On Rater Reliability and Agreement Based Dynamic Active Learning

Yue Zhang*, Eduardo Coutinho*, Zixing Zhang†, Michael Adam‡, and Björn Schuller*†

*Department of Computing, Imperial College London, London, United Kingdom
†Chair of Complex and Intelligent Systems, University of Passau, Passau, Germany
‡Machine Intelligence & Signal Processing Group, MMK, Technische Universität München, Munich, Germany
Email: yue.zhang1@imperial.ac.uk

***Abstract*—In this paper, we propose two novel Dynamic Active Learning (DAL) methods with the aim of ultimately reducing the costly human labelling work for subjective tasks such as speech emotion recognition. Compared to conventional Active Learning (AL) algorithms, the proposed DAL approaches employ a highly efficient adaptive query strategy that minimises the number of annotations through three advancements. First, we shift from the standard majority voting procedure, in which unlabelled instances are annotated by a fixed number of raters, to an agreement-based annotation technique that dynamically determines how many human annotators are required to label a selected instance. Second, we introduce the concept of the order-based DAL algorithm by considering rater reliability and inter-rater agreement. Third, a highly dynamic development trend is successfully implemented by upgrading the agreement levels depending on the prediction uncertainty. In extensive experiments on standardised test-beds, we show that the new dynamic methods significantly improve the efficiency of the existing AL algorithms by reducing human labelling effort up to 85.41 %, while achieving the same classification accuracy. Thus, the enhanced DAL derivations opens up high-potential research directions for the utmost exploitation of unlabelled data.**

I. Introduction

Within the context of Computational Paralinguistics, speech patterns can be characterised using objective and subjective measures [1]. In the case of *objective* measures (e. g. age, gender, weight), the labels attributed to speech are referred to as the 'ground truth'. On the other hand, there are speech phenomena (e. g. voice likeability, degree of interest or nativeness) that can only be reliably assessed (annotated/labelled) by perceptive judgments [2]. In consequence, the reliability of labels for the subjective speech phenomena highly depends on the annotators' stable and transient characteristics, including a myriad of subjective factors [2], [3]. Taking emotion recognition for example, there are clear gender differences in the ability to recognise discrete emotions in a variety of non-verbal domains, which indicate that women perform slightly better overall, especially for negative emotions [4]. Further, some of the variations amongst individuals in perception of emotion in the auditory domain can be attributed to personality differences, which are associated with affective biases in emotion judgment [5] due to the interaction of personality with attention, motivation and mood [6]. Other factors which may influence individual variation in the perception of emotion include emotional intelligence (which is associated with improved emotion perception abilities), and age (emotion recognition is at its peak in young adults and declines with age [7]). Therefore, in contrast to the 'ground truth' that can be measured objectively, subjective annotations lead to what is known as the 'gold standard', and are necessarily assessed by inter-rater agreement procedures. Thus, a large number of annotators is necessary to establish a well grounded reference. Unfortunately, one of the major barriers of today's research is the costly consequences of obtaining human annotations, which are time-consuming and expensive to obtain.

Given this scenario, many researchers in the area of Machine Learning (ML) developed approaches for the exploitation of unlabelled data, which is nowadays pervasive in digital format and relatively easy and inexpensive to collect (e. g. from public resources such as social media). The most common methods include Semi-Supervised Learning (SSL) [8], [9], Active Learning (AL) [10]–[12], as well as diverse combinations thereof (e. g. [13]–[15]). The essence of the conventional ML methods is to train a classifier on a small, labelled data set, and re-train the model iteratively by sequentially adding new (machine or human) labelled instances to the training set. The active learner aims at achieving greater accuracy with fewer training labels by (actively) choosing the data from which it learns, and querying human annotators for labelling. It has been shown that AL strategies significantly reduce human labelling work, while still achieving good performance levels [10]. Despite the success that has been achieved with these techniques, the methodology has converged to a degree of standardisation, and major breakthroughs have been lacking in the past years.

As aforementioned, in the case of subjective labelling tasks, reliability of ratings is of paramount importance and therefore AL techniques are preferred as labels are obtained from human and reliability can be assessed. In conventional AL, the most 'informative' instances are selected and submitted to a *fixed* number of human raters for labelling (hereinafter referred to as 'Static Active Learning' (SAL)). It is evident that applying majority voting on a fixed number of annotators for each instance is a rather inefficient method. For instance, if there are five annotators available and the first three annotate a specific instance with the same label, the annotations of the other

two annotators seem to be redundant. Consequently, there is the possibility of further reducing the amount of human annotations required by SAL, as long as we shift our perspective from standard majority voting methods to agreement-based annotation strategies.

