MACHINE LEARNING-ASSISTED AFFECT LABELLING OF SPEECH DATA

Alicia F. Requardt, Olga Egorow, Andreas Wendemuth

Cognitive Systems Group, Otto-von-Guericke University, 39016 Magdeburg, Germany
alicia.requardt@ovgu.de

Abstract: This paper addresses the assisted annotation of emotions in affective speech data recorded in natural "in the wild" surroundings. Here, affective states with low expressiveness are encountered which makes manual annotation difficult and very time-consuming even for expert human annotators. Further, the training of an automatic emotion recognition system in such a setup requires high amounts of annotated data. We present a machine-learning-assisted semi-automatic annotation procedure, adopted from speech recognition. We give annotation time estimates and evaluate our approach on data of real-life in-vehicle emotions which are prototypical for natural surroundings. The time necessary for the complete data annotation could be substantially reduced to around 80% of the time needed for the fully manual annotation. At the same time, the quality of the obtained annotation remains the same as of the fully manual approach, in contrast to other currently available approaches such as Active Learning or Semi-Supervised Learning. Having shown the time saving effect, our approach is generally highly useful for annotation processes with high annotation effort.

1 Introduction

In natural "in the wild" surroundings, distinct "full-blown" emotions occur rather seldom – instead, we encounter more or less mixed affective states with low expressiveness. This has been acknowledged in the field of affective computing, leading to more and more investigations on real-life data. Especially in noisy environments, like it is the case for in-vehicle applications, an automatic detection of the emotional state of the driver is challenging, as we do not only need to cope with this low-expressive highly natural speech data, but also with a strong distortions of the speech signal itself.

The training of an automatic emotion recognition system in such a setup requires a high amount of annotated data. At the same time, the emotions' naturalness also renders the annotation process very time-consuming, since the emotional states are hard to recognise even for trained human annotators [1]. This problem has been addressed by developing approaches to facilitate the annotation process. The most prominent of these approaches are active learning (AL), passive learning (PL) and semi-supervised learning (SSL) [2, 3]. The general idea behind these approaches is to train the classifier on a small amount of labelled data and to re-evaluate a sub-set of the classification results. This re-evaluated sub-set can then be used to re-train the initial model. In the case of AL, query strategies are used to identify the most relevant samples of the unlabelled data pool for a later manual annotation. PL pursues a similar strategy, but the samples are chosen at random. For these two approaches, the training material of the classifier still relies on manually labelled data contrary to SSL, where only the initial classifier is

trained on already existing manually labelled data. Afterwards, the model relies completely on the results of the initially trained model and classified data samples achieving a high certainty value.

However, it can be assumed that these approaches will not lead to satisfying results, since real-life in-car data comprise not only noise but also highly natural material with low expressiveness. Automatic recognition for such material remains challenging: Even with the latest developments F1-measures of only 29% for a three class classification problem [4] or 42.54% for the classification of positive, negative, neutral and anxious emotional state [5] can be achieved.

For SSL, the problem is evident: it relies completely on the performance of the initially trained model, therefore the algorithm cannot reach the performance of the classifier trained on the fully manually annotated set. The performance of the SSL self-training and co-training algorithm cannot reach the performance of the corresponding baseline support vector machine (SVM) classifier for discrete speech emotion recognition, converging towards a much lower unweighted average recall (UAR) with an increasing number of labelled instances [6].

For AL, it has been shown that the performance of the baseline classifier (i.e. using the full data set) can be achieved with a lower number of labelled instances. For example, for continuously labelled values of valence and arousal, the same performance as the baseline classification could be achieved using only 88% of the manually labelled data samples [7]. For discrete speech emotion recognition, even a reduction to only 15% of manually labelled data while keeping the classification performance stable is reported [8, 6] – however, outperforming the baseline classification was not possible in this case. Furthermore, it should be kept in mind that these numbers do not consider the amount of labelled data needed for the initial training and evaluation of the classifier – including these samples results in the manual annotation of around 56% of the original data. Also, these approaches pursuit a different goal, aiming at limiting the number of samples requiring manual annotation, whereas we aim at decreasing the annotation time while receiving a fully manual, assisted annotation.

