INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Sentence level estimation of psycholinguistic norms using joint multidimensional annotations

Anil Ramakrishna, Shrikanth Narayanan

Signal Analysis and Interpretation Laboratory, University of Southern California, Los Angeles, CA

akramakr@usc.edu, shri@ee.usc.edu

Abstract

Psycholinguistic normatives represent various affective and mental constructs using numeric scores and are used in a variety of applications in natural language processing. They are commonly used at the sentence level, the scores of which are estimated by extrapolating word level scores using simple aggregation strategies, which may not always be optimal. In this work, we present a novel approach to estimate the psycholinguistic norms at sentence level. We apply a multidimensional annotation fusion model on annotations at the word level to estimate a parameter which captures relationships between different norms. We then use this parameter at sentence level to estimate the norms. We evaluate our approach by predicting sentence level scores for various normative dimensions and compare with standard word aggregation schemes.

Index Terms: Psycholinguistic normatives, Annotation fusion, Multidimensional annotations.

1. Introduction

Psycholinguistic norms are numeric ratings assigned to linguistic cues such as words or sentences to measure various psychological constructs. Examples include dimensions such as valence, arousal, and dominance which are used to analyze the affective state of the author (of the spoken or written text), along with norms of higher order mental constructs such as concreteness and imageability which have been associated with improvements in learning [1]. The ease of computing the norms has enabled their application in a variety of tasks in natural language processing such as information retrieval [2], sentiment analysis [3], text based personality prediction [4] and opinion mining. The norms are typically annotated at the word level by psychologists who provide numeric scores to a curated list of seed words, which are then extrapolated to a larger vocabulary using either semantic relationships such as synonymy and hyponymy or using word occurrence based contextual similarity [5].

Most applications of psycholinguistic norms in NLP use sentence or document level scores, but manual annotation of the norms at these levels is difficult and not straightforward to generalize. In these cases, estimation of sentence level norms is done by aggregating the word level scores using simple averaging [6, 5], or by using distribution statistics of the word level scores [7]. However, such strategies do not account for the non-trivial dependencies of sentence level semantics on the words, and may not be accurate at estimating the norms at the sentence level. In this work, we propose a new approach to estimate sentence level norms using inferred relationships between different dimensions along with partial annotations of the sentence level norms.

Annotation of the normatives at the sentence level is a challenging task when compared to word level annotations since it involves evaluating the underlying semantics of the sentence in the abstract space of the corresponding dimension, with some dimensions in particular being more difficult than others. For example, *imageability*, a measure of how easy it is to create a mental image of the input word or sentence, is more difficult to annotate at the sentence level when compared to words. On the other hand, norms such as valence are relatively easier to annotate even at the sentence level in comparison. We use this observation along with the parameters learned from a joint annotation fusion model at word level to predict norms at sentence level.

Annotations are typically performed online using crowdsourcing platforms such as Amazon Mechanical Turk[1] (Mturk), which connect researchers with inexpensive workers from across the globe and provide easy scalability. Annotations are collected from several workers over a large number of instances, often on several related dimensions. These are then combined to obtain estimates for the label of interest, typically using aggregation techniques such as simple averaging or majority voting, or using more nuanced aggregation models which assume a structure for the annotators' behavior [8]. The annotation dimensions are usually modeled independent of each other, but a few recent publications have explored joint modeling of the dimensions and have highlighted the benefits of this approach [9, 10]. These models assume a joint relationship between the dimensions being annotated, and estimate model parameters that capture this relationship for each annotator, which can be used in estimating the sentence level normatives. Specifically, we can use model parameters learned at the word level to estimate the norms at the sentence level using partial sentence level annotations.

We use the model presented in [10], in which the authors assume a matrix factorization model to capture the annotators' behavior, in which the annotations are assumed to be based on a linear transformation of the underlying label vector. Parameters of this model include a linear transformation matrix, F_k, which captures the individual contributions of each dimension in the annotation output. In our work, we assume that the annotator specific relationships between the dimensions captured by the parameter F_k is comparable at both word and sentence levels. We collect word level annotations on valence, arousal and dominance and train the joint global annotation model from [10] to estimate the annotator parameters including F_k; we then use the word level estimates for F_k on sentence level ratings from the same set of annotators. To predict sentence level scores of a given normative dimension, we make use of partial annotations on the remaining dimensions along with F_k. Our proposed ap-

