Towards Intelligent Crowdsourcing for Audio Data Annotation: Integrating Active Learning in the Real World

Simone Hantke[1,2], Zixing Zhang[1], and Björn Schuller[1,3]

[1]Chair of Complex & Intelligent Systems, University of Passau, Germany
[2]Machine Intelligence & Signal Processing Group, Technische Universität München, Germany
[3]Department of Computing, Imperial College London, UK

`simone.hantke@uni-passau.de`

Abstract

In this contribution, we combine the advantages of traditional crowdsourcing with contemporary machine learning algorithms with the aim of ultimately obtaining reliable training data for audio processing in a faster, cheaper and therefore more efficient manner than has been previously possible. We propose a novel crowdsourcing approach, which brings a simulated active learning annotation scenario into a real world environment creating an intelligent and gamified crowdsourcing platform for manual audio annotation. Our platform combines two active learning query strategies with an internally calculated trustability score to efficiently reduce manual labelling efforts. This reduction is achieved in a twofold manner: first our system automatically decides if an instance requires annotation; second, it dynamically decides, depending on the quality of previously gathered annotations, on exactly how many annotations are needed to reliably label an instance. Results presented indicate that our approach drastically reduces the annotation load and is considerably more efficient than conventional methods.

Index Terms: Intelligent Crowdsourcing, Active Learning, Annotation Reduction, Audio Processing

1. Introduction

Supervised machine learning techniques are highly dependent on the amount and quality of labelled training data. Therefore, a crucial step in building supervised classifiers is the manual annotation of data. Conventionally, data annotation has been performed by groups of experts in a traditional laboratory setting, involving a large amount of labour. Therefore, acquiring these manual expert annotations is costly, time-consuming, and tedious work [1, 2, 3, 4].

Technologies such as the *Internet of Things* have made it easier than ever to collect vast and truly big amounts of data. *Crowdsourcing* has been shown to be a viable alternative to conventional labelling paradigms to rapidly collect the mass of annotations needed to leverage new data sources [5]. Whilst crowdsourcing has many positive aspects including efficiency and cost reduction, the online recruitment of anonymous annotators brings new and different issues especially in relation to the annotation quality. In this regard, we recently developed the gamified crowdsourcing platform iHEARu-PLAY [1] [6]. The platform offers audio, video and image labelling for a diverse range of annotation tasks. In addition, the platform ensures a high quality of annotations through optimized data quality management, while the gamification aspect reduces the mental boredom of the annotators.

---

[1]`https://www.ihearu-play.fim.uni-passau.de`

A further issue is that conventional crowdsourcing still requires large amount of human efforts; at least as many labels as there are unlabelled data instances need to be provided. State-of-the art optimization techniques like *Active Learning algorithms* reduce the number of data instances which need manual labelling [7, 8, 9, 10] and are capable of reducing the time-consuming and expensive manual labelling work. However, the benefits that such techniques can bring to crowdsourcing has yet to be fully investigated.

1.1. Related Work

Over the past few years, several crowdsourcing platforms have been established; examples include Mechanical Turk, Crowd-Flower, Turkit, Mob4hire, uTest, Freelancer, eLance, Trada, 99design, Innocentive, CloudCrowd, and Cloud-Flower [11]. These platforms all give monetary compensation to the annotator; however, as this amount can be small and the mundane work there is often a lack of appeal for this type of labelling work. A new strategy named *Games with a Purpose* was introduced to get individuals involved in research projects by motivating through a joyful environment instead of financial rewards [12, 13], resulting in the development of iHEARu-PLAY [6].

Even though the annotator might be more motivated by the gamification elements to perform labelling tasks, the gathered labels may still lack an expert-like quality. Diverse studies on examining the quality of the annotation data from expert annotators in a research lab and non-expert annotators from the internet have been performed and much work was done on the quality management to increase the quality of non-experts labels with special applied mechanisms [14, 15, 16, 17].