Therefore, we cannot rely on automatically obtained annotations, and the results obtained by AL and SSL need to be verified by human annotators, leading to high annotation effort. This is a problem for projects with limited time and fixed financial resources.

In this paper, we present a novel approach for affective labelling of emotion categories. For this, we adapt an already existing procedure from the domain of speech event recognition implementing a semi-automatic labelling procedure for filled pauses [9]. This approach is applied to the current scenario of in-car affective data and achieves a reduction of annotation time of around 20% while maintaining the high quality of the annotation.

2 Data

In our experiments, we use the data of an existing naturalistic in-vehicle real-world data collection [10, 11]. The data comprise around 7 hours of German speech material (16988 speech samples of approximately 2 seconds length each) of 30 drivers in four emotion categories (positive, neutral, frustrated and anxious) of high naturalness and low expressiveness. All speech samples were annotated by three independent, German speaking, female labellers. The annotation results were obtained in a three step procedure: an annotation of the dimensions of valence and arousal using the 5-point self-assessment manikin scale [12]; an annotation of the four emotion categories and a rating of the annotators' level of satisfaction. In order to obtain a sufficient quality of the labels, the measure of the agreement of the annotators, the inter-rater reliability (IRR) of the annotators, was computed, resulting in a "reliability score" assigned to each of the annotators. If the IRR for the considered speech sample was below a certain threshold, the

label of the annotator lowering the IRR was excluded. Afterwards, the labels of the categorial annotation were assigned to the speech samples based on a majority voting of the most reliable annotators and the labels of the dimensional annotation were assigned by determining the average valence and arousal level over the reliable annotators. The process of annotation took a total of 131.35 hours, with an average of 43.78 hours for each annotator. Excluding the time needed for the satisfaction level annotation, an average of 35.62 hours for each annotator can be calculated. Unfortunately, the separation of the time needed for the categorial and dimensional annotation is not possible, since both annotation processes are not independent and happen simultaneously. This is problematic, as we are only interested in the annotation of the emotion categories. Therefore, we will return to this issue later in Section 3.

As the assignment of the categorial labels is based on a majority voting of only the reliable annotators, not for all samples a clear label assignment was possible. This resulted in 11230 categorically labelled samples, comprising 2150 positive, 5139 neutral, 2329 frustrated and 1612 anxious samples. For each speech sample, the annotators took around 6.66s $\pm$ 1.02s, already including the time needed to listen to the sample and excluding outliers. On average, the annotators listened to each speech sample around 1.46 $\pm$ 0.29 times before making the decision regarding the label.

3 Machine Learning-assisted Annotation Method

Our method presented here is based on an existing method for the annotation of rare speech events such as filled pauses [9]. Therefore, we will first present the original method before describing its application to the process of categorial annotation of emotions.

3.1 Filled Pauses Annotation

The semi-automatic approach for the detection and annotation of filled pauses was developed because of the low amount of available training material for such rare speech events, which in turn results in the high effort for a conventional fully manual data annotation. In this approach, a classifier is trained on an already existing annotated data set. This classifier is then applied to unknown data, resulting in an automatic annotation of filled pauses in new material. The obtained results contain high amounts of misclassifications, yet the effort to manually correct these misclassifications by far undercuts the effort necessary for the fully manual annotation. The approach does not aim at achieving high recognition rates, but rather at increasing the amount of reliable training material. Therefore, the number of not detected filled pauses (i.e. false positives (FPs)) is not important for this approach. Instead, the focus is on the recall and the exact label position of the detected filled pauses (i.e. true positives (TP)). Therefore, there is a manual post-processing step subsequent to the automatic classification: Every detected "area of interest" must be evaluated by an annotator in order to either remove it in the case of an FP or to correct it to the exact starting and ending time of a filled pause in case of a TP. By measuring the time needed to check, remove, verify and adjust the automatically obtained annotation, an average time of 20 seconds for the adjustment of a TP and 5 seconds for removing an FP could be determined. Based on these considerations, an estimate for the time of the semi-automatic annotation process was developed:

$$T_{\text{semi-auto}} = \#\text{TP} \cdot 20s + \#\text{FP} \cdot 5s. \tag{1}$$

Using this approach, the annotation effort could be decreased to up to 85% compared to the conventional fully manual approach.

3.2 Application to Affective Labelling

The approach described above is adapted to the annotation of categorial emotional states. In the first step, a classifier is trained on the fully annotated data of the first subject (from the full set of 30 subjects). This classifier is then applied to classify the unknown speech samples of the next subject. The next step is then to re-evaluate these newly obtained annotated samples by annotators. These re-evaluated samples are then added to the initial data and the classifier is re-trained on this increased amount of data. Then, the process is repeated until the data of all 30 subjects are annotated.

In order to calculate the time reduction compared to the fully manual approach, we develop an estimate similar to the estimate for filled pauses presented above.

For the manual verification process, there are only two cases to be considered: In case of a TP, the classifier set the emotion label correctly, in case of an FP, the emotion label was set incorrectly. We can assume that for expert annotators, the time needed to verify a TP (i.e. listen to the sample and confirm the label) is much lower than the time needed to re-evaluate an FP (i.e. listen to the sample and select a new label). In Section 2, we have already discussed the times necessary for the annotation: around 6.66s to annotate a sample, listening to each sample around 1.46 times. Based on this and assuming that 1 second is needed to verify the perceived emotional state, we can estimate the time $\hat{T}_{\text{TP}}$ needed to verify a TP:

$$\hat{T}_{\text{TP}} = 1.46 \cdot 2s + 1s = 3.92s < 4s. \tag{2}$$

The time needed to re-evaluate a FP can be seen as similar to a fully manual annotation of a speech sample:

$$\hat{T}_{\text{FP}} = 6.66s < 7s. \tag{3}$$

These two estimates together with the time needed to manually annotate the first subject's speech samples T_{S_1} = 4020 seconds result in the following estimate:

$$\hat{T}_{\text{semi-auto}} = \#\text{TP} \cdot \hat{T}_{\text{TP}} + \#\text{FP} \cdot \hat{T}_{\text{FP}} + T_{S_1} = \#\text{TP} \cdot 4s + \#\text{FP} \cdot 7s + 4020s. \tag{4}$$

To verify the stated averaged annotation times, the total time needed for the fully manual annotation $\hat{T}_{\text{man}}$ can be estimated using the following estimate:

$$\hat{T}_{\text{man}} = \#\text{samples} \cdot 7s. \tag{5}$$

This results in an estimated manual annotation time of 33.03 hours. Compared to the real value of 35.62 hours for an annotation without a satisfaction level annotation, this seems to be an appropriate estimate. Furthermore, we need to keep in mind that the real value included two ways of estimation, as already mentioned in Section 2: the annotation in categorial emotions and dimensional emotions. However, this is not a problem since both, the estimated and the real value are based on the same averaged annotation times. Therefore, the time for the dimensional annotation can be seen as an inherent bias leading to an over-estimation, which can be neglected, since we are interested in the upper limit of the annotation time.

4 Results & Discussion

The approach presented above was applied to the data described in Section 2 using a SVM with a radial basis function as a kernel for classification. The classifier was not further optimised, using the default values provided by WEKA [13], namely $\gamma = 1/d$ (with d being the dimension

Table 1 – Confusion matrix of the conducted classification experiments to determine the efficiency of the semi-automatic annotation approach. All TPs are highlighted in green. The last column of the table contains "0" entries, as this class was not included in the classification process.