The our previous work [16], we introduced for the first time a novel method called DAL with random query order (rDAL). rDAL is a derivation of the DAL algorithm, in which an adaptive query strategy is used to dynamically determine the number of annotators for each selected instance. The main underlying idea behind this method is to sequentially query human annotators to label a specific instance until a predefined agreement level (i. e. a certain number of votes for one common class label) is achieved, instead of requesting all available raters and then computing the gold standard by majority voting. As a consequence, the number of annotators required for each instance depends solely on the agreement level defined by the user to establish the gold standard. rDAL was shown to significantly reduce the labelling costs up to 79.17 % with the Medium Certainty (MC) strategy in relation to traditional AL methods and warrant the reliability of the subjective labelling procedure without sacrificing the classification performance.

In rDAL, the order in which the raters are queried to label a specific instance was *randomised*. Considering that some annotators (and groups of annotators) are more reliable than others, it is plausible to assume that the efficiency of rDAL can potentially be improved by querying the most reliable raters first in order to reach the predefined agreement level with less annotators. This idea motivated us to introduce and evaluating new query order strategies that consider both the individual rater reliability as well as inter-rater agreement for every possible combination of raters. Henceforth, we will refer to this method as order-based DAL (oDAL). The core underlying idea behind oDAL is that we approximate the gold standard by selecting the most reliable rater or group of raters first, hoping to achieve further reduction of the annotation costs while maximising the reliability of the gold standard. Moreover, the agreement level in these algorithms is set to a fixed value, which lead to a deterioration of classification performance as learning progressed because of the noisier instances left to be labelled at the end of the learning process. In order to overcome this limitation, a possible option is to switch to higher agreement levels for labelling noisy instances. Therefore, in this paper, we also introduce another variant of DAL – uDAL– which implements an oDAL algorithm that dynamically upgrades the agreement levels dependent on the noisiness of the instance to be labelled.

In what follows we describe the algorithms developed to integrate an efficient query order into rDAL and their applications to speech emotion recognition. In Section II, we describe our proposed algorithms and methodology. Then, in Section III and Section IV, we describe the database and feature set, respectively, which are used to demonstrate the potential of our method. The experimental setup and the results are presented in Section V. In Section VI, we discuss our findings and explore possible extensions of this work.

II. Methodology

In this section, we introduce the oDAL and the uDAL algorithms and the method for their evaluation in the context of speech emotion recognition. We employ Support Vector Machines (SVMs) as the classification model. SVMs are introduced in Section II-A as well as the concept of confidence values which are used by the algorithms to select the instances for human annotation. Then we formally define agreement level, rater reliability and inter-rater correlation, which form the basic concepts underlying the DAL algorithms. Finally, we describe the oDAL and uDAL algorithms in Section II-D.

A. SVMs and Confidence Levels

Similar to traditional AL, the dynamic active learner actively selects the data from which it learns by considering the prediction uncertainty of the trained classifier in terms of confidence values. For this purpose, we apply Support Vector Machines (SVMs) that construct decision hyperplanes to separate instances of different classes by using the decision function $f(\mathbf{x})$, while maximizing the functional margin. For each instance, the output distances to the decision boundaries are then transformed into probability values through a parametric method of logistic regression [17]. For binary classification, the sigmoid function with the parameters A and B is defined as:

$$P_1(\mathbf{x}) = \frac{1}{1 + \exp(Af(\mathbf{x}) + B)} \tag{1}$$

$$P_0(\mathbf{x}) = 1 - P_1(\mathbf{x}) \tag{2}$$

The confidence value for the predicted class is obtained by forming the difference of the posterior probabilities $P_0(\mathbf{x}), P_1(\mathbf{x})$ for classes '0' and '1', respectively.

$$C(\mathbf{x}) = |P_1(\mathbf{x}) - P_0(\mathbf{x})| \tag{3}$$

In our experiments, we consider the MC query strategy [18] that has the potential advantage of avoiding the selection of noisy data, which can be caused by distortions of acoustic patterns [19], unreliable or ambiguous annotations [20] as it is usually the case for acoustic emotion recognition tasks due to their subjective nature. Formally, the query function for MC is defined as:

$$\mathbf{x}_{MC} = \arg\min_{\mathbf{x}} |C(\mathbf{x}) - C_m|, \tag{4}$$

where $C(\mathbf{x})$ denotes the confidence value assigned to the predicted label of a given instance $\mathbf{x}$. The confidence values are ranked and stored in a queue (in descending order). Accordingly, C_m represents the confidence value of the instance located in the centre of the ranking queue. Ideally, for uniformly distributed predictions, C_m would be 0.5. Nonetheless, in practice this value is not fixed. In fact, it varies due to the changes on the unlabelled data pool as learning progresses and labelled instances are accumulatively moved to the training set.

