Automatic detection of children's engagement using non-verbal features and ordinal learning

Jaebok Kim, Khiet P. Truong, Vanessa Evers

Human Media Interaction, University of Twente, The Netherlands
{j.kim, k.p.truong, v.evers}@utwente.nl

Abstract

In collaborative play, young children can exhibit different types of engagement. Some children are engaged with other children in the play activity while others are just looking. In this study, we investigated methods to automatically detect the children's levels of engagement in play settings using non-verbal vocal features. Rather than labelling the level of engagement in an absolute manner, as has frequently been done in previous related studies, we designed an annotation scheme that takes the order of children's engagement levels into account. Taking full advantage of the ordinal annotations, we explored the use of SVM-based ordinal learning, i.e. ordinal regression and ranking, and compared these to a rule-based ranking and a classification method. We found promising performances for the ordinal methods. Particularly, the ranking method demonstrated the most robust performance against the large variation of children and their interactions.

Index Terms: children, engagement, ranking, non-verbal

1. Introduction

As a result of emerging studies on social signal processing, it has become feasible to develop social robots facilitating social interaction between people (e.g. collaboration and engagement) [1, 2], in particular, between children [3, 4]. In small group play setting, it was observed that children exhibit different levels and types of involvement, i.e. engagement [5, 6, 7]. In order to develop a social robot that interacts with groups of children and that can anticipate to a child's level of engagement, we aim to study ways to automatically detect a child's level of engagement in small group play settings.

Modelling of children's social behaviours such as engagement and dominance has been extensively studied in the field of Human Robot Interaction (HRI) [4, 8] but their detection methods relied on hand-coded features. More importantly, in many studies, the cues or status of the child are not considered with respect to the other group participants but rather individually, despite the observation that group participants adjust their behaviours to each other [9, 10].

In this paper, we aim to develop the automatic detection of engagement in groups of children in playful settings which has remained unexplored so far. In contrast to previous related studies, we argue that it is necessary to model a child's level of engagement relatively with respect to the other children in the group. We propose an annotation scheme that takes "relative levels" into account and that can be used in ordinal learning methods. In addition, our feature extraction focuses on vocalic non-verbal cues, which are strongly associated with social behaviours e.g. engagement and dominance [11, 12, 13]. Lastly, our method is based on a fine resolution to capture temporal dynamics of engagement [10].

This paper is structured as follows. In section 2, related works are presented. We explain our audiovisual corpus and annotation scheme in section 3. We define our method and features in section 4. In section 5, results of experiments are presented, and conclusions are addressed in section 6.

2. Related Work

Automatic analysis of children's individual (e.g., [14]) and group behaviours have been studied before [4, 3], but these studies are limited to the detection of engagement and often rely on manual annotations. In addition, many of the related works [4, 8] addressed classification of *absolute* engagement or dominance qualifications such as "high" and "low" and did not take the behaviours of the other group members into account. However, it is known that social behaviours largely rely on participants and specific interactions or situations [9]. For instance, [10] found variability of temporal engagement depending on the group. We believe that engagement models should be able to express relative differences at specific moments and identify individuals who are more or less engaged with respect to the other members at certain moments in time. The annotation scheme and learning methods adopted in our current study reflect these properties and will be elaborated on in the following sections.

http://dx.doi.org/10.21437/WOCCI.2016-5

3. Data

3.1. Corpus of groups of children in a playful setting

For our analyses, we used a corpus containing audiovisual recordings of groups of children playing with each other [13]. In this corpus, a 3D puzzle playful activity was used to facilitate children's spontaneous social behaviours. Using cube blocks, we asked children to build given shapes of animals together. Dutch children aged 5 - 8 (6.95 ± .95) were recruited from a primary school. They were first clustered according to age and then assigned randomly to a group of three for each session. Eight out of ten sessions were considered in our analyses (two sessions were discarded due to technical malfunctions of recording). More details about this corpus can be found in [13].

3.2. Annotation

In our task, we employed the concept of engagement incorporating with verbal and non-verbal behaviours supporting the perception of connection between participants [3]. In our task, different levels of engagement are identified by observing verbal and non-verbal exchanges of attention among a group of children. During pilot coding sessions, we provided two coders with only the description of engagement and videos. The coders were asked to label levels of engagement in an absolute manner, i.e. {low, medium, high}, it became clear that annotators had difficulty of labelling these classes, resulting in poor inter-rater agreement (kappa) between two coders (.57). Hence, we devised an annotation scheme in consideration of the relative levels of engagement children exhibit using the following ordinal descriptions (from low to high level of engagement):

1 giving relatively less attention to others and getting relatively less attention from others.