		Predicted				
		anxious	frustrated	neutral	positive	*N/A*
True	anxious	565	140	655	149	*0*
	frustrated	324	667	853	370	*0*
	neutral	663	467	3199	580	*0*
	positive	308	246	654	882	*0*
	N/A	1297	703	2706	937	*0*

of the feature space) and $C = 1$. The employed feature set is the *emobase* feature set with 988 features [14].

Performed as described in Section 3, the classification achieved the following results: Out of the 16988 samples, 5313 samples could be identified correctly, and 11052 samples were classified incorrectly, requiring manual re-evaluation. Applying the estimates presented in Section 3, the semi-automatic annotation would take 28.51 hours. Compared to the fully manual approach resulting in 35.62 hours, this corresponds to a time reduction of 19.96% for each annotator.

The classification results are shown in Table 1. The categorial annotation was performed by a majority voting, therefore not all samples obtained a label. This case was added as an additional class, referred to as *N/A*. This class was not included in the training process of the classifier – the classifier would still deliver one of the four true classes even for instances labelled as *N/A*, counting all such instances as an FP. This leads to an over-estimation of the annotation time, such that our estimate is the upper limit for the required annotation time.

The procedure also has one additional positive side effect, namely the continuous improvement of the classifier's performance as the amount of samples increases. This is of great interest for applied research, for instance if a certain performance of the classification must be ensured before switching to an AL or SSL approach. This performance improvement is shown in Figure 1. We can observe that the classification performance increases with the number of subjects, however not linearly, due to the natural differences in the data of different subjects. Even with high quality headset recordings, we cannot expect "in the wild" data to be exactly the same over different subjects. Furthermore, not all subjects were able to express the induced emotional state in a recognisable way (e.g. the subject no. 27, as indicated by the performance drop seen in Figure 1).

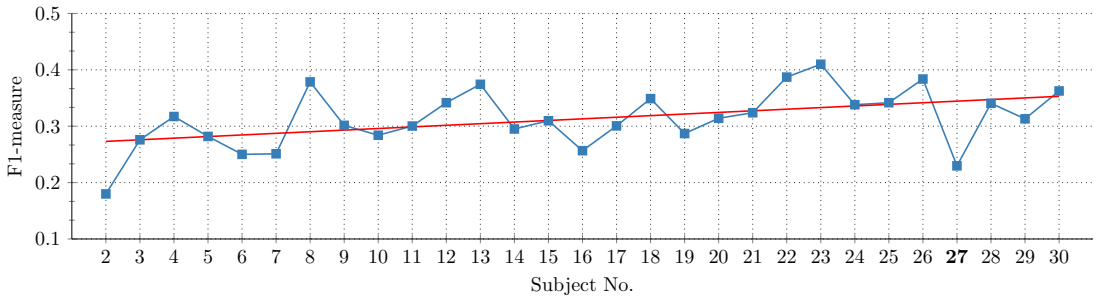

Figure 1 – Improvement of the classification performance in terms of F1-measure with increasing amount of training data. The red line represents a linear regression of the data points, indicating a high variance but a clear increase in F1 ($R^2 = 0.2203$).

5 Conclusion

We presented an application of a semi-automatic approach from the domain of speech recognition to the domain of real-life in-vehicle emotions. Using the developed approach, the time necessary for the complete data annotation could be substantially reduced to around 80% of the time needed for the fully manual annotation. This is highly useful for annotation processes with high annotation effort, such as natural affects. At the same time, the quality of the obtained annotation remains the same as of the fully manual approach, in contrast to other currently available approaches such as AL and SSL.

However, it should be noted that the results presented here are upper estimates of the actual annotation time needed to conduct the machine learning-assisted affective labelling. As already mentioned above, the data set that served as the base for the obtained results leads to an overestimate of the annotation time. We should keep in mind that this data set is a special case of in-car emotional data. Therefore, the achieved reduction of annotation time cannot be assumed to be universal. To obtain a more general information on the performance of the annotation approach, the approach should be tested on different data sets.