Active learning has been a rich literature in machine learning to efficiently exploit unlabelled data for model training. It has been applied to many diverse domains such as machine translation [1, 18], medical imaging [19], classification tasks [20], sentiment detection [9, 21], and text classification [22]. Its effectiveness was shown in multimedia retrieval [23], typical classification tasks such as automatic speech recognition [24], and speech emotion recognition [8, 10], resulting in e. g., 79.17% less manual annotation effort as can be seen in [10].

1.2. Contributions of this Paper

This paper proposes and explores a novel *intelligent crowdsourcing* approach which combines a state-of-the-art active learning algorithm with a gamified crowdsourcing platform, in order to combine the accuracy of manual labelling with the speed and cost-effectiveness of machine learning classifiers. Whilst – to the best of the authors knowledge – most studies on active learning have only simulated the active learning annotation scenario, herein, we present the first gamified intelligent
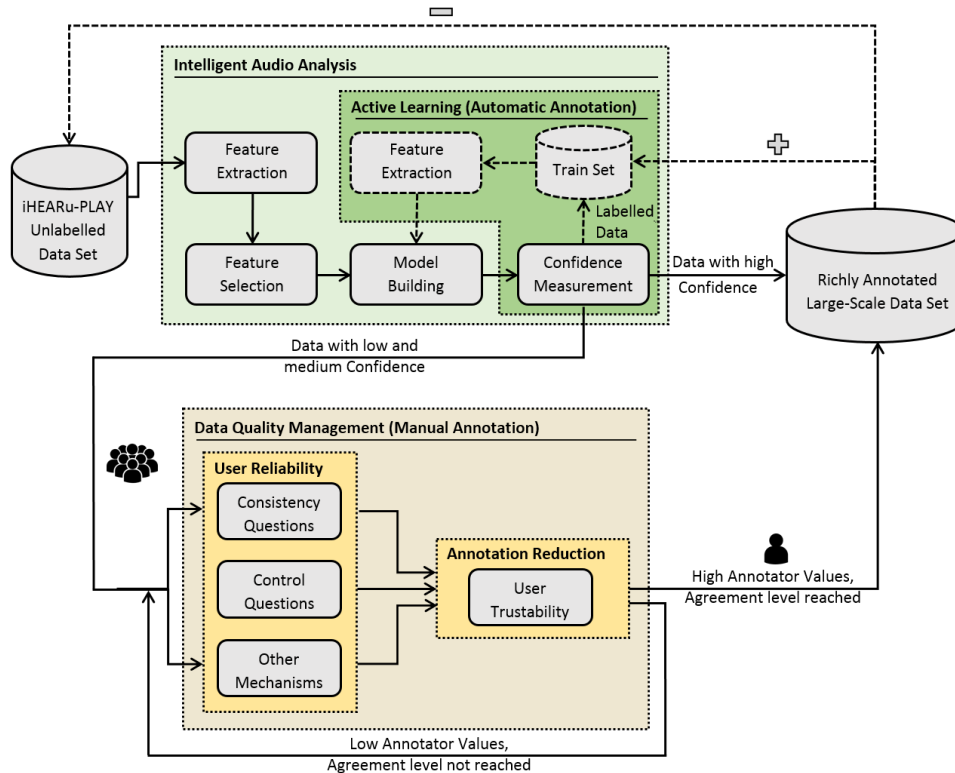
Figure 1: *iHEARu-PLAY's interaction between the intelligent audio analysis, the active learning, and the data quality management components, including the user reliability calculation and the annotation reduction components [6].*

crowdsourcing platform for gathering expert-like manual audio annotations, exemplified through an emotion recognition experiment. We show the combination of two active learning strategies and an integrated data quality management – including a user trustability calculation depending on the correctness of the given answers to *Consistency* and *Control* Questions – which is capable of (i) deciding in the first place whether an instance needs labelling, or not, and (ii) dynamically deciding on how many annotators opinions are needed per instance, depending on the trustability of the user and the quality of the label. By incorporating these methods, our platform makes it feasible to reduce a large amount of unneeded annotations to obtain labelled training data in a faster, cheaper and higher quality way than it was previously possible.