2a giving relatively less attention to others but getting attention from others.

2b giving attention to others but getting relatively less attention from others.

3 giving attention to others and getting attention from others.

In this way, children in a group can be ordered from a low to a high level of engagement. For our analyses, the classes: {2a} and {2b} were equally ranked (in level of engagement) and merged into one class {2}. Moreover, if any differences could not be observed among the three children, ties were allowed (e.g. {1, 1, 1}, {3, 3, 3}). In order to annotate all the recordings, a proper size for an annotation segment needed to be determined. In previous work [4, 15, 12], segments of between 0.5s and 5min long were used to predict engagement, roles, and dominance.

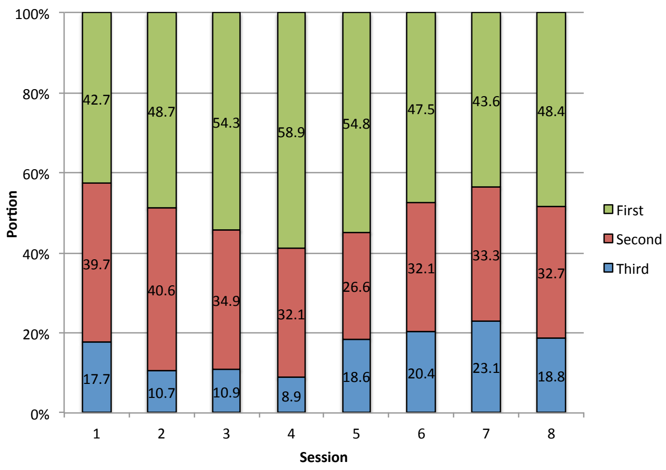

Figure 1: *Proportions of segments that each child is ranked as the one with highest level of engagement. Green=child with largest proportion, red=child with 2nd largest proportion, blue=child with 3rd largest proportion.*

Through several pilot coding sessions, we concluded on an empirical basis that 5s-long segments were suitable for the annotators to observe various levels of engagement. Again, without having access to any individual speech recordings, only the videos and descriptions were given to two annotators to code each child for level every 5s using ELAN [16]. This labelling resulted in a total of 1571 segments and the average inter-rater agreement was 0.82 (kappa). Finally, 1171 segments agreed upon by all annotators were used in subsequent analyses.

Results of the annotations show, as expected, that each group of children displays different patterns of engagement, see Fig. 1. We can observe that different groups of children indeed display different behaviours of engagement which warrants our motivation to model children's behaviours with respect to others in a group.

4. Method

4.1. Approach

The detection of ordinal engagement levels can be formulated as classification, ordinal regression and ranking tasks. While classification does not take orders into account, "ordinal regression" and "ranking" assign classes in orders. More specifically, "ordinal regression" assigns samples into ordinal intervals and predicts values of ordinal categories while "ranking" predicts orders between instances [17]. One of simple forms of ranking is rule-based ranking. Although one of participants in a group might be less active in engagement than the others even if the level of speaking activity is high [18], the high level of speech activity could be still one of indicators of engagement [19]. Based on this, a rule-based ranking will be our baseline **BASE**. On the other hand, we utilise SVM-based learning methods which are widely applied to automatic detection of social behaviour [20]. In particular, SVM ranking was successfully adopted in various fields e.g. document retrieval and speech emotion recognition [21, 22]. In summary, we will compare different methods: rule-based ranking (**BASE**), SVM classification (**CLASSIFY**), and SVM-based ordinal methods: ordinal regression (**REG**) and ranking (**RANK**). Particularly, SVM ranking compares only instances in certain restrictions, which might be effective for our tasks where children are compared at every segment not in an entire session. We assume that SVM ranking will result in the best performance among all methods mainly due to this property. To give more understanding, we explain SVM raking in the following section.

