Annotator trustability-based cooperative learning solutions for intelligent audio analysis

Simone Hantke, Christoph Stemp, Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Hantke, Simone, Christoph Stemp, and Björn Schuller. 2018. "Annotator trustability-based cooperative learning solutions for intelligent audio analysis." In *Interspeech 2018, September 2-6, Hyderabad, India: Speech Research for Emerging Markets in Multilingual Societies*, edited by B. Yegnanarayana, C. Chandra Sekhar, Shrikanth Narayanan, S. Umesh, S. R. M. Prasanna, Hema A. Murthy, Preeti Rao, Paavo Alku, and Kumar Ghosh, 3504–8. Baixas: ISCA. https://doi.org/10.21437/interspeech.2018-1019.

Interspeech 2018
2-6 September 2018, Hyderabad

Annotator Trustability-based Cooperative Learning Solutions for Intelligent Audio Analysis

Simone Hantke[1,2]*, Christoph Stemp*[1,3]*, and Björn Schuller*[1,4]

[1]Chair for Embedded Intelligence on Health Care and Wellbeing, University of Augsburg, Germany
[2]Machine Intelligence & Signal Processing Group, Technische Universität München, Germany
[3]Chair of Complex & Intelligent Systems, University of Passau, Germany
[4]GLAM – Group on Language, Audio & Music, Imperial College London, UK

simone.hantke@informatik.uni-augsburg.de

Abstract

A broad range of artificially intelligent applications are nowadays available resulting in a need for masses of labelled data for the underlying machine learning models. This annotated data, however, is scarce and expensive to obtain from expert-like annotators. Crowdsourcing has been shown as a viable alternative, but it has to be carried out with adequate quality control to obtain reliable labels. Whilst crowdsourcing allows for the rapid collection of large-scale annotations, another technique called Cooperative Learning, aims at reducing the overall annotation costs, by learning to select only the most important instances for manual annotation. In this regard, we investigate the advantages of this approach and combine crowdsourcing with different iterative cooperative learning paradigms for audio data annotation, incorporating an annotator trustability score to reduce the labelling effort needed and, at the same time, to achieve better classification results. Key experimental results on an emotion recognition task show a considerable relative annotation reduction compared to a 'non-intelligent' approach of up to 85.3 %. Moreover, the proposed trustability-based methods reach an unweighted average recall of 74.8 %, while the baseline approach peaks at 61.2 %. Therefore, the proposed trustability-based approaches efficiently reduce the manual annotation load, as well as improving the model.

Index Terms: Audio Processing, Crowdsourcing, Cooperative Learning, Machine Learning, Annotator Trustability

1. Introduction

Machine learning algorithms came a long way in recent years and are applied in a wide variety of fields, such as cancer prognosis and prediction in medicine [1], image processing for autonomous driving [2], stock market analysis for financial trading [3], or sentiment analysis and emotion recognition in computational paralinguistics [4]. All algorithms in these different fields have in common that they rely on labelled data to train their machine learning models. Although raw data is readily available through social media or the Internet-of-Things-enabled devices, such data mostly comes without any labels. Conventionally, experts in the corresponding field were hired to perform the labelling process, which is time-consuming and can get quite expensive [5].

To help alleviate this labelled data shortage, crowdsourcing has emerged as a new trend to outsource these tedious and costly annotation tasks to the masses for large-scale parallel labelling [6, 7]. However, crowdsourcing comes with a large amount of anonymous annotators, who mostly do not have much knowledge about the specific task, potentially resulting in lower quality of annotations [8, 9]. As a consequence – to still obtain reliable annotations – the quality of the labels has to be assured by developing valid quality control mechanisms.

Furthermore, even with the help of crowdsourcing, a large amount of annotation effort is still required and at least as many annotations as there are unlabelled data instances need to be collected. Therefore, different machine learning algorithms have been developed for exploitation of unlabelled (speech) data [4, 10] aiming to reduce the number of data instances which need manual labelling in the first place [11, 12].

1.1. Related Work

The idea of crowdsourcing was already applied in several tasks, such as relevance evaluation for information retrieval systems [13], for natural language processing [14], for labelling emotional speech assets [15], or for sentiment analysis [16]. As this recruitment of anonymous annotators has to be done with an adequate quality control to obtain reliable labels, different methods were developed, including test questions, automatic filtering, and carefully designed tasks or its descriptions [9]. In this context, a novel method of evaluating the collected labels was introduced, making use of the annotators' trustability score [17, 4].