2. Proposed Intelligent Crowdsourcing Platform

In this section, we describe the proposed audio annotation scheme in iHEARu-PLAY; an overview of the system is depicted in Figure 1.

2.1. Intelligent Audio Analysis

Data owners can upload their audio data to iHEARu-PLAY, which will then automatically run through the *Intelligent Audio Analysis* component. After having chosen a feature set out of a pool of different available feature sets, the acoustic features will be extracted by using the integrated openSMILE toolkit [25]. Then, a classifier is automatically trained with the already on iHEARu-PLAY pre-labelled small amount of training data and the results are automatically transferred to the *Active Learning* component.

2.2. Active Learning Algorithms

The main idea of an active learning algorithm is to improve the performance accuracy with as little training data as possible, if the algorithm is allowed to choose the data from which it learns [26, 27]. The integrated active learning algorithm extracts a subset of instances, based on the prediction confidence values, and creates a sorted list from the highest confidence on the instances to the lowest. This subset is then removed from the unlabelled data and automatically passed on for manual labelling by creating a new dataset within iHEARu-PLAY.

Table 1 describes the active learning algorithms used in this work which are based on the *least certainty* query strategy and on the *medium certainty* query strategy, which have been introduced in [28]. For each query strategy, the algorithm starts by classifying all instances of the unlabelled data pool U using a model trained on a small pool of data L, previously labelled using iHEARu-PLAY. Upon the posterior probability, the confidence values assigned to each instance are ranked and stored in a queue Q in descending order. Assuming there are u instances in the unlabelled dataset, then, in the centre of the queue would be the $\lceil u/2 \rceil th$ one. Finally, a subset N_a of U corresponding to those instances predicted with least or medium confidence values are sent to manual annotation within iHEARu-PLAY. Thenceforth, these instances are added to the training set and removed from the unlabelled data set U. This sequential process is repeated until a predefined number of instances are selected or until some stopping criterion is met [28].

2.3. User Reliability and Annotation Reduction

Whilst active learning is mainly used to expedite the learning process and aims to reduce the labelling efforts required by a human expert, crowdsourcing, on the other hand, reduces the

Table 1: *Pseudocode description of the active learning algorithm based on the least certainty query strategy and on the medium certainty query strategy.*

Algorithm: Active Learning with least certainty query strategy and with medium certainty query strategy

Loop:

1. (Optional) Upsample the training set L to obtain even class distribution L_D
2. Use L/L_D to train classifier h, and then classify the unlabelled set U
3. Rank the data based on the prediction confidence values C and store them in queue Q
4. • **Least certainty query strategy:** Select subset N_a whose elements are 'on the bottom' of the ranking queue Q

   • **Medium certainty query strategy:** Select subset N_a whose elements are 'in the middle' of the ranking queue Q
5. Submit the selected instances N_a to manual annotation

   Loop:

   (a) Compute the aggregated manual labels and assess the users' confidences using the designed User Reliability checks

   (b) Add instances with high users' confidence $N_{a(h)}$ to the labelled dataset L, $L = L \cup N_{a(h)}$

   (c) Remove $N_{a(h)}$ from the unlabelled set U, $U = U - N_{a(h)}$

   (d) Keep the instances with low users' confidence $N_{a(l)}$ for the next iteration in order to get more manual labels (until a stop criteria is fulfilled)
6. Add the high confidence instances $N_{a(h)}$ and their aggregated labels to the training set L, $L = L \cup N_a(h)$

---

cost or workload per annotation in a fast and efficient way. The goal of iHEARu-PLAY is to obtain annotations from non-expert annotators that are qualitatively close to gold standard annotations created by experts. Therefore, several data quality mechanisms have been applied, such as pre-time quality checks and tracking the users' behaviour. Further, iHEARu-PLAY offers (i) *Control Questions*, which are definitely wrong answers used to detect players who do not read the question and/or just select a random answer and (ii) *Consistency Questions*, which are repeated certain questions from within the data files currently being annotated by a particular user, and then comparing the answer with the previous answer(s). Giving 'inconsistent' or 'wrong' answers decreases a *Users Trustability* whilst a 'correct' answer increases or maintains it.

