Show Me What You've Learned: Applying Cooperative Machine Learning for the Semi-Automated Annotation of Social Signals

Johannes Wagner[1], Tobias Baur[1], Dominik Schiller[1], Yue Zhang[2], Björn Schuller[2], Michel Valstar[3], Elisabeth André[1]

[1] Augsburg University

[2] Imperial College London

[3] University of Nottingham

{wagner,baur,schiller,andre}@hcm-lab.de

{yue.zhang1,bjoern.schuller}@imperial.ac.uk

michel.valstar@nottingham.ac.uk

Abstract

In this paper we suggest the use of Cooperative Machine Learning (CML) to reduce manual labelling efforts while simultaneously generating an intuitive understanding of the learning process of a classification system. To this end, we introduce the open-source tool NOVA, which aims to combine human intelligence and machine learning to annotate social signals in large multi-modal corpora. NOVA features a semi-automated labelling process in which users are provided with immediate visual feedback on the predictions, which affords insights into the strengths and weaknesses of the underlying classification system. Following an interactive and exploratory workflow, the performance of the model can be improved by manual revision of the predictions, a process that uses confidence values to guide the inspection.

1 Introduction

In various research disciplines (Behavioural Psychology, Medicine, Anthropology, ...) the annotation of social behaviours is a common task. This process includes manually identifying relevant behaviour patterns in audio-visual material and assigning descriptive labels. Generally speaking, segments in the signals are labelled using sets of discrete classes or continuous scores, e. g., a certain type of gesture, a social situation (e. g., conflict), or the emotional state of a person. In Social Signal Processing (SSP), a subset of these events – the so called *social signals* – are used to augment the spoken part of a message with non-verbal information to enable a more natural human-computer interaction [Vinciarelli *et al.*, 2009][1]. To automatically detect social signals from raw sensory input (e. g., speech signals) it is common practice to apply machine learning (ML) techniques. That is, sensory input is transformed into a compact set of relevant features and a classifier is trained on manually labelled examples to optimise a learning function. Once trained, the classifier is used to automatically predict labels on unseen data.

However, since humans transmit non-verbal messages through a number of channels (voice, face, gestures, etc.) and due to the complex interplay between these channels, large amounts of annotated data are necessary to cover those phenomena. Therefore, the progress in the field of SSP is directly linked to the availability of large and well transcribed multi-modal databases rich of human behaviour under varying context and different environmental settings [Douglas-Cowie *et al.*, 2003]. Common challenges in creating such datasets lie in the high degree of naturalness required of the recording scenarios, how well one recording scenario generalises to other settings, the number of human raters needed to reach a consensus on labels, and of course the sheer amount of data. When one considers the many hours of labelled data that are required, gathering such large amounts of annotated training samples may seem like an infeasible task, with respect to time, cost and effort.

An obvious solution is to exploit computational power to accomplish some of the annotation work automatically. However, to ensure the quality of the predicted annotations this still requires human supervision to identify and correct errors. To keep the human effort as low as possible, it is useful to understand why a model makes wrong assumptions. Therefore, it is not only important to provide tools that ease the use of semi-automated labelling, but also to increase the transparency of the decision process (a non trivial task given that most modern classifiers come as black boxes). By visualising the predictions, for instance, even non ML experts get an idea about the strengths and weaknesses of the underlying classification model and can immediately decide which parts of a prediction are worth keeping. If a particular label is regularly missed, a user could actively provide more training examples for this phenomenon, or redesign the ML system to capture its relevant characteristics better. Ideally, the system even guides his or her attention towards parts where manual revision is necessary. Once an annotation has been revised, the model can be retrained to improve its performance for the next cycle. This procedure can be repeated until a desired

[1] To give an example of a social signal, think of a situation where we say something in a sarcastic voice to indicate that we actually mean the opposite.

performance is reached.

In this paper, we introduce an annotation tool called NOVA ((Non-)Verbal Annotator), which implements the described workflow that interactively incorporates the 'human in the loop'. In particular, NOVA offers an interface to acquire semi-automated annotations and provides visual feedback to inspect and correct machine-generated labels. In that sense, our work combines three hot topics of ML: *Explainable Artificial Intelligence*, as the transparency of the decision process is increased via visualisation of the predictions; *Active Learning*, since labels with low confidence are highlighted to guide the user towards relevant parts; and finally, *Interactive Machine Learning*, because human intelligence and machine power can cooperate and improve each other. We subsume our approach under the term *Cooperative Machine Learning* (see Section 3).