On the other hand, the combination of crowdsourcing and Active Learning (AL) algorithms has already been successfully used to build an automatic translation system for low-resource language pairs [18], for entity recognition and sentiment analysis [19], or diverse other research fields [12, 19]. Due to the promising results of AL, considerable research has been done on this topic [11, 12, 19]. Then, Dynamic Active Learning was introduced to further reduce the number of manual annotations by using an adaptive query strategy that decides dynamically on an instance basis how many manual annotations are needed [20], followed by the combination of AL with semi-supervised learning resulting in a Cooperative Learning algorithm [21].

1.2. Contributions of this Work

In this context, we have previously shown the success of combining our developed crowdsourcing-based annotator trustability measurements with two basic active learning query strategies in order to exploit the advantages of AL and most importantly to tackle the problem of unreliable annotations [4]. Based on this initial work, we herein expand our AL algorithms by introducing novel *Trustability-based Cooperative Learning* algorithms, which are cooperative learning algorithms including the calculated annotators' trustability values. Within this contribution, these novel trustability-based algorithms are exemplary integrated into our gamified intelligent crowdsourcing platform iHEARu-PLAY [22, 23] to perform several emotion recognition experiments. We aim at identifying techniques which combine

10.21437/Interspeech.2018-1019

Algorithm 1: Trustability-based Dynamic Active Learning with least and medium certainty query strategy.

Input: Pre-labelled training set $\mathcal{T}$.
repeat
  Upsample the training set $\mathcal{T}$ to obtain even class distribution $\mathcal{T}_D$.
  Use $\mathcal{T}/\mathcal{T}_D$ to train enhancing classifier $\langle$, then classify pool set $\mathcal{P}$.
  Rank data based on the prediction confidence values C and store them in queue Q.
  Choose a query strategy:
    – Least certainty query strategy: Select subset $\mathcal{N}_p$ whose elements are 'at the bottom' of ranking queue Q.
    – Medium certainty query strategy: Select subset $\mathcal{N}_p$ whose elements are 'in the middle' of ranking queue Q.
  Submit selected instances $\mathcal{N}_p$ to manual annotation.
  repeat
    Compute aggregated manual labels l and assess annotators' trustability values T_u using the trustability score calculation (cf. [4]).
    Add instances with high annotators' trustability scores $N_{p(high)}$ to labelled dataset $\mathcal{T}$, $\mathcal{T} = \mathcal{T} \cup N_{p(high)}$.
    Remove $N_{p(high)}$ from unlabelled set P, $P = P - N_{p(high)}$.
    Keep instances with low $N_{p(low)}$ or medium $N_{p(medium)}$ annotators' confidence for the next iteration in order to obtain more manual labels.
    Add high confidence instances $N_{p(high)}$ and their aggregated labels to training set $\mathcal{T}$, $\mathcal{T} = \mathcal{T} \cup N_{p(high)}$.
  until *until a stop criteria is fulfilled.*
until *a predefined number of iterations is met.*
Obtain the final classifier $\mathcal{H}$.

the advantages of trustability-based crowdsourcing and cooperative learning to efficiently reduce the number of needed annotations by following valid quality control procedures.

2. Intelligent Audio Analysis

The herein proposed algorithms aim at collecting only reliable labels, while at the same time preventing the acquisition of unnecessary annotations. Therefore, the *Trustability-based Dynamic Active Learning Algorithm* (cf. [4]) and the novel *Trustability-based Cooperative Learning Algorithms* include an annotator trustability-based agreement level j. This agreement level is computed by using the trustability value T_u of an annotator (cf. [4]) by determining how many annotations have to be gathered in order to compute the final label of an instance. The advantage of the agreement level j in this approach is that it is repeated until the defined annotator trustability sum of the answers reaches the agreement level instead of stopping when n annotations for one label have been collected.

For the learning approaches presented in this work, a small set of already labelled data is required, which can be obtained from experts or by transfer learning [24].