4.2. SVM ranking

In order to learn an order of engagement between children, we compare only feature vectors of two children in the same constraint, q, which is the period of time in our corpus. Let us denote feature vectors of two children: (x_i, x_j) where i and j are indices of each child. Note that pairwise feature vectors $(x_i - x_j)$ are generated only in the same period. In other words, SVM ranking does not compare features with different periods. Next, learning ranks can be designed as finding a function f that generalises beyond the instances for query k as follows:

$$(x_i, x_j) \in f_{\overrightarrow{w}}(q_k) \iff \overrightarrow{w}\Phi(q_k, x_i) > \overrightarrow{w}\Phi(q_k, x_j) \tag{1}$$

where $\overrightarrow{w}$ is a weight vector learned by training and Φ is a kernel function. Finding f is equivalent to finding $\overrightarrow{w}$ that maximizes the number of the following inequalities:

$$\begin{aligned} \forall(x_i, x_j) \in r_1 &: \overrightarrow{w}\Phi(q_1, x_i) > \overrightarrow{w}\Phi(q_1, x_j) \\ &\quad \dots \\ \forall(x_i, x_j) \in r_n &: \overrightarrow{w}\Phi(q_n, x_i) > \overrightarrow{w}\Phi(q_n, x_j) \end{aligned} \tag{2}$$

where n is the total number of queries and r is the rank of the corresponding query. Then, we can approximate this generalisation by introducing slack variables $\xi_{i,j,k}$ and minimising the upper bound $\sum \xi_{i,j,k}$. Finally, we can transform it to classification on pairwise difference vector $\Phi(q_k, x_i) - \Phi(q_k, x_j)$, so the optimisation becomes equivalent to:

$$\overrightarrow{w}(\Phi(q_k, x_i) - \Phi(q_k, x_j)) \geq 1 - \xi_{i,j,k} \tag{3}$$

In order to rank children using the trained model, we simply sort children by their scores i.e. output of function f.

4.3. Features

Data representation is often considered a key for successful machine learning applications [23]. In this paper, we use non-verbal vocal features and acoustic features based on previous works [11, 12, 13]. We carefully selected our features and categorised them into individual non-verbal, relational non-verbal, and acoustic features as summarised in Table 1.

Category	Features	Functions
individual (8)	speech (4), self-silence(4)	mean-duration SD-duration total-duration total-count
relational (20)	speaker change (4), speaker change with overlaps (4), successful interruptions (4), unsuccessful interruptions (4), overlaps (4)	
acoustic (36)	F0 (6), energy (6), ZCR (6), HNR (6), jitter (6), shimmer (6)	mean SD

Table 1: *Feature sets (number of features by functions) for each child*

All of these features are extracted from every 5s-long segment using each child's voice stream [13]. First of all, we extract each child's speech segment using voice activity detection from each voice stream. Then, in order to correct errors caused by noise and channel-inferences, we applied iterative automatic speaker identification and manual correction. In a similar ways as [24], we used MFCC features and Gaussian-Mixture-Model (GMM) to detect segments of different speakers. Next, all non-verbal features and other acoustic features are extracted on the segments. For acoustic features, we select F0, energy, HNR, ZCR, jitter, and shimmer and their $\{\Delta, \Delta\Delta\}$ by using openSMILE [25] as a representative set to classify social behaviours [20]. Then, statistical functions and normalisation are applied. For individual and relational features, mean-duration, standard-deviation (SD) of duration, total-duration, and total-count are applied. For acoustic features, we simply obtained mean and SD values of the features. Lastly, all values are scaled into the range {0.0 - 1.0} over the entire corpus. All features are combined at feature-level. Although a more in-depth analysis and comparison of features would be interesting, we consider this out of scope for the current paper and leave this for future work.

Note that SVM ranking requires multiple feature vectors to predict an order of their instances. Therefore, it has feature constraints to compare instances within the same group. As mentioned previously, the constraint is period of time, i.e. the segment index representing "given moments" in the range of [0, the total number of segments for each child]. Therefore, it generates pairwise feature vectors at every segment.

5. Experiments and Results

5.1. Setup

In this section, we present results from our experiments by using rule-based ranking (**BASE**) [12], SVM classification (**CLASSIFY**), SVM ordinal regression (**REG**), and SVM ranking (**RANK**). Based on [19], we composed a rule that more speech features (e.g. speech, speaker-changes, and overlaps) indicates a higher level of engagement for **BASE**. For reproduction purposes, we used the implementation of libsvm and their extensions [26, 27]. Parameters of each model were tuned by a grid search for the optimal performances. For the evaluation, we used the normalised tau distance, which is one of the evaluation methods for detection of ordinal categories [28]. If we calculate it for two different lists (e.g. X_1 and X_2), it is defined as follows:

$$K(X_1, X_2) = \frac{D}{N(N-1)/2} \tag{4}$$

where D is the total number of discordant or swapped pairs and N is the total number of elements in a list. If all orders are wrong, then it becomes 1.0 while it indicates 0.0 for all correct orders. Since neither **REG** nor **RANK** predicts "ties", we do not evaluate "ties" in the current study. In order to test the statistical significance of differences between the methods, we employed a paired corrected t-test [29] (p-values are separately provided with the results). For validating robustness of each method, we conducted two types of experiments: session-independent and session-dependent experiments.