2 Related Work

Despite vast resources of raw data, nowadays pervasive in digital format and relatively easy and inexpensive to collect, e. g., from public resources such as social media, the problem of efficiently gathering relevant annotations still needs to be overcome. One approach is *Active Learning* (AL) [Zhu, 2005], a type of algorithm that interactively queries the user to manually label certain data points. The core idea of AL is to extract the most informative instances from a pool of unlabelled data based on a specific query strategy [Settles, 2010]. These selected instances are then passed to human annotators for labelling and a model is derived from this subset. This approach significantly reduces the labelling effort.

The work by Zhang *et al.* [2015c] takes the idea of AL a step forward and combines it with *Semi-Supervised Learning* (SSL) techniques to efficiently share the labelling work between a human annotator and a machine: a pre-existing classifier is used to predict labels for the unlabelled data. For each of those predictions a confidence level is calculated by the classifier. Only if this level falls below a certain threshold a human annotator is asked to revise the annotation. To further save labelling efforts, one can apply *Dynamic Active Learning* (DAL) by choosing the most reliable raters first [Zhang *et al.*, 2015b]. Zhang *et al.* [2015a] developed an agreement-based annotation technique that dynamically determines how many human annotators are required to label a selected instance. The technique considers individual rater reliability and inter-rater agreement to decide on a combination of raters to be allocated to an instance.

However, little emphasis is given to the question of how to assist users in the application of these techniques for the creation of their own corpora. While the benefits of integrating active learning with annotation tasks has been demonstrated in a variety of experiments, annotation tools that provide users with access to active learning techniques are rare. Recent developments for audio, image and video annotation that make use of active learning include CAMOMILE [Poignant *et al.*, 2016] and iHEARu-PLAY [Hantke *et al.*, 2015]. However, systematic studies focusing on the potential benefits of the active learning approach within the annotation environment from a user's point of view have been performed only rarely [Cheng and Bernstein, 2015; Kim and Pardo, 2017].

Interactive Machine Learning (IML) [Fails and Olsen, 2003; Amershi *et al.*, 2014] aims to involve users actively in the creation of models for recognition tasks. Most approaches integrate automated data analysis and interactive visualisation tools in order to enable users to inspect data, process features and tune models. An example includes ModelUI [Wagner *et al.*, 2010]. It presents users with a graphical user interface that allows them to test different ML algorithms on labelled data. Labels are acquired by stimuli which may include textual instructions, but also images or videos. Afterwards, users can review the recordings and correct the annotations.

Rosenthal and Dey [2010] investigated which kind of information should be provided to users in order to reduce annotation errors. They found out that contextual information and predictions of the learning algorithms were in particular useful for the annotation of activity data. In contrast, uncertainty information had no effect on the accuracy of the labels, but just indicated to the labellers that classification was difficult. Amershi *et al.* [2009] investigated how to empower users to select samples for training by appropriate visualisation techniques. They found that a representative overview of best and worst matching examples is of higher value than a set of high-certainty images and conjecture that high-certainty images do not provide much information to the learning processing due to their similarity to already labelled images. In another paper by Amershi *et al.* [2015], the authors suggest an interactive visualization technique in order to assess a models' performance. By sorting samples according to their prediction score, the user can directly retrieve additional information and annotate them for better performance tracking. This way, the tool allows users to monitor the performance of individual samples while the model is iteratively retrained.

In addition to presenting the outcome of a classification in a structured way, one may aim at opening up what is usually perceived as a 'black box': the classifier itself. *Explainable Artificial Intelligence* (XAI) deals with the problem of making AI decisions transparent and explainable [Samek *et al.*, 2017]. For example, displaying the closest match to an instance can be a simple, yet effective way to increase transparency of the classification process. However, in complex AI systems where reasoning is no longer based on instance matching such an approach is not sufficient. A detailed discussion of different theories of explanation is given in [Sørmo *et al.*, 2005].

Today, explainability becomes increasingly important as we rely more and more on AI in our everyday life. For instance, before trusting a 'black box' model in a mission-critical applications, e. g., the diagnosis of Pneumonia, we have to ensure that the prediction is not based on random factors such as overfitting and spurious correlation [Caruana *et al.*, 2015]. However, gaining insights into the inner workings of a classifier may not just prevent misapplication, but also bears potential for improving the system. Or as noted by Samek *et al.* [2017]: "the first step towards improving an AI system is to understand its weaknesses".