2.1. Trustability-based Dynamic Active Learning

Giving the highly promising results obtained in our earlier work [4], we consider two basic Dynamic Active Learning (DAL) algorithms with two different certainty query strategies (cf. Algorithm 1). Starting with an upsampling procedure, an equal class distribution $\mathcal{T}_D$ in the training set $\mathcal{T}$ can be obtained. Both algorithms start by classifying all instances of the unlabelled data pool $\mathcal{P}$ using the model previously trained on the labelled data. Then, the confidence values C (cf.[25]) assigned to each instance are ranked and stored in a queue Q. Afterwards, a subset N_p of $\mathcal{P}$ corresponding to those instances predicted with least and medium confidence values are sent for manual annotation. Finally, these instances are added to the training set $\mathcal{T}$ and are removed from the unlabelled data set $\mathcal{P}$. This sequential process is repeated until a predefined number of instances are selected or until some stopping criterion is met [4, 21].

Algorithm 2: Trustability-based Single-view Cooperative Learning.

Input: Pre-labelled training set $\mathcal{T}$.
repeat
  Perform **Active Learning** (cf. [4]) on an initial training set $\mathcal{T}$ and obtain subset $\mathcal{N}_p$ for manual labelling.
  For each instance x in $\mathcal{N}_p$:
  repeat
    Submit x to all annotators.
  until *trustability-based agreement level* $\sum_{a_u \in \mathcal{A}_{xl}} (T_u + a_t) \geq j$ *is fulfilled.*
  Remove manual labelled subset $\mathcal{N}_p$ from unlabelled set $\mathcal{P}$, $\mathcal{P} = \mathcal{P} \setminus \mathcal{N}_p$ and add $\mathcal{N}_p$ to labelled training set $\mathcal{T}$, $\mathcal{T} = \mathcal{T} \cup \mathcal{N}_p$.
  Execute **Self-Training** (cf. [21]) based on the new training set and obtain subset $\mathcal{N}_{st}$ for automatic labelling.
  Remove automatic labelled subset $\mathcal{N}_{st}$ from unlabelled $\mathcal{P}$, $\mathcal{P} = \mathcal{P} \setminus \mathcal{N}_{st}$ and add $\mathcal{N}_{st}$ to labelled training set $\mathcal{T}$, $\mathcal{T} = \mathcal{T} \cup \mathcal{N}_{st}$.
until *a predefined number of iterations is met.*

2.2. Trustability-based Cooperative Learning

Cooperative learning combines AL and semi-supervised learning (SSL), using previously labelled data to find corresponding labels for the still unlabelled data in an iterative process to reduce the potential downsides of the individual algorithms [21]. We differentiate between three cooperative learning scenarios:

(i) Trustability-based single-view Cooperative Learning: The single-view cooperative learning (svCL), shown in Algorithm 2, is a combination of AL with a medium certainty query strategy and Self-Training. Self-Training (cf. [21]) is based on the principle of highest certainty or agreement, in such a way that the predicted classes with higher certainty levels are automatically labelled and added to the training set $\mathcal{T}$. First, the AL algorithm executes the initial training set and obtains a subset of files N_p with medium certainty to hand over to manual annotation. Then, the labelled files $\mathcal{N}_p$ are removed from the unlabelled pool set $\mathcal{P}$ and are added to the training set $\mathcal{T}$. With this newly updated training set, the Self-Training algorithm is executed to generate labels for a subset of files $\mathcal{N}_{st}$ with the highest confidence values C. When the labels are selected, the associated files $\mathcal{N}_{st}$ are removed from the unlabelled pool set $\mathcal{P}$ and added to the training set $\mathcal{T}$. This algorithm is repeated until all files, or a predefined number of files are annotated, or the trained model reaches a certain performance threshold.

(ii)Trustability-based mixed-view Cooperative Learning: Mixed-view cooperative learning (xvCL) is the combination of AL with a medium certainty query strategy and Co-Training (cf. [21]), hence the name "mixed-view" (cf. Algorithm 3). In comparison with the earlier described Self-Training, Co-Training uses two models trained and tested on two different "views" of the data. In each iteration of the algorithm, the two "views" select the instances independently. The rest of the algorithm structure is similar to the earlier described svCL algorithm.