B. Agreement Levels

Given the number n of annotators who are available for labelling a specific database, we define the *agreement level* as the minimum number of raters agreeing on one common category. Accordingly, $j \in \{1, ..., \lfloor \frac{n+1}{2} \rfloor\}$, with $j, n \in \mathbb{N}$, agreement levels can be selected. For the upper limit of the interval, the floor is considered with regard to even numbers of annotators. Specifically, $n' \in \{j, ..., 2j-1\}, n' \in \mathbb{N}$ raters might be needed until a certain agreement level j is achieved. In practice, j raters would be required simultaneously in the first round of queries to minimise the related time-consumption as j is the minimum number of ratings to achieve the respective agreement level. The SAL performance that is achieved through majority voting among all n raters is set to the baseline in our experiments.

C. Rater Reliability and Correlation

In existing crowd-sourcing platforms (e. g. Amazon Mechanical Turk [21]), annotation tasks are distributed to paid clickworkers to complete [22], [23]. For work screening in these annotation systems, the rater reliability is usually assessed and guaranteed through a pretest comprising different questions to determine the annotator is taking his task seriously or just clicking haphazardly. Inspired by the quality management system, we implement a preliminary stage preceding the learning algorithm to assess the rater reliability and the representativeness of every possible rater subset in relation to the respective gold standard labels. For this purpose, we randomly select a test set of labelled instances and train a rater-specific model for each single rater. The obtained classification accuracy is used to rank the raters according to their reliability. Additionally, the correlation between the arithmetic mean of the votes within the rater subsets and the respective gold standard label is computed. By this means, we obtain a measure for the inter-rater agreement and the reliability of a rater group, respectively. Table I depicts an example of the correlation values of the rater subgroups with the highest correlations. As it can be seen, the larger rater groups with the highest correlation values always include the smaller ones, which can be explained by the high coherence between rater reliability and inter-rater correlation. This ranking list is particularly advantageous for the implementation of the adaptive query strategy because the raters are sequentially requested. Depending on the defined agreement level, the subgroup with the minimum number j of raters is selected. If no consensus is reached, the raters who are most representative by forming a group with the preceding raters are enlisted one after another.

In this example, there are five raters available and rater 3 achieves the highest performance in the pretest. Consequently, the most efficient query order considering both rater reliability and correlation is defined as $3, 4, 1, 5, 2$. Besides, it should be noted that once a minimum subset of j raters is selected for an agreement level j, the internal order is not relevant since all raters will be queried in one turn. The principle of the adaptive query strategy is illustrated in Figure 1. The solid line indicates the minimum number of annotations required to achieve the respective agreement level, while the dotted line implies the backup raters who might be enlisted one after another.

TABLE I
CORRELATION BETWEEN THE ARITHMETIC MEAN OF THE RATINGS OF THE MOST RELIABLE RATER SUBGROUPS AND THE GOLD STANDARD LABELS, OBTAINED BY MAJORITY VOTING AMONG ALL FIVE AVAILABLE RATERS.

# Raters	Rater Subgroup	Correlation
2	3 4	0.852
3	3 4 1	0.891
4	3 4 1 5	0.890

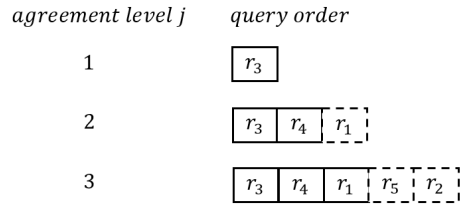


Fig. 1. Adaptive query strategy of the DAL method according to different agreement levels

D. Algorithms

Figure 2 shows the pseudo-code description of the oDAL and uDAL algorithms based on the MC strategy. Let $\mathcal{L} = ([\mathbf{x}_1, y_1], \ldots, [\mathbf{x}_l, y_l]), i = 1, 2, ..., l$, denote a small set of labelled training data, where $\mathbf{x}_i$ is a d-dimensional feature vector, and y_i is the assigned emotion-related label. Additionally, a large pool of unlabelled data $\mathcal{U} = (\mathbf{x}'_1, \ldots, \mathbf{x}'_u), k = 1, 2, ...u$, exists where $u \gg l$ and $\mathbf{x}'_k$ is a d-dimensional feature vector. The number of votes for a specific class label y' that is manually assigned to an example instance $\mathbf{x}' \in \mathcal{N}_a$ is named v'.