Summing up, it may be said that multiple studies empirically investigated the potential of novel techniques in order to minimise human labelling effort. In addition, some studies

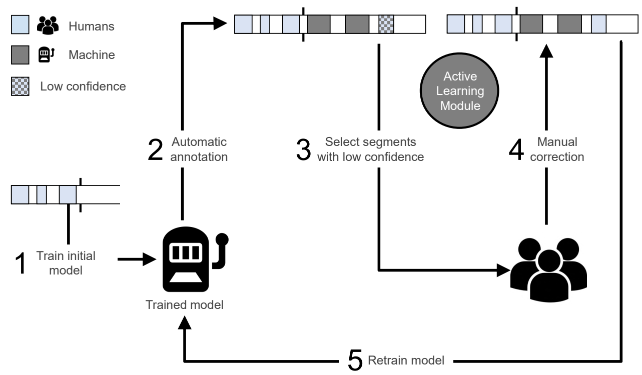

Figure 1: The scheme depicts the general idea behind Cooperative Machine Learning (CML): (1) An initial model is trained on partially labelled data. (2) The initial model is used to automatically predict unseen data. (3) Labels with a low confidence are selected and (4) manually revised. (5) The initial model is retrained with the revised data.

were conducted to actually label novel data, rather than test whether such method could save effort. Relatively little attention has been paid, however, to the question of how to make these techniques available to human labellers. There is a high demand for annotation tools that integrate ML techniques in order to reduce human effort – in particular in the area of social signal processing where human raters typically disagree on the labels [Lotfian and Busso, 2017].

3 Cooperative Machine Learning

In this paper, we subsume learning approaches that efficiently combine human intelligence with the machine's ability of rapid computation under the term *Cooperative Machine Learning* (CML) [Dong and Sun, 2003; Zhang *et al.*, 2015c]. In Figure 1, we illustrate our approach to CML, which creates a loop between a machine learned model and human annotators: an initial model is trained (1) and used to predict unseen data (2). An active learning module then decides which parts of the prediction are subject to manual revision by human annotators (3+4). Afterwards, the initial model is retrained using the revised data (5). Now the procedure is repeated until all data is annotated. By actively incorporating the user into the loop it becomes possible to interactively guide and improve the automatic predictions while simultaneously obtaining an intuition for the functionality of the classifier. In [Wagner *et al.*, 2018], we report an experiment that measures the increase in speed when the described CML strategy is applied within a realistic annotation task. Results showed that manual work was reduced by a factor of $\frac{5}{8}$.

However, the approach not only bears the potential to considerably cut down manual efforts, but also to come up with a better understanding of the capabilities of the classification system. For instance, the system may quickly learn to label some simple behaviours, which already facilitates the work load for human annotators at an early stage. Then, over time, it could learn to cope with more complex social signals as well, until at some point it is able to finish the task in a completely automatic manner. Such an iterative approach may even help bridging the gap between quantitative and qualitative coding, which still defines a great challenge in many fields in social science [Chen *et al.*, 2016].

To efficiently apply the described strategy, we would like to know the *sweet spot* for handing an annotation task over to the machine. On the one hand, if we do it too early, the model becomes unstable and predictions will be poor. On the other hand, if we annotate more data than necessary, we give away precious time. To avoid any of the described situations, we are interested in finding a good trade-off between machine performance and human effort. Unfortunately, we cannot easily guess what is the ideal moment to hand over the task to a machine. This is because the amount of training data that is required to build a robust model depends on a number of factors, such as the homogeneity of the data, the discrimination ability of the extracted features, the number of subjects and classes, and, not least, the complexity of the recognition problem. Alternatively, instead of trying to determine a sweet spot beforehand (and possibly miss it), we could iteratively test the applicability of the strategy and stop when the performance seems promising. Therefore, we opt to make the described strategy an integral part of a graphical interface (see Section 4). This allows annotators to visually examine the results at any time and to individually decide whether more labelling is required or not. This procedure can be further accelerated by providing visual feedback on the quality of the predictions. This way, annotators can concentrate on parts with *low confidence*, i. e., correcting only labels with a high uncertainty[2].