(iii) Trustability-based multi-view Cooperative Learning: The multi-view cooperative learning (mvCL) approach combines Co-active learning (cf. [21]) as the AL part and Co-Training as the SSL part (cf. Algorithm 4). Otherwise, the algorithm follows the same structure as the previously described xvCL and svCL algorithms.

Algorithm 3: Trustability-based Mixed-view Cooperative Learning.

Input: A learning domain with features $\mathcal{X}$.
repeat
  Perform **Active Learning** (cf. [4]) on an initial training set $\mathcal{T}$ and obtain subset $\mathcal{N}_p$ for manual labelling.
  For each instance x in $\mathcal{N}_p$:
  repeat
    Submit x to all annotators.
  until *trustability-based agreement level* $\sum_{a_u \in \mathcal{A}_{xl}} (T_u + a_t) \geq j$ *is fulfilled.*
  Remove manual labelled subset $\mathcal{N}_p$ from unlabelled set $\mathcal{P}, \mathcal{P} = \mathcal{P} \setminus \mathcal{N}_p$ and add $\mathcal{N}_p$ to labelled training set $\mathcal{T}, \mathcal{T} = \mathcal{T} \cup \mathcal{N}_p$.
  Execute **Co-Training** (cf. [21]) based on the new training set and obtain subset $\mathcal{N}_{ct}$ for automatic labelling.
  Remove automatic labelled subset $\mathcal{N}_{ct}$ from unlabelled $\mathcal{P}, \mathcal{P} = \mathcal{P} \setminus \mathcal{N}_{ct}$ and add $\mathcal{N}_{ct}$ to labelled training set $\mathcal{T}, \mathcal{T} = \mathcal{T} \cup \mathcal{N}_{ct}$.
until *a predefined number of iterations is met.*

Algorithm 4: Trustability-based Multi-view Cooperative Learning algorithm.

Input: A learning domain with features $\mathcal{X}$.
repeat
  Perform **Co-Active Learning** (cf. [21]) on an initial training set $\mathcal{T}$ and obtain subset $\mathcal{N}_{ca}$ for manual labelling.
  For each instance x in $\mathcal{N}_{ca}$:
  repeat
    Submit x to all annotators.
  until *trustability-based agreement level* $\sum_{a_u \in \mathcal{A}_{xl}} (T_u + a_t) \geq j$ *is fulfilled.*
  Remove manual labelled subset $\mathcal{N}_{ca}$ from unlabelled set $\mathcal{P}, \mathcal{P} = \mathcal{P} \setminus \mathcal{N}_{ca}$ and add $\mathcal{N}_{ca}$ to labelled training set $\mathcal{T}, \mathcal{T} = \mathcal{T} \cup \mathcal{N}_{ca}$.
  Execute **Co-Training** (cf. [21]) based on the new training set and obtain subset $\mathcal{N}_{ct}$ for automatic labelling.
  Remove automatic labelled subset $\mathcal{N}_{ct}$ from unlabelled set $\mathcal{P}, \mathcal{P} = \mathcal{P} \setminus \mathcal{N}_{ct}$ and add $\mathcal{N}_{ct}$ to labelled training set $\mathcal{T}, \mathcal{T} = \mathcal{T} \cup \mathcal{N}_{ct}$.
until *a predefined number of iterations is met.*

3. Experiments

3.1. Tasks

The proposed algorithms were evaluated by conducing several emotion recognition experiments comparing the performance of svCL with the least and medium query strategy, xvCL with the least and medium strategy, and mvCL, all repeatedly without and with trustability-based annotations.

3.2. Dataset

We evaluated on the *FAU Aibo Emotion Corpus* [26], which has been part of the INTERSPEECH emotion challenge 2009 [27] and consists of recordings of German children playing with Sony's pet-robot Aibo. The audio data were recorded at two different schools in Germany, featuring 51 children at the age of 10 to 13 and cut into over 15 000 instances for our purposes, resulting in nearly 7.5 hours of continuous speech. To assure speaker-independence, the instances from one school were used as pool and training sets, whereas the recordings from the second school were used as test set.

3.3. Acoustic Feature Set

We applied the INTERSPEECH 2009 feature set (IS09) [27], which has been applied successfully for different kinds of emotion tasks [28, 29] and extracted our audio features using the open-source toolkit openSMILE [30].