In a preliminary stage, a rater-specific model is trained on a test set comprising a number t of randomly selected labelled instances. The classification accuracy of the model is used to determine the reliability of each rater, while the inter-rater agreement corresponds to the correlation between the arithmetic mean of the votes within a subset of raters and the respective gold standard label (Table I). Based on the reliability assessment, a specific query order is defined by ranking the raters and forming the rater subgroups for different agreement levels. When using uDAL algorithm, the initial agreement level is set to $j = 1$. The learning process starts by training a model on the labelled data $\mathcal{L}$ and subsequently using this model to classify all instances of the unlabelled data pool $\mathcal{U}$. According to the MC query strategy, an subset $\mathcal{N}_a \subset \mathcal{U}$ is selected. Step 5) pertains to the proposed adaptive query strategy. Optionally, the uDAL algorithm can be applied at this stage. Starting at $j = 1$, if the confidence value of a selected instance decreases below a prediction uncertainty level p, the agreement level is upgraded to a higher one until the next prediction uncertainty limit is reached. Depending on the selected agreement level, the most reliable rater or rater subgroup is requested to annotate the

selected instances. The stopping criterion for manual labelling of each instance is fulfilled when the respective agreement level is achieved. Finally, the human labelled instances are removed from $\mathcal{U}$ and added to $\mathcal{L}$. The sequential process is repeated until a predefined number of instances are annotated.

Taking again our previous example, assuming there are five expert labellers available for a specific binary classification task, the majority is attained if the same opinion occurs three times. In the existing SAL algorithm, all five labellers would be needed in the query process regardless of the actual distribution of votes. In comparison, choosing $j = 3$, the DAL approach first queries the most reliable group consisting of three raters since this is the minimum number of annotations to obtain the same result as with majority voting among all five raters. If the first three raters agree on one common category, the opinions of the two others are irrelevant, in the sense that they will not affect the final annotation. If this is not the case, the query will be continued by iteratively requesting one more rater according to the ranking until the same opinion occurs three times or there is no rater left. Following the line of thought, the term 'majority' can be regarded as *relative* to the number of the raters actually enlisted, which does not necessarily correspond to the number of all available raters. In this example, $j = 3$ is the highest achievable agreement level as explained before. However, it is also possible to set the stopping condition to the first or second agreement level, requiring only one or two vote(s) for one common category. The respective numbers of potentially required raters would be $n' = \{2, 3\}$ for $j = 2$ and $n' = 1$ for $j = 1$, respectively. The related trade-off between learning performance and cost reduction will be further investigated in Section V-B.

III. Database

In our experiments, we use the FAU Aibo Emotion Corpus (AEC) [24] of the INTERSPEECH 2009 Emotion Challenge (IS09 EC) [25], [26]. The database consists of recordings of children interacting with Sony's pet robot Aibo, which performs a fixed, predetermined sequence of actions. Spontaneous German speech that is emotionally coloured is provoked by leading the children to believe that the Aibo was responding to their commands, whereas the robot was actually controlled by a human operator and sometimes behaved disobediently. The recordings were taken from 51 children (age 10-13, 21 male, 30 female, about 9.2 hours of speech without pauses) at two different schools, referred to as 'MONT' and 'OHM'. Five labellers (advanced students of linguistics) annotated each word independently from each other as neutral (default) or as one of ten emotional states: *angry*, *touchy*, *reprimanding*, *emphatic*, *surprise*, *joyful*, *helpless*, *motherese*, *bored*, and *rest*. We use the same natural speech corpus as in the IS09 EC that comprises 18 216 instances. Each instance corresponds to a manually defined chunk that consists of multiple words according to the syntactic-prosodic criteria. For binary classification, the 11-class labels are mapped onto two-class labels by defining states with negative valence (*angry*, *touchy*, *reprimanding*, *emphatic*) as **NEG**(agative), and all other states as **IDL**(e).