4 NOVA Tool

We will now introduce our novel annotation tool NOVA. The interface is inspired by existing software, such as EUDICO Linguistic Annotator (ELAN) [Wittenburg *et al.*, 2006] and Annotation of Video and Language (ANVIL) [Kipp, 2013], which offer layer-based tiers to insert time-anchored labelled segments – that is *discrete* annotations. In addition, NOVA also supports *continuous* annotations, which allow an observer to track the content of an audiovisual stimulus over time along one ore more dimensions – a feature inspired by software like GTRACE (General Trace) [Cowie *et al.*, 2012], CARMA (Continuous Affect Rating and Media Annotation) [Girard, 2014] and DARMA (Dual Axis Rating and Media Annotation) [Girard and Wright, 2016]. However, whereas the mentioned tools offer none or only little automation, NOVA has been advanced with features to create semi-automated annotations (see Section 3).

NOVA is open-source and available on Github: `https://github.com/hcmlab/nova`.

4.1 Main Interface

The NOVA user interface has been designed with a special focus on the annotation of long and continuous recordings involving multiple modalities and subjects. A screenshot of a loaded recording session is shown in Figure 2. On the top,

[2]Uncertainty can be derived from the distance a predicted sample has to the decision boundaries of the other classes. In regression problems, dropout can be used [Gal and Ghahramani, 2016]).