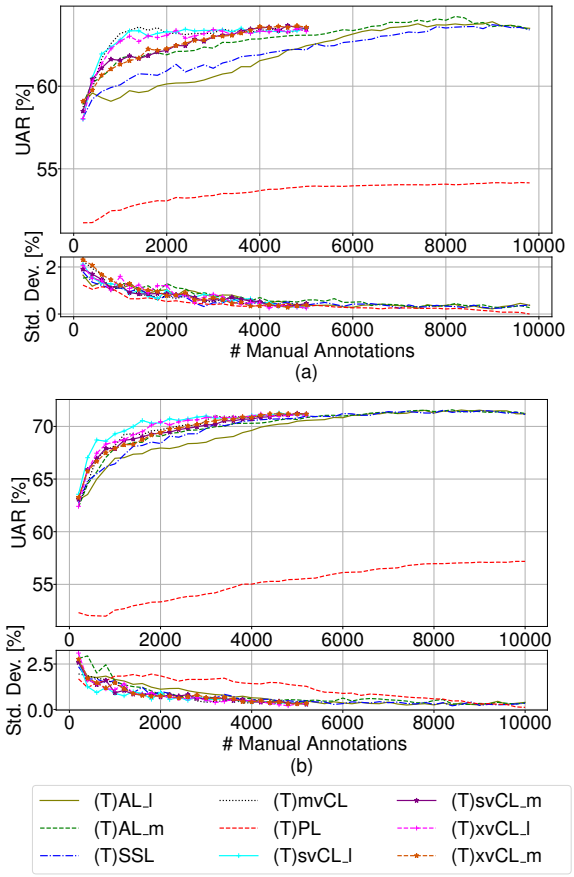

Figure 1: *Comparison of the (trustability-based) learning algorithms and the Passive Learning algorithm. Average UAR and number of manual annotations are measured across 20 independent runs of each algorithm. Results of the algorithms without trustability score are shown in (a), with trustability score in (b).*

3.4. Experimental Setup

We applied the open-source machine-learning and data-mining software WEKA [31] and chose a Support-Vector Machine (SVM) being trained with a Sequential Minimal Optimization (SMO) algorithm and a complexity constant C of 0.1 to construct a hyperplane to separate the instances of different classes.

Our model was initially trained with 200 randomly selected instances, while the remaining data was used as the unlabelled data pool $\mathcal{P}$. At each iteration, the algorithms choose 200 instances for manual annotation from the pool set. We made use of upsampling [32], wherein multiple copies of existing instances were added to the classes which have a low amount of instances. A SSL step was performed in each iteration, getting 200 files from the pool set labelled by the machine. In addition, to ensure statistical accuracy, every algorithm was repeated 20 times with randomly selected initial training instances.

3.5. Evaluation

Annotations. The FAU Aibo Emotion Corpus was labelled on iHEARu-PLAY [22] with the help of the proposed algorithms taking into account the trustability of the annotator. All in all, we had 14 annotators (3 female and 11 male) between 20 and 27 years old, excluding three annotators who did not reveal their

Table 1: *Numeric results of 20 independent runs with 200 randomly selected initial training instances for the approach with conventional and trustability-based annotations. UAR[%] denotes the maximum achieved UAR of the algorithms and their standard deviations SD [%]. NA is the number of annotations needed to achieve the maximal UAR. CR [%] is the relative cost reduction when achieving the maximum compared to Passive Learning; TR [h] is the duration of the instances for which the annotation costs can be saved.*

