In: *Current J. of Applied Science and Technology* (2016).

[20] J. Hopkins and D. Kiela. “Automatically generating rhythmic verse with neural networks”. In: *ACL*. 2017.

[21] Yufang Hou and Anette Frank. “Analyzing Sentiment in Classical Chinese Poetry”. In: *Proc. of the SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. 2015.

[22] R E. Jack et al. “Four not six: Revealing culturally common facial expressions of emotion.” In: *J. of Experimental Psychology: General* (2016).

[23] K. Krippendorff. “Computing Krippendorff’s alpha-reliability”. In: (2011).

[24] A L. Maas et al. “Learning word vectors for sentiment analysis”. In: *ACL*. 2011.

[25] A. Mehrabian. “Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament”. In: *Current Psychology* (1996).

[26] J. Misztal and B Indurkhya. “Poetry generation system with an emotional personality.” In: *Proc. of ICCC*. 2014.

[27] S. Mohammad et al. “Semeval-2018 task 1: Affect in tweets”. In: *Proc. of Int’l Workshop on Semantic Evaluation*. 2018.

[28] G. Mohammadi et al. “Towards Understanding Emotional Experience in a Componential Framework”. In: *Proc. of Int’l Conf. on ACII*. 2019.

[29] A. Natarajan et al. “Detecting divisions of the autonomic nervous system using wearables”. In: *Engineering in Medicine and Biology Society*. IEEE. 2016.

[30] T H Nguyen et al. “Sentiment analysis on social media for stock movement prediction”. In: *Expert Systems with Applications* (2015).

[31] *Poetry Foundation*. https://www.poetryfoundation.org/. Accessed: 2020-05-30.

[32] S. Poria et al. “MELD: A multimodal multi-party dataset for emotion recognition in conversations”. In: *arXiv preprint arXiv:1810.02508* (2018).

[33] J. Posner et al. “The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology”. In: *Development and psychopathology* (2005).

[34] Richard S. et al. *Emotional Musical Prosody: Validated Vocal Dataset for Human Robot Interaction*. 2020.

[35] V. Sacharin et al. “Geneva emotion wheel rating study”. In: (2012).

[36] KR. Scherer. “What are emotions? And how can they be measured?” In: *Social science information* (2005).

[37] A. Sherman and C Morrissey. “What is art good for? the socio-epistemic value of art”. In: *Frontiers in human neuroscience* 11 (2017).

[38] E. Wassiliwizky et al. “The emotional power of poetry: Neural circuitry, psychophysiology and compositional principles”. In: *Social cognitive and affective neuroscience* (2017).