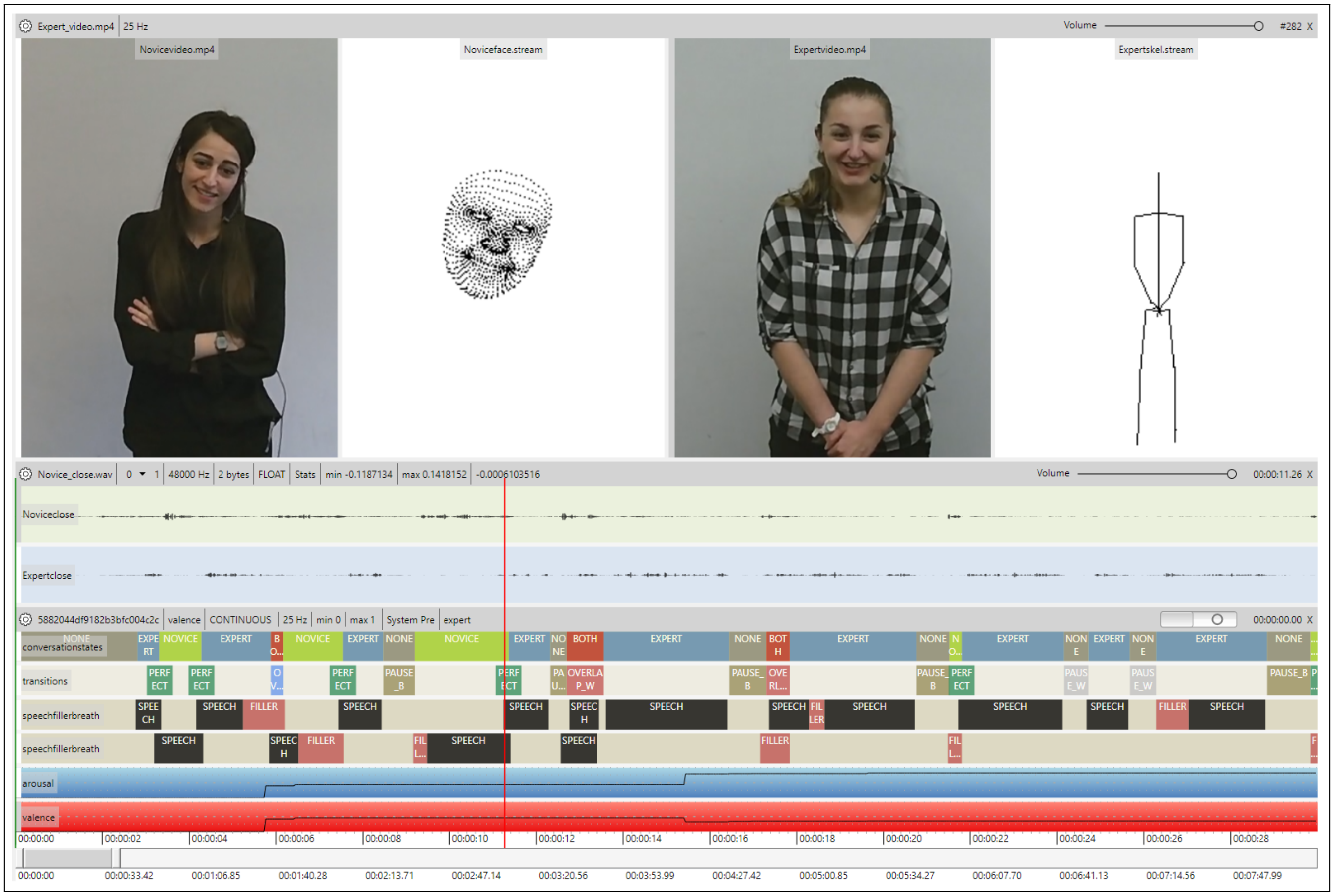

Figure 2: NOVA allows to visualise various media and signal types and supports different annotation schemes. From top downwards: full-body videos along with skeleton and face tracking, and audio streams of two persons during an interaction. In the lower part, several discrete and continuous annotation tiers are displayed. Annotations can be edited on a static fraction of the recording or interactively during playback.

several media tracks are visualised and ready for playback. Note that the number of tracks that can be displayed at the same time is not limited and various types of signals (video, audio, facial features, skeleton, depth images, etc.) are supported. In the lower part, we see multiple annotation tracks of different types (discrete, continuous and transcriptions) describing the visualised content.

To support a collaborative annotation process, NOVA maintains a database back-end, which allows users to load and save annotations from and to a MongoDB[3] running on a central server. This gives annotators the possibility to immediately commit changes and follow the annotation progress of others. Beside human annotators, a database may also be visited by one or more "machine users". Just like a human operator, they can create and access annotations. Hence, the database also functions as a mediator between human and machine. NOVA provides instruments to create and populate a database from scratch. At any time new annotators, schemes and additional sessions can be added.

NOVA provides several functions to process the annotations created by multiple human or machine annotators. For instance, statistical measures such as Cronbach's α or Cohen's κ can be applied to identify inter-rater agreement. Finally, multiple annotations can be merged to a Gold Standard. However, in the following we will concentrate on another feature of NOVA: the use of machine learning to support the user during the annotation process.

4.2 Machine Learning

For best possible performance, tasks related to machine learning (ML) are outsourced and executed in a background process. As backend we use our open-source Social Signal Interpretation (SSI) framework[4]. Since SSI is primarily designed to build online recognition systems, a trained model can be directly used to detect social cues in real-time [Wagner *et al.*, 2013].

A typical ML pipeline starts by preprocessing data to input data for the learning algorithm, a step known as *feature extraction*. An XML template structure is used to define extraction chains from individual SSI components. For instance, the following template extracts the commonly used Mel-frequency cepstral coefficients (MFCCs) from audio sig-

[3] https://www.mongodb.com/

[4] http://openssi.net

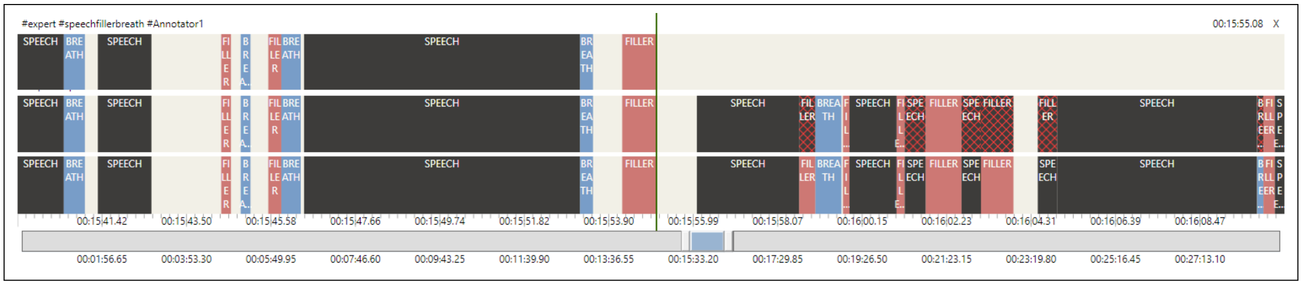

Figure 3: The upper tier shows a partly finished annotation. Machine learning is now used to predict the remaining part of the tier (middle), where segments with a low confidence are highlighted with a red pattern. The lower tier shows the final annotation after manual revision.