Conventional						**Trustability-based**					
Algorithm	UAR [%]	SD [%]	NA [#]	CR [%]	TR [h]	Algorithm	UAR [%]	SD [%]	NA [#]	CR [%]	TR [h]
PL	54.17	0.00	15 000	–	–	TPL	57.18	0.11	15 000	–	–
DAL_l	**63.9**	0.37	9 000	40.0	3.0	$TDAL_l$	**71.56**	0.39	8 600	42.6	3.2
DAL_m	63.52	0.27	8 200	45.3	3.4	$TDAL_m$	71.58	0.35	8 400	44.0	3.3
SSL	63.72	0.38	9 000	40.0	3.0	TSSL	71.53	0.35	8 200	45.3	3.4
$svCL_l$	63.51	0.29	5 000	66.7	5.0	$TsvCL_l$	71.28	0.31	4 600	69.3	5.2
$svCL_m$	63.68	0.45	4 600	69.3	5.2	$TsvCL_m$	71.24	0.28	5 000	66.7	5.0
$xvCL_l$	63.48	0.25	4 800	68.0	5.1	$TxvCL_l$	71.17	0.33	**4 200**	72.0	5.4
$xvCL_m$	63.66	0.38	4 800	68.0	5.1	$TxvCL_m$	71.2	0.33	4 800	68.0	5.1
mvCL	63.54	0.29	**4 400**	70.6	5.3	TmvCL	71.23	0.39	5 200	65.3	4.9

age. The annotators were asked to label the data into the emotions *motherese*, *touchy*, *surprised*, *neutral*, *joyful*, *emphatic*, *angry*, *helpless*, *bored*, and *other* as was originally performed on the data, following IS2009 [27]. We adapted the proposed binary emotion classes in [27]; NEG(ative) which includes all negative emotions (angry, touchy, and emphatic), and IDL(e), containing all other emotions.

Baseline. A *Passive Learning* (PL) approach acted as a baseline, which simply chooses a subset $\mathcal{N}_p$ of the pool set $\mathcal{P}$ for manual labelling randomly and is therefore considered as a 'non-intelligent' approach. After having collected the label l for an instance, this instance is removed from the pool set and added together with the label to the training set $\mathcal{T}$. This procedure is repeated until all instances of the pool set are labelled.

Measurement. To determine the classification performance, we used the unweighted average recall (UAR), following the recommendation in [27].

4. Results and Discussion

As shown in Figure 1, for all approaches, the sequential addition of manually labelled instances to an initial training set leads to continuous improvements in the performance of the classifier. Further, for all algorithms the UAR first increases steeply with the number of manual annotations and stagnates at some point.

In order to study the effectiveness of the proposed methods, we compare the different learning approaches to PL. The results, as presented in Figure 1 and Table 1, demonstrate that all conventional approaches outperform the baseline PL with 54.17 % UAR. The highest UAR was achieved by DAL_l with a UAR of 63.9 %, directly followed by $svCL_m$ with a maximum UAR of 63.68 %. The maximum UAR, the number of annotations needed to reach this maximum, and the relative annotation cost reduction (CR) of reaching the maximum UAR compared to PL are given in Table 1. As can be seen, the PL algorithm collects more than 15 k annotations to achieve its maximum UAR, while mvCL stops after only 4.4 k manual labelled instances, resulting in the highest overall CR of 70.6 %.

More importantly, when comparing the conventional machine learning approaches to the novel trustability-based ones, a considerable improvement on the UAR can be observed for all algorithms, indicating that the trustability-based annotations result in a more coherent and robust model than the conventional ones. Furthermore, the observed shorter TsvCL, TxvCL, and TmvCL curves indicate that these methods require markedly less manual annotations to achieve the same performance for both, untrusted and trusted algorithms. In order to demonstrate the CR, the costs in terms of the numbers of manual annotations at the highest UAR achieved are compared by each method (cf. Table 1). The overall highest UAR was achieved by the trustability-based DAL_l approach with a maximum UAR of 71.56 %, directly followed by the TSSL method with 71.53 %, and the $TsvCL_l$ approach with a maximum UAR of 71.28 %. More importantly, these findings indicate that the trustability-based CL methods reduce the annotation load compared to PL (15 k annotations) with up to 4.2 k annotations for the $TxvCL_l$ approach and outperforms PL by reducing the annotations drastically by up to 72.0 %, as well as $TDAL_l$ (8.6 k annotations, 42.6 % CR) and TSSL (8.2 k annotations, 45.3 % CR).

5. Conclusion and Outlook

Motivated by a scarcity of annotated data, active learning strategies have been investigated to reduce the cost of gathering labels for audio datasets. In this regard, we introduced the idea of incorporating an annotator trustability score into different machine learning approaches. Making use of the advantages of crowdsourcing to collect annotations in a fast and cost-effective manner, we integrated the proposed trustability-based algorithms into the crowdsourcing platform iHEARu-PLAY, to evaluate our algorithms by performing a range of emotion recognition studies. As a result, our conventional, baseline passive learning (PL) approach required more than 15 k annotations to achieve a maximum UAR of 54.17 %. The trustability-based TXvCL approach achieved a relative annotation cost reduction of up to 72.0 % while achieving a UAR of 71.17 %.