In session-independent experiments, we look into the overall performances of each method using Leave-One-Session-Out-Cross-Validation, resulting in session-independent models (**SI**). In each fold, one session is used for validation and all other sessions are used for training. Since we have 3 children's samples per segment, a total of 3571 samples were used and the average number of test samples is 438 and that of training is 3072.

In session-dependent experiments, we study the effectiveness of session-dependent models (**SD**) and their intra-session variations. As a large intra-session variation of engagement was reported in [10], we expect each child to present large intra-session variation of the levels and cues. To validate the robustness of each method, we randomly ordered our data and applied 10 fold cross-validation for each session (totalling 10 x 8 = 80 experiments for each method). Note that **BASE** does not distinguish **SD** from **SI** since it does not use any statistical learning.

5.2. Results and discussion

First of all, we examined the overall performances (average over sessions) of the methods. Fig. 2 summarises all results of the session-independent models (**SI**) and session-dependent models (**SD**). **RANK** and **REG** outperformed **BASE** and **CLASSIFY** significantly ($p < .0001$) while **CLASSIFY** performed worse than **BASE**. The main reason of the low performance of **CLASSIFY** is that it does not model differences between instances at a given condition or moment.

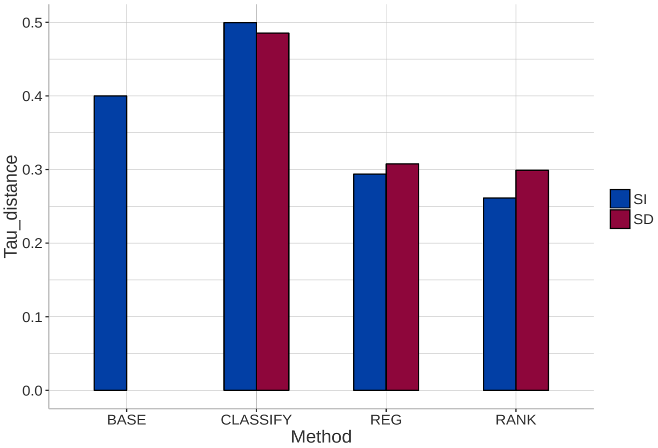

Figure 2: *Overall performances of the methods.*

Moreover, **RANK** reduced the error rate of **BASE** by nearly 14%. As **BASE** did not show any promising performance, the rule assuming more speech activity higher engagement turned out to be unreliable in engagement detection and supports the previous findings in [18]. Furthermore, **RANK** was slightly superior to **REG** ($p < .05$). We assume that it was mainly due to feature-constraints of **RANK** while **REG** does not have such a constraint. These findings support our assumption that models reflecting orders will be more suited for detection of engagement level in a relative manner.

In addition, if a model is robust against the variation of children, the performance of **SI** should be as good as that of **SD**. We can observe in Fig. 2 that **SD** slightly outperformed **SI** in only the case of **CLASSIFY** while **SI** outperformed **SD** in all other cases: **REG** and **RANK**. Since gaps of the performances between **SD** and **SI** are not statistically significant ($p < .1$), our findings are not conclusive. However, at least, **SI** of **RANK** showed the best performance among all and the robustness against the variation of sessions, and this is because it is able to capture characteristics of interactions at only given moments and generalise them into the model. In the following sections, we present results of session-independent and dependent experiments and investigate variations of performances of the methods in order to see the robustness.

5.2.1. Session-independent experiments

Fig. 3 shows the performances of each session separately in **SI**. First of all, we can observe that large variations over sessions exist regardless of the method. Particularly, **CLASSIFY** indicated the largest variation (.01) while it showed the worst performance. On the other hand, **RANK** showed the lowest variation (.002) compared to **REG** (.004). We found variability of performances depending on the session. For example, session