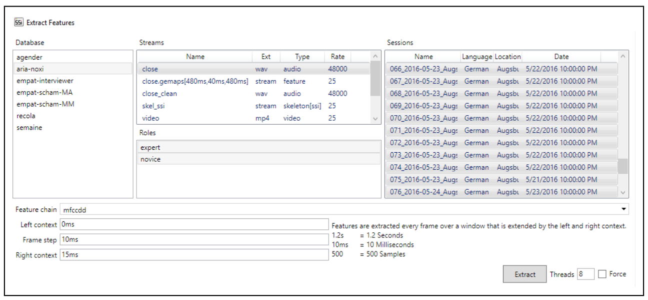

Figure 4: Feature extraction dialogue.

nals. It feeds the input signal through a pre-emphasis filter and afterwards extracts the features over a sliding window of 25 ms with a frame step of 10 ms.

```
<chain>
  <register name="audio"/> <!--load components-->
  <meta frameStep="10ms" rightContext="15ms"/>
  <filter> <!--apply filtering-->
    <item create="PreEmphasis"/>
  </filter>
  <feature> <!--extract features-->
    <item create="Mfcc" option="mfcc"/>
  </feature>
</chain>
```

A dialogue helps users to extract features by selecting an input stream and a number of sessions (see Figure 4). The result of the operation is stored as a new signal in the database. This way, feature streams can be reviewed in the GUI and accessed by all users. Based on the extracted features, a classifier can be trained. Again, templates are used to define classification schemes, e. g., :

```
<trainer>
  <register name="model"/>
  <meta balance="under"/> <!--apply under sampling-->
  <normalize>
    <item method="Scale"/> <!--scale the features-->
  </normalize> <!--Support Vector Machine-->
  <model create="SVM" option="svm"/>
</trainer>
```

To automatically finish an annotation, the user either selects a previously trained model or temporarily builds one using the labels on the current tier. An example before and after the completion is shown in Figure 3. Note that labels with a low confidence are highlighted with a pattern. This way, the annotator can immediately see how well the prediction worked. He or she can now either revert the operation or continue based on the automated generated annotation. At any time, usually after correcting a couple of false predictions or adding some missing labels, the procedure can be repeated. Over time, this should lead to increasingly stable predictions.

Summing up, the described methodology offers transparency from two directions. By observing the output of the classifier, the user can assess its performance and also trace how it changes with new input. In addition, visualising the input to the classifier (raw media or feature streams) can provide hints why a prediction was successful in one place but failed in another. For instance, the user may find out that predictions were wrong due to failure of the tracking algorithm. This way, users also learn in which situations they can trust the model.

Note that users can extend NOVA's ML tools by simply adding new templates. SSI supports a variety of features sets for different types of signals. For instance, it allows to extract a large number of audio parameters based on the widely used OPENSMILE toolkit [Eyben *et al.*, 2013]. For the computation of facial points and action units from video streams, the OPENFACE tool by Baltrušaitis *et al.* [2016] has been integrated. In terms of classification models, SSI supports (among others) Google's neural network framework TENSORFLOW[5] or the popular THEANO[6] library.

5 Conclusion

The goal of the presented work is to foster the application of *Cooperative Machine Learning* (CML) strategies to support the annotation of social signals in large multi-modal databases. Well described corpora that are rich of human behaviour are needed in a number of disciplines, such as Social Signal Processing and Behavioural Psychology. However, populating captured user data with adequate descriptions can be an extremely exhausting and time-consuming task. To this end, we have presented a novel annotation tool NOVA. It allows to distribute annotation tasks among multiple human raters and offers an interface to ML algorithms for semi-automated annotation.

The core idea of the presented work is to create a loop, in which humans start solving a task (here labelling social signals), and over time, a machine learns to automatically com-

[5]https://www.tensorflow.org/

[6]https://github.com/Theano/

plete the task. In conventional approaches, this involves at least two parties: an end-user, who has knowledge about the domain, and a machine learning practitioner, who can cope with the learning system. However, to make the process more rapid and focused, our tool combines a traditional annotation interface with techniques for automated labelling that can be applied out of the box requiring no knowledge on ML. For an optimised workflow, coders have the possibility to individually decide when and how to use them in the labelling process. Further, to assess the reliability of automatic predictions immediate visual feedback is provided, which gives annotators the chance to gain insights into the ML model and adapt their strategies at times. By interactively guiding and improving automatic predictions, an efficient integration of human expert knowledge and rapid mechanical computation is achieved.