All these methods can be used as a measurement to weigh the gathered annotations for example with the help of the *Weighted Trustability Evaluator* [14]. In addition, under the assumption that a 'good' user with a high User Trustability creates better annotations than a 'bad' user with a low User Trustability, the number of redundant annotations can be minimised, since this information can be used to define a quality threshold for the number of needed annotations. Annotations from trusted users will then be added to the final labelled dataset. If a user is not trusted, the file will be sent back to a new round of labelling until the given minimum number of required annotations is met. This procedure of the whole components is iterated until a defined criterion is met and all annotations are given.

3. Experiments

We provide proof-of-concept results on the integration of the active learning algorithm into iHEARu-PLAY. The exemplary experiments performed are a set of binary arousal (describing how strong or weak an emotion is) and valence (describing how positive or negative an emotion is) recognition tasks.

3.1. Database

Since the perception of emotions differs between native speakers and non-native speakers [29], we chose a database which presents nonsensical utterances, and therefore does not contain any contextual meaning to ensure the annotators were able to only annotate the emotions without taking the content into account. Therefore, we chose the GEMEP database, which is an acted database, including portrayals of very intense emotion utterances pronounced by five male and five female professional native-French speaking actors [30].

3.2. Acoustic Features

The acoustic feature set used in our experiments corresponds to the feature set of the INTERSPEECH 2010–2016 Challenges [31]. We used the open-source openSMILE feature extraction toolkit [32] to 'brute-force' the high-dimensional feature set by applying statistical functionals to frame-wise low-level descriptors, which comprise energy, spectral, and voicing related LLDs. Regarding functionals, the feature set is a compromise between a broad variety of functionals, including mean, min, max, and moments. Altogether, the INTERSPEECH 2013–2016 Challenge feature set contains 6 373 features.

3.3. Setup

Using the open-source Weka toolkit [33], *Support Vector Machines* (SVMs) were employed as the modelling paradigm, initially trained with a Sequential Minimal Optimisation (SMO) algorithm with a polynomial kernel and a constant complexity of 0.01. For our uncertainty based active learning experiments we require prediction confidences. Therefore, we converted the distances to probability estimates within the range of [0,1] and further calculated the prediction confidence value. More details about the conversion procedure can be found in [28].

3.4. Evaluation

Annotations: The internally running data quality management component gave us reliable annotations per instances used as a starting training set; i. e., all annotators used had a trustability score of 100%. All in all we received seven reliable annotations per file. Out of the annotators: 3 were female and 4 male; one Indian, one Chinese, three German, one Italian and one British; and their age ranking from 27 to 34 years (only 5 users gave us the information on their age), with a mean of 29.4 years and a standard deviation of 3.13 years.

Baseline: We chose a passive learning algorithm as a baseline to compare the effectiveness of the active learning algorithm. The passive learning algorithm randomly selects a subset from the unlabelled set and asks a human to label it. The labelled files will then be removed from the unlabelled set and added to the labelled set. This approach is traditionally followed by 'non-intelligent' crowdsourcing and should be interrupted as our baseline result. For a better comparison of the algorithms itself, we also used the trusted annotations for the passive learning algorithm.