Algorithms: *Order-based Dynamic Active Learning (oDAL)*

Pretest:

1) Select t random test instances
2) Train a model for each single rater
3) Compute the correlation value for each rater subgroup
4) Define the query order based on rater reliability and inter-rater agreement

Repeat:

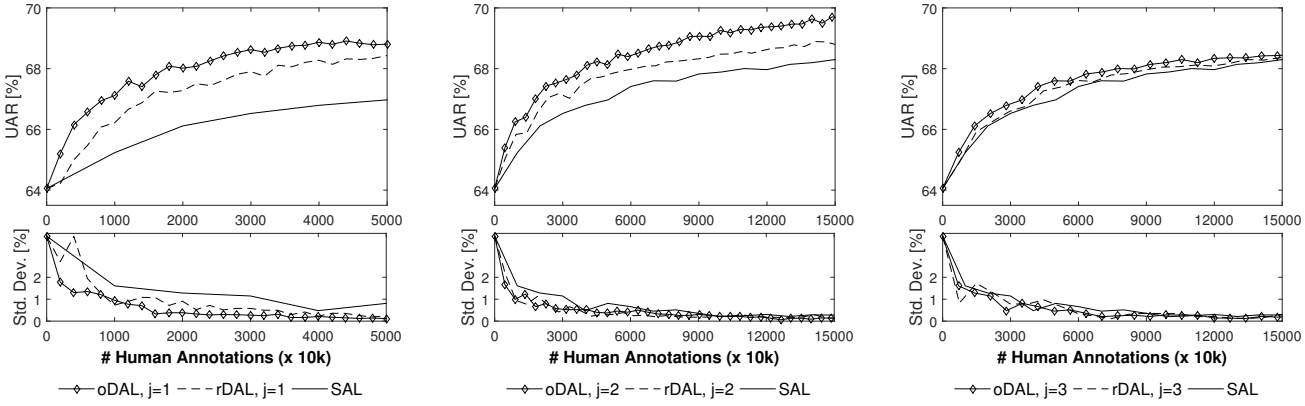
1) (Optional) Upsample the training set $\mathcal{L}$ to obtain even class distribution $\mathcal{L}_D$
2) Use $\mathcal{L}/\mathcal{L}_D$ to train a classifier $\mathcal{H}$, and then classify the unlabelled data set $\mathcal{U}$
3) Rank the data based on the prediction confidence values C and store them in a queue
4) Select a subset $\mathcal{N}_a$ with medium certainty
5) **For** each instance $\mathbf{x}'$ in $\mathcal{N}_a$
   a) Optional: *Upgraded Dynamic Active Learning (uDAL)*
      If $C > p$; $j = j$;
      else $j++$;
   b) Submit $\mathbf{x}'$ to the first j raters
   c) If $v' = j$; **STOP**
      else **repeat**: select one rater for annotation
      until agreement level j is achieved
   d) Assign y' to $\mathbf{x}'$
6) Remove $\mathcal{N}_a$ from the unlabelled set $\mathcal{U}$, $\mathcal{U} = \mathcal{U} \setminus \mathcal{N}_a$
7) Add $\mathcal{N}_a$ to the labelled set $\mathcal{L}$, $\mathcal{L} = \mathcal{L} \cup \mathcal{N}_a$

Fig. 2. Pseudocode description of the oDAL and uDAL algorithms based on the medium certainty query strategy.

A heuristic approach is applied to map the labels from the word-level to the chunk-level for each of the five labellers, where a chunk is defined as NEG if it contains at least one word with negative valence. To define the gold standard for the baseline results, we resort to majority voting to combine the labels from all five labellers to one single label for each chunk. The frequencies for the two-class problem are given in Table II. Speaker independence is guaranteed by using the speech samples of the school 'OHM' for training and the data of the other school 'MONT' for validation. Specifically, the training data referred to as 'Pool' contains both the labelled training set $\mathcal{L}$ and the unlabelled data pool $\mathcal{U}$.

TABLE II
DISTRIBUTION OF SPEAKERS AND INSTANCES PER PARTITION OF THE FAU AEC. M: MALE; F: FEMALE; NEG: NEGATIVE EMOTIONS; IDL: NEUTRAL AND POSITIVE EMOTIONS.

	# speakers		# instances per class		
FAU AEC	M	F	NEG	IDL	Σ
Pool	13	13	3 358	6 601	9 959
Validation	8	17	2 465	5 792	8 257
Σ	21	30	5 823	12 393	18 216



(a) (b) (c)

Fig. 3. Comparison between Random Dynamic Active Learning (rDAL) and Order-based Dynamic Active Learning (oDAL): the performance measures show the UAR values averaged across 20 runs of the algorithm and the respective standard deviations vs the number of human annotations for the FAU AEC with IS09 EC feature set by 200 initial training instances for agreement level a) $j = 1$, b) $j = 2$, and c) $j = 3$