Our experiences with NOVA show that CML strategies not only have the potential to speed up coding, but can also have a positive influence coding quality. Because of the preciseness machine-aided techniques introduce into the coding process, the level-of-detail is improved while at the same time human efforts are reduced. However, while a machine is able to annotate social signals much faster and more consistently than humans can do, human raters still bring a better understanding for the application in which the models to be trained will eventually be applied. Furthermore, human raters do not just look at the behaviours to be labelled, but also reason about the context in which they occur [Baur *et al.*, 2017]. Being presented with the results of an automated labelling process might influence human labellers in a positive manner. Nevertheless, one should be aware of the risk that a machine-like style of annotation might not always result in better systems. This is in particular true when social signals are analysed where raters usually disagree on the labels and no objective ground truth can be established. In order to benefit from the complementary skills of machines and human raters, annotation tools like NOVA are needed that aim for a smooth integration of human intelligence and resources.

In the future, we aim at further improving the explanation capabilities of the system by providing more information about the inner workings of the classifiers. This, for instance, could be achieved by adopting explanation approaches like the LIME-System of Ribeiro *et al.* [2016] or the Explicable-Boundary-Tree-Explainer by Wu *et al.* [2018]. The idea here is to not only visualize final predictions, but also disclose what has lead to a specific decision. We believe that this way, human resources could be applied even more effectively, which may further shorten the time it takes to achieve a stable classification performance.

Acknowledgements

This work is Funded by European Union Horizon 2020 research and innovation programme, grant agreement No. 645378 (ARIA-Valuspa).

References

Saleema Amershi, James Fogarty, Ashish Kapoor, and Desney S. Tan. Overview based example selection in end user interactive concept learning. In *Proceedings of the 22nd Annual ACM Symposium on User Interface Software and Technology, Victoria, BC, Canada, October 4-7, 2009*, pages 247–256, 2009.

Saleema Amershi, Maya Cakmak, W. Bradley Knox, and Todd Kulesza. Power to the people: The role of humans in interactive machine learning. *AI Magazine*, 35(4):105–120, 2014.

Saleema Amershi, Max Chickering, Steven M. Drucker, Bongshin Lee, Patrice Simard, and Jina Suh. Modeltracker: Redesigning performance analysis tools for machine learning. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 337–346. ACM, 2015.

Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–10. IEEE, 2016.

Tobias Baur, Dominik Schiller, and Elisabeth André. Modeling user's social attitude in a conversational system. In Marko Tkalcic, Berardina De Carolis, Marco de Gemmis, Ante Odic, and Andrej Kosir, editors, *Emotions and Personality in Personalized Services - Models, Evaluation and Applications*, Human-Computer Interaction Series, pages 181–199. Springer, 2017.

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 1721–1730, New York, NY, USA, 2015. ACM.

Nan-Chen Chen, Rafal Kocielnik, Margaret Drouhard, Vanessa Peña-Araya, Jina Suh, Keting Cen, Xiangyi Zheng, and Cecilia R. Aragon. Challenges of applying machine learning to qualitative coding. In *CHI 2016 workshop on Human Centred Machine Learning*, 2016.

Justin Cheng and Michael S. Bernstein. Flock: Hybrid crowd-machine learning classifiers. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 600–611. ACM, 2015.

Roddy Cowie, Gary McKeown, and Ellen Douglas-Cowie. Tracing emotion: An overview. *International Journal of Synthetic Emotions (IJSE)*, 3(1):1–17, January 2012.

Miaobo Dong and Zengqi Sun. On human machine cooperative learning control. In *Proceedings of the 2003 IEEE International Symposium on Intelligent Control*, pages 81–86, Oct 2003.

Ellen Douglas-Cowie, Nick Campbell, Roddy Cowie, and Peter Roach. Emotional speech: Towards a new generation of databases. *Speech Communication*, 40(c):33–60, 2003.

Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings* *of the 21st ACM International Conference on Multimedia*, MM '13, pages 835–838, New York, NY, USA, 2013. ACM.