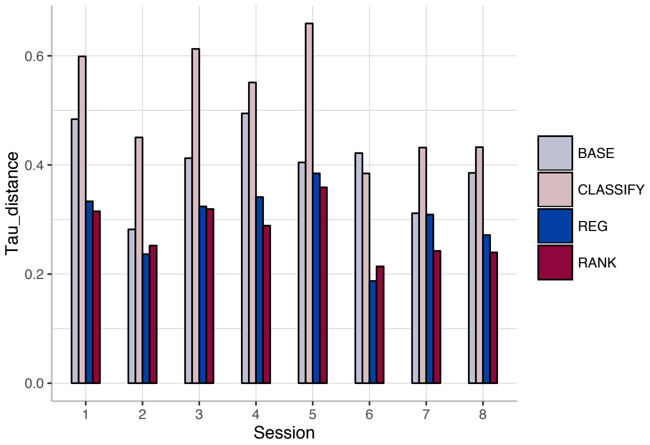

Figure 3: *Performances of **SI** models over sessions.*

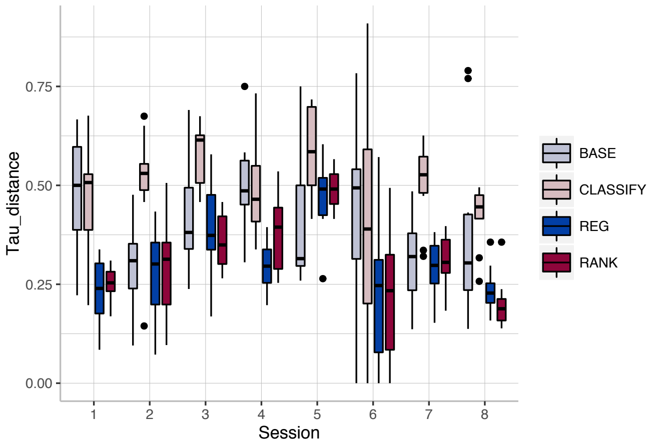

Figure 4: *Box plots displaying distribution of performances of **SD** models.*

6 showed the best performance while session 5 showed the worst regardless of its method. However, SVM ranking modelling social behaviours as a group and considering relative differences resulted in the best performance compared to the individual classification methods mainly used in [4, 8].

5.2.2. Session dependent experiments

In this section, we discuss the performances of **SD** models and their distribution. Figure 4 contains box plots that describe the distribution of the performances. Note that each session has 10-fold cross-validation and we calculated variation of the performance in each session based on the 10 folds. **RANK** and **REG** seem to show narrower distributions of the performance compared to **CLASSIFY**: the average of intra-session variations are .0285, .0233, .0168, .0159 for **BASE**, **CLASSIFY**, **REG**, and **RANK**, respectively. In other words, **RANK** and **REG** achieved the most stable performance in each session. Furthermore, we found that each session (group) showed a different variation of performance in a similar as the **SI** models did. As studied in [10, 9], children might interact in variant ways at different period of time. In this sense, **RANK** was more robust against variant interactions compared to the other methods.

6. Conclusions

In this study, we explored automatic ordinal detection of engagement between three children in collaborative play settings using non-verbal vocal cues. We showed that levels of engagement can be characterised by differences between children and that children's individual levels of engagement can be detected through ordinal learning methods. Using non-verbal vocal features, we compared rule-based ranking, SVM classification, SVM ordinal regression and SVM ranking methods. We found that SVM-based ordinal learning, ordinal regression and ranking, demonstrated promising results in comparison to the other methods studied. In particular, SVM ranking showed the best performance and the lowest inter and intra-session variation. In other words, SVM ranking was the most robust against variation of children and interactions at a given period of time. As future work, we will conduct more in-depth feature analyses and selection. Moreover, we will investigate how other aspects such as gender, personality traits, relation to other children, could be taken into account to advance our model.

7. Acknowledgements

The research leading to these results was supported by the European Community's 7th Framework Programme under Grant agreement 610532 (SQUIRREL - Clearing Clutter Bit by Bit).

8. References

[1] B. Robins, K. Dautenhahn, R. Te Boekhorst, and A. Billard, "Robotic assistants in therapy and education of children with autism: can a small humanoid robot help encourage social interaction skills?" *Universal Access in the Information Society*, vol. 4, no. 2, pp. 105–120, 2005.

[2] Y. Matsuyama, I. Akiba, S. Fujie, and T. Kobayashi, "Four-participant group conversation: A facilitation robot controlling engagement density as the fourth participant," *Computer Speech & Language*, vol. 33, no. 1, pp. 1–24, 2015.

[3] C. L. Sidner, C. Lee, C. D. Kidd, N. Lesh, and C. Rich, "Explorations in engagement for humans and robots," *Artificial Intelligence*, vol. 166, no. 1, pp. 140–164, 2005.

[4] I. Leite, M. McCoy, D. Ullman, N. Salomons, and B. Scassellati, "Comparing models of disengagement in individual and group interactions," in *Proceedings of the 10th Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2015, pp. 99–105.