The performed experiments indicate that the proposed trustability-based algorithms have clear advantages over the PL method and conventional active learning approaches. While achieving better performances, the main aspect is the effective way of drastically reducing the number of needed annotations and therefore the need for manual labellers. Hence in conclusion, the trustability-based algorithms are an effective approach combining active learning and the annotator trustability and can therefore be used in crowdsourcing platforms in order to reduce the annotation costs while at the same time potentially improving classification results. Future work will focus on evaluating the algorithms on more databases and with even more diverse trustability scores of the annotators to demonstrate the robustness and performance improvements.

6. Acknowledgment

This research has received funding from the European Community's Seventh Framework Programme (grant agreement No. 338164 – ERC Starting Grant iHEARu). We thank all iHEARu-PLAY users for their annotations and Dr Zixing Zhang for providing initial baseline code for the conventional algorithms.

7. References

[1] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and Structural Biotechnology Journal*, vol. 13, pp. 8 – 17, 2015.

[2] C. Chen, A. Seff, A. Kornhauser, and J. Xiao, "Deepdriving: Learning affordance for direct perception in autonomous driving," in *Proceedings of the International Conference on Computer Vision*, Washington, DC, USA, 2015, pp. 2722–2730.

[3] E. A. Gerlein, M. McGinnity, A. Belatreche, and S. Coleman, "Evaluating machine learning classification for financial trading: An empirical approach," *Expert Systems with Applications*, vol. 54, pp. 193 – 207, 2016.

[4] S. Hantke, Z. Zhang, and B. Schuller, "Towards Intelligent Crowdsourcing for Audio Data Annotation: Integrating Active Learning in the Real World," in *Proceedings INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association*, Stockholm, Sweden, 2017, pp. 3951–3955.

[5] N. Ide and J. Pustejovsky, *Handbook of Linguistic Annotation*. Springer, 2017.

[6] M. Sabou, K. Bontcheva, L. Derczynski, and A. Scharl, "Corpus annotation through crowdsourcing: Towards best practice guidelines," in *Proceedings of the International Conference on Language Resources and Evaluation*, Reykjavik, Iceland, 2014, pp. 859–866.

[7] A. Wang, C. D. V. Hoang, and M.-Y. Kan, "Perspectives on crowdsourcing annotations for natural language processing," *Language resources and evaluation*, vol. 47, pp. 9–31, 2013.

[8] U. Gadiraju, R. Kawase, S. Dietze, and G. Demartini, "Understanding malicious behavior in crowdsourcing platforms: The case of online surveys," in *Proceedings of the Annual Conference on Human Factors in Computing Systems*, New York, NY, USA, 2015, pp. 1631–1640.

[9] D. E. Difallah, G. Demartini, and P. Cudré-Mauroux, "Mechanical Cheat: Spamming Schemes and Adversarial Techniques on Crowdsourcing Platforms," in *Proceedings of the International Workshop on Crowdsourcing Web Search*, Lyon, France, 2012, pp. 26–30.

[10] Y. Zhang, E. Coutinho, Z. Zhang, C. Quan, and B. Schuller, "Dynamic Active Learning Based on Agreement and Applied to Emotion Recognition in Spoken Interactions," in *Proceedings of the International Conference on Multimodal Interaction*, Seattle, WA, USA, 2015, pp. 275–278.

[11] E. Boiy and M.-F. Moens, "A machine learning approach to sentiment analysis in multilingual Web texts," *Information retrieval*, vol. 12, pp. 526–558, 2009.

[12] Y. Yan, R. Rosales, G. Fung, and J. G. Dy, "Active learning from crowds," in *Proceedings of the International Conference on Machine Learning*, Bellevue, Washington, USA, 2011, pp. 1161–1168.

[13] O. Alonso, D. E. Rose, and B. Stewart, "Crowdsourcing for relevance evaluation," *SIGIR Forum*, vol. 42, pp. 9–15, 2008.