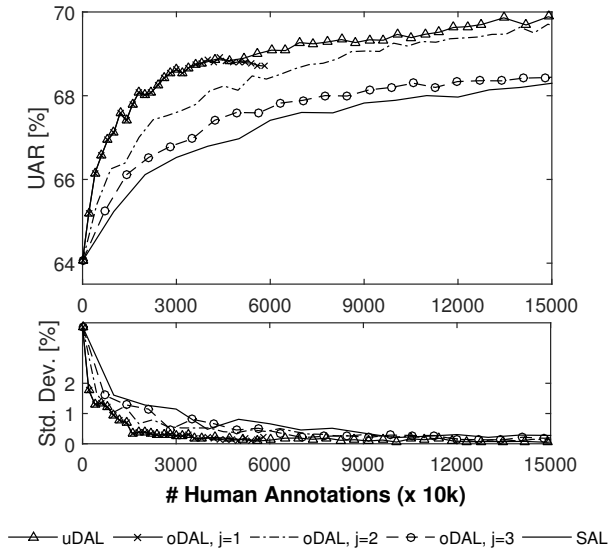


Fig. 4. Order-based Dynamic Active Learning(oDAL) vs Upgraded Dynamic Active Learning (uDAL): the performance measures show the UAR values averaged across 20 runs of the algorithm vs the number of human annotations for the FAU AEC with IS09 EC feature set by 200 initial training instances.

IV. Acoustic Features

The acoustic features used in our experiments are adopted from the baseline feature set of IS09 EC. This is created with the openSMILE framework [27], [28] by applying statistical functionals to frame-wise low-level-descriptors (LLDs) as depicted in Table III. To each of the 16 LLDs, the delta coefficients are computed. Finally, the 12 functionals are applied on a per-chunk level. As result of the 'brute-forcing' method, the total feature vector per chunk contains $16 \times 2 \times 12 = 384$ attributes.

TABLE III
The IS09 EC Acoustic feature set: low-level descriptors (LLDs) and respective functionals.

LLD (Δ)	**Functionals**
ZCR	mean
RMS Energy	standard deviation energy
F0	kurtosis, skewness
HNR	extremes: value, rel. position, range
MFCC 1-12*	linear regression: offset, slope, MSE

V. Empirical Evaluation

In this section, we investigate the performance of the various DAL algorithms by evaluating the classification accuracy in relation to the number of human annotations. Specifically, the rDAL and oDAL algorithms are confronted with regard to different agreement levels. Furthermore, the uDAL algorithm which combines the first and second agreement levels of the oDAL method is evaluated. Additionally, the robustness of the trained model is examined. All results are compared with the baseline performance achieved through the conventional SAL method.

A. Experimental Setup

For transparency and reproducibility, we used open-source classifier implementations of SVMs from the WEKA data mining toolkit [29]. As classifiers, we chose linear kernel SVMs trained with a complexity parameter C constant of 0.05 and with Sequential Minimal Optimization (SMO), as they are robust against over-fitting in high dimensional feature spaces. For initial training of the model, 200 instances were randomly selected from the training data, whereas the remaining instances were used as the unlabelled data pool. At each learning iteration, we selected a subset $\mathcal{N}_a$ comprising 200 instances to be submitted to manual annotation. The learning process stopped after a predefined number of iterations is reached. The training

process was repeated in 20 independent runs. As the evaluation measure, we considered the unweighted average recall (UAR) in accordance with the previous IS challenges.

B. Discussion of Results

In Figure 3, we show the UAR measures across 20 independent runs of the learning process and the respective standard deviations by the use of the rDAL and oDAL algorithms. According to the characteristic curve progression of AL, the sequential addition of human-labelled instances to the initial training set (200 per iteration) leads to continuous improvements in the performance of the classifier. The UAR first increases steeply with the number of total human annotations before reaching a plateau. It can be clearly seen that higher classification accuracy can be achieved through the DAL methods with the same annotation effort as in SAL. For a more detailed analysis of the various algorithms, we computed Student's t-test to statistically compare the performances. For $j = 1$, we examined the interval between $1\,000$ and $5\,000$ annotations. As for $j = 2, 3$, the UAR performance measures are compared between $7\,000$ and $15\,000$ annotations. The analysis of the significance levels ($p < .0001$) confirms our previous observation and indicates that the DAL approaches generally lead to significantly better performance than SAL. This is particularly evident for oDAL that led to the best performance for all three agreement levels by consistently and robustly outperforming the other methods. Further, it is worth noting that the highest UAR is obtained at agreement level $j = 2$, followed by $j = 1$ and $j = 3$, which can be explained by the different rater reliability levels. Above all, Figure 4 demonstrates that the highest efficiency is realised through the uDAL method, which starts at agreement level $j = 1$ before jumping to $j = 2$ after reaching $4\,000$ annotated instances and finishing 20 iterations, respectively. The transition point can be noted by the slight click in the learning curve.