Jerry Alan Fails and Dan R. Olsen, Jr. Interactive machine learning. In *Proceedings of the 8th International Conference on Intelligent User Interfaces*, IUI '03, pages 39–45, New York, NY, USA, 2003. ACM.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 1050–1059, 2016.

Jeffrey M. Girard and Aidan G C Wright. DARMA: Dual Axis Rating and Media Annotation. *Manuscript submitted for publication*, 2016.

Jeffrey M. Girard. Carma: Software for continuous affect rating and media annotation. *Journal of Open Research Software*, 2(1):e5, 2014.

Simone Hantke, Florian Eyben, Tobias Appel, and Björn Schuller. ihearu-play: Introducing a game for crowd-sourced data collection for affective computing. In *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*, pages 891–897. IEEE, 2015.

Bongjun Kim and Bryan Pardo. I-sed: An interactive sound event detector. In *Proceedings of the 22Nd International Conference on Intelligent User Interfaces*, IUI '17, pages 553–557, New York, NY, USA, 2017. ACM.

Michael Kipp. Anvil: The video annotation research tool. In *Handbook of Corpus Phonology*. Oxford University Press, Oxford, UK, 2013.

Reza Lotfian and Carlos Busso. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, PP(99):1–1, 2017.

Johann Poignant, Mateusz Budnik, Hervé Bredin, et al. The CAMOMILE collaborative annotation platform for multi-modal, multi-lingual and multi-media documents. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016.*, 2016.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016.

Stephanie Rosenthal and Anind K. Dey. Towards maximizing the accuracy of human-labeled sensor data. In *Proceedings of the 2010 International Conference on Intelligent User Interfaces, February 7-10, 2010, Hong Kong, China*, pages 259–268, 2010.

Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *CoRR*, abs/1708.08296, 2017.

Burr Settles. Active learning literature survey. 52(55–66):11 pages, 2010.

Frode Sørmo, Jörg Cassens, and Agnar Aamodt. Explanation in case-based reasoning–perspectives and goals. *Artificial Intelligence Review*, 24(2):109–143, Oct 2005.

Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. Social signal processing: Survey of an emerging domain. *Image Vision Computing*, 27(12):1743–1759, November 2009.

Johannes Wagner, Elisabeth André, Michael Kugler, and Daniel Leberle. SSI/ModelUI - A tool for the acquisition and annotation of human generated signals. In *DAGA 2010*, Berlin, Germany, 2010. TU Berlin.

Johannes Wagner, Florian Lingenfelser, Tobias Baur, Ionut Damian, Felix Kistler, and Elisabeth André. The social signal interpretation (ssi) framework: multimodal signal processing and recognition in real-time. In *Proceedings of the 21st ACM international conference on Multimedia*, MM '13, pages 831–834, New York, NY, USA, 2013. ACM.

Johannes Wagner, Tobias Baur, Yue Zhang, Michel F. Valstar, Björn W. Schuller, and Elisabeth André. Applying cooperative machine learning to speed up the annotation of social signals in large multi-modal corpora. *CoRR*, abs/1802.02565, 2018.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. Elan: A professional framework for multimodality research. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, pages 879–896, 2006.

Huijun Wu, Chen Wang, Jie Yin, Kai Lu, and Liming Zhu. Sharing deep neural network models with interpretation. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 177–186. International World Wide Web Conferences Steering Committee, 2018.

Yue Zhang, Eduardo Coutinho, Björn Schuller, Zixing Zhang, and Michael Adam. On rater reliability and agreement based dynamic active learning. In *International Conference on Affective Computing and Intelligent Interaction, ACII*, pages 70–76, Xi'an, China, September 2015.

Yue Zhang, Eduardo Coutinho, Zixing Zhang, Caijiao Quan, and Björn Schuller. Agreement-based dynamic active learning with least and medium certainty query strategy. In A Krishnamurthy, A Ramdas, N Balcan, and A Singh, editors, *Proceedings of the 32nd International Conference on Machine Learning (ICML 2015). JMLR W&CP volume 37*, pages 1–5, Lille, France, 2015. Lille, France.

Zixing Zhang, Eduardo Coutinho, Jun Deng, and Björn Schuller. Cooperative learning and its application to emotion recognition from speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(1):115–126, Jan 2015.

Xiaojin Zhu. Semi-supervised learning literature survey. Technical report, Computer Sciences, University of Wisconsin-Madison, 2005.