[14] S. Novotney and C. Callison-Burch, "Cheap, fast and good enough: Automatic speech recognition with non-expert transcription," in *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Los Angeles, USA, 2010, pp. 207–215.

[15] A. Tarasov, S. J. Delany, and C. Cullen, "Using crowdsourcing for labelling emotional speech assets," Paris, France, pp. 1–4, 2010.

[16] P.-Y. Hsueh, P. Melville, and V. Sindhwani, "Data quality from crowdsourcing: a study of annotation selection criteria," in *Proceedings of the NAACL HLT 2009 workshop on active learning for natural language processing*, Boulder, Colorado, USA, 2009, pp. 27–35.

[17] S. Hantke, E. Marchi, and B. Schuller, "Introducing the Weighted Trustability Evaluator for Crowdsourcing Exemplified by Speaker Likability Classification," in *Proceedings of the International Conference on Language Resources and Evaluation*, Portoroz, Slovenia, 2016, pp. 2156–2161.

[18] V. Ambati, S. Vogel, and J. Carbonell, "Active learning and crowd-sourcing for machine translation," in *Proceedings of the International Conference on Language Resources and Evaluation*, Valletta, Malta, 2010, pp. 2169–2174.

[19] F. Laws, C. Scheible, and H. Schütze, "Active learning with amazon mechanical turk," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA, 2011, pp. 1546–1556.

[20] Y. Zhang, E. Coutinho, Z. Zhang, M. Adam, and B. Schuller, "On Rater Reliability and Agreement Based Dynamic Active Learning," in *Proceedings of the biannual Conference on Affective Computing and Intelligent Interaction*, Xi'an, P. R. China, 2015, pp. 70–76.

[21] Z. Zhang, E. Coutinho, J. Deng, and B. Schuller, "Cooperative Learning and its Application to Emotion Recognition from Speech," *ACM Transactions on Audio, Speech and Language Processing*, vol. 23, pp. 115–126, 2015.

[22] S. Hantke, F. Eyben, T. Appel, and B. Schuller, "iHEARu-PLAY: Introducing a game for crowdsourced data collection for affective computing," in *Proceedings of the International Workshop on Automatic Sentiment Analysis in the Wild, satellite of the biannual Conference on Affective Computing and Intelligent Interaction*, Xi'an, P. R. China, 2015, pp. 891–897.

[23] S. Hantke, T. Appel, and B. Schuller, "The Inclusion of Gamification Solutions to Enhance User Enjoyment on Crowdsourcing Platforms," in *Proceedings of the Asian Conference on Affective Computing and Intelligent Interaction*, Beijing, P. R. China, 2018, 6 pages.

[24] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: transfer learning from unlabeled data," in *Proceedings of the International Conference on Machine learning*, Corvallis, Oregon, 2007, pp. 759–766.

[25] J. Platt *et al.*, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, vol. 10, pp. 61–74, 1999.

[26] S. Steidl, "Automatic classification of emotion related user states in spontaneous children's speech," Ph.D. dissertation, University of Erlangen-Nuremberg, 2009.

[27] B. Schuller, S. Steidl, and A. Batliner, "The Interspeech 2009 Emotion Challenge," in *Proceedings INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association*, Brighton, UK, 2009, pp. 312–315.

[28] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, and M. Pantic, "AVEC 2017 real-life depression, and affect recognition workshop and challenge," in *Proceedings of the International Workshop on Audio/Visual Emotion Challenge*, Mountain View, CA, USA, 2017, pp. 1–8.

[29] M. Schmitt, E. Marchi, F. Ringeval, and B. Schuller, "Towards cross-lingual automatic diagnosis of autism spectrum condition in children's voices," in *Proceedings of the ITG Symposium on Speech Communication*, Paderborn, Germany, 2016, pp. 1–5.

[30] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor," in *Proceedings of the International Conference on Multimedia*, Barcelona, Spain, 2013, pp. 835–838.

[31] M. A. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *SIGKDD Explorations*, vol. 11, pp. 10–18, 2009.

[32] F. Provost, "Machine learning from imbalanced data sets 101," in *Proceedings of the AAAI Workshop on Learning from Imbalanced Data Sets*, Austin, TX, USA, 2000, pp. 1–3.