In order to substantiate our findings, we additionally compare the relative cost reduction by measuring the number of human annotations at $UAR = 68.2\,\%$, which is the top performance achieved with the SAL method. According to Table IV, the relative cost reduction (CR) increases with lower agreement levels regarding all applied algorithms. This can be explained by the fact that all five labellers can be considered relatively reliable. Consequently, selecting lower agreement levels results in a dramatic cost reduction without affecting much the overall performance. Furthermore, as expected, the rDAL algorithm requires generally more human annotations than oDAL, resulting in lower CR. Moreover, it is important to mention that the average number of annotators per selected instance (AA) inclines to the minimum that is necessary to achieve a certain agreement level (Section II-B). This finding suggests that the maximum number of raters is not required in most annotation cases and reinforces the inefficiency of majority voting. Finally, the analysis of standard deviation shows that the stability of the model is enhanced during the learning process.

TABLE IV
COST CORRESPONDING TO THE NUMBER OF HUMAN ANNOTATIONS AT UAR = $68.2\,\%$ AND THE RELATIVE COST REDUCTION (CR) BY COMPARING THE AGREEMENT LEVELS $j = 1, 2, 3$ OF THE oDAL ALGORITHM AND THE uDAL PERFORMANCE WITH THE SAL BASELINE. THE AVERAGE NUMBER OF ANNOTATORS PER SELECTED INSTANCE (AA) IS ALSO PROVIDED.

	cost (x 10k)	CR (%)	AA
SAL	1.6	–	5
rDAL j=3	1.27	20.53	3.72
rDAL j=2	0.72	54.69	2.35
rDAL j=1	0.38	76.33	1
oDAL j=3	0.99	38.34	3.52
oDAL j=2	0.42	73.73	2.25
oDAL j=1	0.23	85.41	1
uDAL	0.23	85.41	1.96

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed two novel approaches for Dynamic Active Learning that further reduce the amount of the costly human labelling work in comparison with our previous work [16] that introduced an adaptive annotation strategy with random query order. In the enhanced DAL derivations, we additionally consider the reliability of each single rater and of every possible rater subgroup in order to identify the most efficient query order. Moreover, a highly dynamic approach is proposed that upgrades the agreement level to handle noisy data on approaching the end of the learning process. Our results demonstrate that the novel features of the DAL method lead to improvement of the original DAL algorithm for all tested agreement levels, requiring up to $85.41\,\%$ less human annotations while obtaining the same performance. Or to put it the other way round, higher classification accuracy can be achieved with the same annotation effort.

The implementation of the preliminary stage to access the rater reliability and inter-rater correlation is also compatible with the currently emerging and popular crowd-sourcing systems. In this way, we combine enhanced machine learning methods with highly efficient data annotation resources, reaching a new milestone for highly efficient exploitation of unlabelled data.

For future research, we will explore the link between the agreement level, the quality of annotators, and the subjectivity of the annotations. Besides, the reliability and correlation values will be updated after each annotated instance in order to enable more dynamic DAL algorithms. Finally, the robustness of the DAL methods will be investigated by conducting experiments on multiple corpora, different feature sets, and varying amount of initial training instances.

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Unions Framework Programme for Research and Innovation HORIZON 2020 under the Grant No. 645378 (ARIA-VALUSPA) and the European Unions Seventh Framework Programme under the ERC Starting Grant No. 338164 (iHEARu).

REFERENCES

[1] B. Schuller and A. Batliner, *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. New York, NY: John Wiley & Sons, 2014.

[2] S. Steidl, M. Levit, A. Batliner, E. Nöth, and H. Niemann, ""of all things the measure is man": Automatic classification of emotions and inter-labeler consistency," in *Proc. of ICASSP*, Philadelphia, PA, 2005, pp. 317–320.

[3] B. Schuller, "Multimodal Affect Databases - Collection, Challenges & Chances," in *Handbook of Affective Computing*, ser. Oxford Library of Psychology, R. A. Calvo, S. DMello, J. Gratch, and A. Kappas, Eds. Oxford University Press, 2015, ch. 23, pp. 323–333, invited contribution.

[4] A. E. Thompson and D. Voyer, "Sex differences in the ability to recognise non-verbal displays of emotion: A meta-analysis," *Cognition and Emotion*, vol. 28, no. 7, pp. 1164–1195, 2014.