Learning Process: We randomly selected 30 instances for both active learning and passive learning as the initial training set, which was approximately 3 % of the whole pool set. We
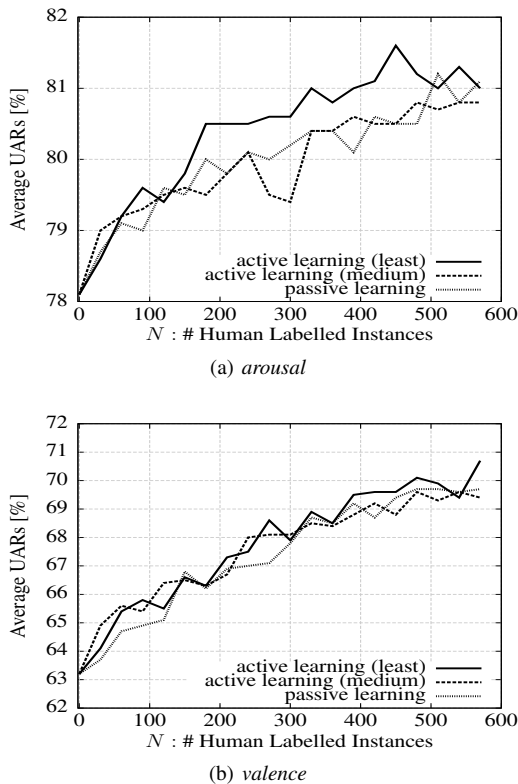
(a) *arousal*

(b) *valence*

Figure 2: *Comparison between intelligent crowdsourcing (integrated active learning component) and traditional crowdsourcing (passive learning – without active learning) for audio data annotation, evaluated by the tasks of binary arousal and valence recognition. The performance measures show the average unweighted average recalls (UARs) values averaged across 30 independent runs of the algorithm vs the number of selected instances N (human annotations).*

further randomly selected additional 30 instances for passive learning or actively selected additional 30 instances for active learning, respectively. For our experiments, we set the iteration times to be as 20 for both algorithms. Further, in order to reduce the statistical impact, we conducted the whole iterations with 30 independent runs.

Measurement: As an evaluation measurement of the performance of the proposed intelligent crowdsourcing platform, we utilised the *Unweighted Average Recall* (UAR). The motivation to consider UAR rather than other measures such as weighted average recall (WAR) is that it can better reflect the overall accuracy in the presence of imbalanced classes as well as for more than two classes, and is widely used for emotion, and even other computational paralinguistics recognition tasks [31, 34].

4. Results

Figure 2 presents the results of our proof-of-concept experiments on intelligent crowdsourcing, showing the average UAR measure for the passive and active learning algorithms, including both the least and medium certainty query strategies. According to the characteristic curve progression of the active learning algorithm, the sequential addition of manually labelled instances to the initial training set (using only 30 per iteration) leads to continuous improvements in the performance of the classifier. The UAR increases steeply with the number of to-

Table 2: *Maximum classification results by unweighted average recall (UAR) in percent for active learning (AL) with (L)east and (M)edium query strategies and passive learning (PL) for (Aro.)usal and (Val.)ence. Results verify the proof-of-concept.*

	AL-L		AL-M		PL	
	Aro.	Val.	Aro.	Val.	Aro.	Val.
UAR	**81.6**	**70.7**	80.8	69.6	81.2	69.7

tal manual annotations and more importantly, it can be seen that overall higher classification accuracy can be achieved especially through the active learning least certainty query strategy with the same annotation effort as in passive learning (UAR = 81.6% for arousal, 70.7% for valence). For the medium certainty query strategy we archived an UAR = 80.8% for arousal and 69.6% for valence (cf. also Table 2). Active learning appears to have greater benefit when performing arousal recognition, which is not unexpected, as this is a speech-based study and it is a well-known fact that arousal is easier to recognize in speech than valence.

For a more detailed analysis and to statistically compare the performances, we computed the Student's t-test, which was executed on the averaged UARs of each algorithm for both arousal and valence. We found, that especially the least certainty query strategy for the active learning algorithm can achieve significant performance improvement for both arousal and valence tasks compared with the passive learning algorithm ($p < 0.05$). However, the performance improvement cannot be seen as significant for the medium certainty query strategy ($p > 0.05$). The analysis of the significance levels confirms our previous observation and indicates that the active learning least certainty query strategy approach generally leads to significantly better performance than passive learning.