[5] J. Piaget, *The psychology of the child*. New York: Basic Books, 1972.

[6] C. Stangor, *Social groups in action and interaction*. Psychology Press, 2004.

[7] M. B. Parten, "Social participation among pre-school children." *The Journal of Abnormal and Social Psychology*, vol. 27, no. 3, p. 243, 1932.

[8] S. Strohkorb, I. Leite, N. Warren, and B. Scassellati, "Classification of children's social dominance in group interactions with robots," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ser. ICMI '15. New York, NY, USA: ACM, 2015, pp. 227–234. [Online]. Available: http://doi.acm.org/10.1145/2818346.2820735

[9] J. A. Hall, E. J. Coats, and L. S. LeBeau, "Nonverbal behavior and the vertical dimension of social relations: a meta-analysis." *Psychological bulletin*, vol. 131, no. 6, p. 898, 2005.

[10] S. Al Moubayed and J. Lehman, "Toward better understanding of engagement in multiparty spoken interaction with children," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 211–218.

[11] M. Heldner and J. Edlund, "Pauses, gaps and overlaps in conversations," *Journal of Phonetics*, vol. 38, no. 4, pp. 555–568, 2010.

[12] D. B. Jayagopi, S. Ba, J.-M. Odobez, and D. Gatica-Perez, "Predicting two facets of social verticality in meetings from five-minute time slices and nonverbal cues," in *Proceedings of the 10th international conference on Multimodal interfaces*. ACM, 2008, pp. 45–52.

[13] J. Kim, T. K. P., V. Charisi, C. Zaga, M. Lohse, D. Heylen, and V. Evers, "Vocal turn-taking patterns in groups of children performing collaborative tasks: an exploratory study," in *Proceedings of the INTERSPEECH*, 2015, pp. 1645–1649.

[14] R. Gupta, C.-c. Lee, S. Lee, and S. Narayanan, "Assessment of a child's engagement using sequence model based features," in *Workshop on Affective Social Speech Signals 2013*, 2013.

[15] F. Pianesi, M. Zancanaro, B. Lepri, and A. Cappelletti, "A multimodal annotated corpus of consensus decision making meetings," *Language Resources and Evaluation*, vol. 41, no. 3-4, pp. 409–429, 2007.

[16] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, "Elan: a professional framework for multimodality research," in *Proceedings of LREC*, 2006, pp. 5–8.

[17] L. Hang, "A short introduction to learning to rank," *IEICE TRANSACTIONS on Information and Systems*, vol. 94, no. 10, pp. 1854–1862, 2011.

[18] K. Jokinen, "Turn taking, utterance density, and gaze patterns as cues to conversational activity," in *Proceedings of ICMI 2011 Workshop Multimodal Corpora for Machine Learning: Taking Stock and Road mapping the Future, Alicante, Spain*, 2011, pp. 31–36.

[19] N. Campbell and S. Scherer, "Comparing measures of synchrony and alignment in dialogue speech timing with respect to turn-taking activity." in *INTERSPEECH*, 2010, pp. 2546–2549.

[20] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schröder, "Bridging the gap between social animal and unsocial machine: A survey of social signal processing," *Affective Computing, IEEE Transactions on*, vol. 3, no. 1, pp. 69–87, 2012.

[21] T. Joachims, "Optimizing search engines using clickthrough data," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2002, pp. 133–142.

[22] H. Cao, R. Verma, and A. Nenkova, "Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech," *Computer speech & language*, vol. 29, no. 1, pp. 186–202, 2015.

[23] P. Domingos, "A few useful things to know about machine learning," *Communications of the ACM*, vol. 55, no. 10, pp. 78–87, 2012.

[24] C. Busso, P. G. Georgiou, and S. S. Narayanan, "Real-time monitoring of participant's interaction in a meeting using audio-visual sensors," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2007.

[25] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[26] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[27] L. Li and H.-T. Lin, "Ordinal regression by extended binary classification," in *Advances in neural information processing systems*, 2006, pp. 865–872.

[28] R. Fagin, R. Kumar, and D. Sivakumar, "Comparing top k lists," *SIAM Journal on Discrete Mathematics*, vol. 17, no. 1, pp. 134–160, 2003.

[29] R. R. Bouckaert and E. Frank, "Evaluating the replicability of significance tests for comparing learning algorithms," in *Advances in knowledge discovery and data mining*. Springer, 2004, pp. 3–12.