Acknowledgement

The authors would like acknowledge support by the project "Intention-based Anticipatory Interactive Systems" (IAIS) funded by the Federal State of Sachsen-Anhalt, Germany and the project "Mod3D" (grant number: 03ZZ0414) funded by 3Dsensation within the Zwanzig20 funding program by the German Federal Ministry of Education and Research (BMBF). Further, this paper has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 688900.

References

[1] TRUONG, K. P., D. A. VAN LEEUWEN, and F. M. DE JONG: *Speech-based recognition of self-reported and observed emotion in a dimensional space*. *Speech Communication*, 54, pp. 1049–1063, 2012.

[2] SETTLES, B.: *Active learning literature survey*. Computer Sciences Technical Report 1648, Computer Sciences, University of Wisconsin-Madison, 2010.

[3] ZHU, X.: *Semi-supervised learning literature survey*. Computer Sciences Technical Report 1530, Computer Sciences, University of Wisconsin-Madison, 2008.

[4] CEVHER, D., S. ZEPF, and R. KLINGER: *Towards multimodal emotion recognition in german speech events in cars using transfer learning*. In *15th Conference on Natural Language Processing (KONVENS)*, pp. 79–90. Deutsche Gesellschaft für Sprachwissenschaften (DGfS), Erlangen, Germany, 2019. Accepted.

[5] REQUARDT, A. F., K. IHME, M. WILBRINK, and A. WENDEMUTH: *Towards affect-aware vehicles for increasing safety and comfort: Recognizing driver emotions from audio recordings in a realistic driving study*, 2020. Submitted.

[6] ZHANG, Z., E. COUTINHO, J. J. DENG, and B. SCHULLER: *Cooperative learning and its application to emotion recognition from speech*. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(1), pp. 115–126, 2015.

[7] HAN, W., H. LI, H. RUAN, L. MA, J. SUN, and B. SCHULLER: *Active learning for dimensional speech emotion recognition*. In *Proceedings of the INTERSPEECH 2013*. 2013.

[8] ZHANG, Z. and B. SCHULLER: *Active learning by sparse instance tracking and classifier confidence in acoustic emotion recognition*. In *Proceedings of the INTERSPEECH 2012*. 2012.

[9] EGOROW, O., A. F. LOTZ, I. SIEGERT, R. BÖCK, and A. WENDEMUTH: *Accelerating manual annotation of filled pauses by automatic pre-selection*. In *Proc. of the 2017 Int. Conf. on Companion Technology (ICCT)*, pp. 1–6. IEEE, Ulm, Germany, 2017.

[10] LOTZ, A. F., F. FALLER, I. SIEGERT, and A. WENDEMUTH: *Emotion recognition from disturbed speech – towards affective computing in real-world in-car environments*. In *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2018*, pp. 208–215. TUDpress, Dresden, 2018.

[11] REQUART, A. F., M. WILBRINK, I. SIEGERT, M. JIPP, A. WENDEMUTH, and K. IHME: *An experimental paradigm for inducing emotions in a real world driving scenario evidence from self-report, annotation of speech data and peripheral physiology*. *Kognitive Systeme*, 2018(1), 2018.

[12] BRADLEY, M. and P. LANG: *Measuring emotion: The self-assessment manikin and the semantic differential*. *Journal of Behavior Therapy and Experimental Psychiatry*, 25, pp. 49–59, 1994.

[13] HALL, M., E. FRANK, G. HOLMES, B. PFAHRINGER, P. REUTEMANN, and I. H. WITTEN: *The weka data mining software: An update*. *ACM-SIGKDD Exploitations Newsletter*, 11(1), pp. 10–18, 2009.

[14] EYBEN, F., M. WÖLLMER, and B. SCHULLER: *openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor*. In *Proc. of the ACM MM-2010*, p. s.p. Firenze, Italy, 2010.