5. Conclusion and Outlook

We introduced a novel approach, where active learning comes together with gamified crowdsourcing for audio data annotation, resulting in the first real-world intelligent crowdsourcing platform iHEARu-PLAY. We introduced first baseline results, showing the effectiveness of the proposed system containing the active learning algorithm with least and medium certainty query strategy. Our proof-of-concept experiments show that a higher classification accuracy is achieved especially through the active learning least certainty query strategy (UAR = 81.6% for arousal, 70.7% for valence) with the same annotation effort as in passive learning.

Future work will focus on extending and improving the classification accuracy of the active learning algorithm, make it even more dynamic based on the user reliability, and run it on larger datasets. Since iHEARu-PLAY is multi-modal and images and videos can also be annotated, we would like to make the active learning component available for these data types as well. Thanks to a soon up-coming researcher portal, iHEARu-PLAY will open for research teams from all over the world to get their data annotated freely.

6. Acknowledgements

7. References

[1] V. Ambati, S. Vogel, and J. Carbonell, "Active learning and crowd-sourcing for machine translation," in *Proc. of Int. Conf. on Language Resources and Evaluation*. Valletta, Malta: ELRA, May 2010, pp. 2169–2174.

[2] V. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy, "Learning from crowds," *Journal of Machine Learning Research*, vol. 11, pp. 1297–1322, 2010.

[3] A. Kittur, E. Chi, and B. Suh, "Crowdsourcing for usability: Using micro-task markets for rapid, remote, and low-cost user measurements," in *Proc. of Int. Conf. on Human Factors in Computing Systems*. Florence, Italy: ACM, April 2008.

[4] A. Tarasov, S. J. Delaney, and C. Cullen, "Using crowdsourcing for labelling emotional speech assets," in *W3C Workshop on Emotion Markup Language*, Paris, France, October 2010, no pagination.

[5] H. Zhai, T. Lingren, L. Deleger, Q. Li, M. Kaiser, L. Stoutenborough, and I. Solti, "Web 2.0-Based Crowdsourcing for High-Quality Gold Standard Development in Clinical Natural Language Processing," *J Med Internet Res*, vol. 15, p. 73, 2013.

[6] S. Hantke, F. Eyben, T. Appel, and B. Schuller, "iHEARu-PLAY: Introducing a game for crowdsourced data collection for affective computing," in *Proc. of Int. Workshop on Automatic Sentiment Analysis in the Wild held in conjunction with the biannual Conf. on Affective Computing and Intelligent Interaction*. Xi'an, China: IEEE, September 2015, pp. 891–897.

[7] Z. Zhang, N. Cummins, and B. Schuller, "Advanced Data Exploitation in Speech Analysis – An Overview," *IEEE Signal Processing Magazine*, vol. 34, p. 24, 2017.

[8] Z. Zhang and B. W. Schuller, "Active Learning by Sparse Instance Tracking and Classifier Confidence in Acoustic Emotion Recognition," in *Proc. of INTERSPEECH*. Portland, USA: ISCA, September 2012, pp. 362–365.

[9] 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.

[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 *Proc. of Int. Conf. on Multimodal Interaction*. Seattle, USA: ACM, November 2015, pp. 275–278.

[11] A. Doan, R. Ramakrishnan, and A. Y. Halevy, "Crowdsourcing systems on the world-wide web," *Commun. ACM*, vol. 54, pp. 86–96, 2011.

[12] B. M. Good and A. I. Su, "Games with a scientific purpose," *Genome Biol*, vol. 12, p. 135, 2011.

[13] L. Von Ahn, "Games with a purpose," *Computer*, vol. 39, pp. 92–94, 2006.

[14] S. Hantke, E. Marchi, and B. Schuller, "Introducing the Weighted Trustability Evaluator for Crowdsourcing Exemplified by Speaker Likability Classification," in *Proc. of Int. Conf. on Language Resources and Evaluation*. Portoroz, Slovenia: ELRA, May 2016, pp. 2156–2161.