[5] C. L. Rusting, "Personality, mood, and cognitive processing of emotional information: three conceptual frameworks." *Psychological bulletin*, vol. 124, no. 2, p. 165, 1998.

[6] C. Edgar, M. McRorie, and I. Sneddon, "Emotional intelligence, personality and the decoding of non-verbal expressions of emotion," *Personality and Individual Differences*, vol. 52, no. 3, pp. 295–300, 2012.

[7] T. Ruffman, J. D. Henry, V. Livingstone, and L. H. Phillips, "A meta-analytic review of emotion recognition and aging: Implications for neuropsychological models of aging," *Neuroscience & Biobehavioral Reviews*, vol. 32, no. 4, pp. 863–881, 2008.

[8] X. Zhu, "Semi-supervised learning literature survey," Department of Computer Sciences, University of Wisconsin at Madison, Madison, WI, Tech. Rep. TR 1530, 2006.

[9] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. of 11th annual conference on Computational Learning Theory*, Madison, WI, 1998, pp. 92–100.

[10] B. Settles, "Active learning literature survey," Department of Computer Sciences, University of Wisconsin–Madison, Wisconsin, WI, Tech. Rep., 2009.

[11] M. Li and I. Sethi, "Confidence-based active learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 8, pp. 1251–1261, 2006.

[12] R. Liere, "Active learning with committees: An approach to efficient learning in text categorization using linear threshold algorithms," Ph.D. dissertation, Oregon State University, OR, Portland, OR, 2000.

[13] G. Tur, D. Hakkani-Tür, and R. E. Schapire, "Combining active and semi-supervised learning for spoken language understanding," *Speech Communication*, vol. 45, no. 2, pp. 171–186, 2005.

[14] X. Zhu, J. Lafferty, and Z. Ghahramani, "Combining active learning and semi-supervised learning using Gaussian fields and harmonic functions," in *ICML 2003 workshop on the continuum from labeled to unlabeled data in machine learning and data mining*, Washington DC, 2003, pp. 58–65.

[15] Z. Zhang, E. Coutinho, J. Deng, and B. Schuller, "Cooperative learning and its application to emotion recognition from speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 115–126, 2014.

[16] Y. Zhang, E. Coutinho, Z. Zhang, C. Quan, and B. Schuller, "Agreement-based Dynamic Active Learning with Least and Medium Certainty Query Strategies," in *Proc. of Advances in Active Learning : Bridging Theory and Practice Workshop, ICML 2015*, Lille, France, 2015, 5 pages.

[17] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in large margin classifiers*, A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, Eds. Cambridge, MA: MIT Press, 1999, pp. 61–74.

[18] Z. Zhang and B. Schuller, "Active learning by sparse instance tracking and classifier confidence in acoustic emotion recognition," in *Proc. of INTERSPEECH*, Portland, OR, 2012, 4 pages.

[19] M. You, C. Chen, J. Bu, J. Liu, and J. Tao, "Emotion recognition from noisy speech," in *Proc. of ICME*. Toronto, Canada: IEEE, 2006, pp. 1653–1656.

[20] M. Grimm and K. Kroschel, "Evaluation of natural emotions using self assessment manikins," in *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding*, Cancun, Mexico, 2005, pp. 381–385.

[21] A. Kittur, E. H. Chi, and B. Suh, "Crowdsourcing user studies with mechanical turk," in *Proc. of the SIGCHI conference on human factors in computing systems*, Florence, Italy, 2008, pp. 453–456.

[22] J. Howe, "The rise of crowdsourcing," *Wired magazine*, vol. 14, no. 6, pp. 1–4, 2006.

[23] M.-C. Yuen, I. King, and K.-S. Leung, "A survey of crowdsourcing systems," in *Proc. of Privacy, Security, Risk and Trust (PASSAT) and Proc. of Social Computing (SocialCom)*. Boston, MA: IEEE, 2011, pp. 766–773.

[24] S. Steidl, *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech*. Berlin: Logos Verlag, 2009.

[25] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Proc. of INTERSPEECH*, Brighton, UK, 2009, pp. 312–315.

[26] B. Schuller, "The computational paralinguistics challenge," *IEEE Signal Processing Magazine*, vol. 29, no. 4, pp. 97–101, 2012.

[27] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE – the Munich versatile and fast open-source audio feature extractor," in *Proc. of ACM MM*, Florence, Italy, 2010, pp. 1459–1462.

[28] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proc. of ACM MM*, Barcelona, Spain, 2013, pp. 835–838.

[29] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.