[15] O. F. Zaidan and C. Callison-Burch, "Crowdsourcing translation: Professional quality from non-professionals," in *Proc. of Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon: ACL, June 2011, pp. 1220–1229.

[16] M. D. Smucker and C. P. Jethani, "The crowd vs. the lab: A comparison of crowd-sourced and university laboratory participant behavior," in *Proc. of Workshop on crowdsourcing for information retrieval*. Beijing, China: ACM, July 2011, no pagination.

[17] P.-Y. Hsueh, P. Melville, and V. Sindhwani, "Data Quality from Crowdsourcing: A Study of Annotation Selection Criteria," in *Proc. of Workshop on Active Learning for Natural Language Processing*. Boulder, Colorado: ACL, June 2009, pp. 27–35.

[18] M. Bloodgood and C. Callison-Burch, "Bucking the trend: Large-scale cost-focused active learning for statistical machine translation," in *Proc. of Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden: ACL, July 2010, pp. 854–864.

[19] S. C. Hoi, R. Jin, J. Zhu, and M. R. Lyu, "Batch mode active learning and its application to medical image classification," in *Proc. of Int. Conf. on Machine learning*. Orlando, USA: ACM, December 2006, pp. 417–424.

[20] G. Schohn and D. Cohn, "Less is more: Active learning with support vector machines," in *Proc. of Int. Conf. on Machine Learning*, Stanford, USA, June 2000, pp. 839–846.

[21] A. Brew, D. Greene, and P. Cunningham, "Using crowdsourcing and active learning to track sentiment in online media," in *Proc. of Conf. on Artificial Intelligence*. Lisbon, Portugal: ACM, August 2010, pp. 145–150.

[22] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *Journal of Machine Learning research*, vol. 2, pp. 45–66, 2001.

[23] M. Wang and X.-S. Hua, "Active learning in multimedia annotation and retrieval: A survey," *Transactions on Intelligent Systems and Technology*, vol. 2, pp. 1–21, 2011.

[24] G. Riccardi and D. Hakkani-Tur, "Active learning: Theory and applications to automatic speech recognition," *Transactions on speech and audio processing*, vol. 13, pp. 504–511, 2005.

[25] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor," in *Proc. of ACM Multimedia*. Barcelona, Spain: ACM, October 2013, pp. 835–838.

[26] J. Costa, C. Silva, M. Antunes, and B. Ribeiro, "On using crowdsourcing and active learning to improve classification performance," in *Proc. of Int. Conf. on Intelligent Systems Design and Applications*. Cordoba, Spain: IEEE, November 2011, pp. 469–474.

[27] A. K. McCallumzy and K. Nigamy, "Employing em and pool-based active learning for text classification," in *Proc. of Conf. on Machine Learning*, Madison, Wisconsin, July 1998, pp. 359–367.

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

[29] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA Multimodal Corpus of Remote Collaborative and Affective Interactions," in *Proc. of Int. Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space held in conjunction with the Conf. on Automatic Face and Gesture Recognition*. Shanghai, China: IEEE, April 2013.

[30] T. Bänziger, H. Pirker, and K. Scherer, "GEMEP-GEneva Multimodal Emotion Portrayals: A corpus for the study of multimodal emotional expressions," in *Proc. of Int. Conf. on Language Resources and Evaluation*. Genova, Italy: ELRA, May 2006, pp. 15–19.

[31] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity & Native Language," in *Proc. of INTERSPEECH*. San Francisco, USA: ISCA, September 2016, pp. 2001–2005.

[32] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The Munich Versatile and Fast Open-source Audio Feature Extractor," in *Proc. of Int. Conf. on Multimedia*. Firenze, Italy: ACM, 2010, pp. 1459–1462.

[33] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explor. Newsl.*, vol. 11, pp. 10–18, 2009.

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