[1] mturk.com

http://dx.doi.org/10.21437/Interspeech.2020-1841

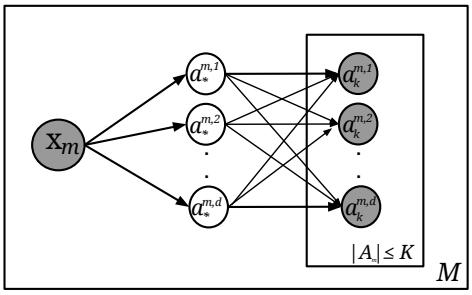


Figure 1: *Proposed model.* $\boldsymbol{x}^m$ *is the set of features for the* m^{th} *data point,* $a_*^{m,d}$ *is the latent label for the* d^{th} *dimension and* $a_k^{m,d}$ *is the rating provided by the* k^{th} *annotator. Vectors* $\boldsymbol{x}^m$ *and* $\boldsymbol{a}_k^m$ *(shaded) are observed variables, while* $\boldsymbol{a}_*^m$ *is latent.* $\boldsymbol{A}_m$ *is the set of annotator ratings for the* m^{th} *instance.*

proach shows improved performance in predicting the sentence level norms when compared to various word level aggregation strategies.

The rest of the paper is organized as follows. In Section 2, we expand on the joint multidimensional annotation model and detail our data annotation approach in Section 3, followed by experiments in Section 4 and results in Section 5 before concluding in Section 6.

2. Joint multidimensional annotation model

The annotation model is represented in plate notation in Figure 1. In this model, the underlying label vector $\mathbf{a}_*^m$ for each data instance is defined as a linear regression model as shown in Equation 1. An annotator, indexed by k, is assumed to apply a linear transformation function on vector $\mathbf{a}_*^m$ to produce the annotation vector $\mathbf{a}_k^m$ using the matrix F_k as shown in Equation 2.

$$\mathbf{a}_*^m = \Theta^T \mathbf{x}_m + \boldsymbol{\epsilon}_m \quad (1)$$

$$\mathbf{a}_k^m = F_k \mathbf{a}_*^m + \boldsymbol{\eta}_k \quad (2)$$

where, $\mathbf{x}_m \in \mathbb{R}^{\mathrm{P}}$; $\Theta \in \mathbb{R}^{\mathrm{P}\times\mathrm{D}}$; $\boldsymbol{\epsilon}_m \sim N(\mathbf{0}, \sigma^2 I)$; $\sigma^2 \in \mathbb{R}$; $\boldsymbol{\eta}_k \sim N(\mathbf{0}, \tau_k^2 I)$; $\tau_k^2 \in \mathbb{R}$. $F_k \in \mathbb{R}^{\mathrm{D}\times\mathrm{D}}$ is the annotator specific linear transformation matrix. Each annotation dimension value $a_k^{m,d}$ for annotator k is defined as a weighted average of the vector $\mathbf{a}_*^m$ with weights given by $F_k(d,:)$.

In this model, the feature vector $\mathbf{x}^m$ corresponding to each instance is assumed to be available, along with the annotations $\mathbf{a}_k^m$, while the label vectors $\mathbf{a}_*^m$ are assumed hidden, as shown in the Figure 1. We use the EM algorithm from [10] to estimate the parameters, listed below for ease of exposition. Detailed derivations for the update equations below can be found in [10].

We use Maximum Likelihood Estimation (MLE) to estimate the model parameters, in which we maximize the model likelihood shown below in Equation 3.

$$\begin{aligned}\log \mathcal{L} &= \sum_{m=1}^{M} \log p(\mathbf{a}_1^m \dots \mathbf{a}_K^m; F_k, \Theta, \sigma^2, \tau_k^2) \\ &= \sum_{m=1}^{M} \log \int_{\mathbf{a}_*^m} p(\mathbf{a}_1^m \dots \mathbf{a}_K^m | \mathbf{a}_*^m; F_k, \tau_k^2) p(\mathbf{a}_*^m; \Theta, \sigma^2)\, d\mathbf{a}_*^m\end{aligned} \quad (3)$$

Optimizing the above objective is non-trivial due to the presence of the integral within the log function. To address this, we use the well known Expectation Maximization algorithm [11], which uses Jensen's inequality to derive a lower bound (shown below in Equation 4) on the objective based on current parameter estimates, by computing the expectation with respect to the conditional distribution $p(\mathbf{a}_*^m | \mathbf{a}_1^m \dots \mathbf{a}_K^m)$.

$$\log \mathcal{L} = \sum_{m=1}^{M} \mathbb{E}_{\mathbf{a}_*^m | \mathbf{a}_1^m \dots \mathbf{a}_K^m} \left[\log \frac{p(\mathbf{a}_1^m \dots \mathbf{a}_K^m | \mathbf{a}_*^m) p(\mathbf{a}_*^m)}{q(\mathbf{a}_*^m)} \right] \quad (4)$$

This is followed by parameter estimation using maximization. The alternating expectation and maximization steps form the iterations of the EM algorithm.

2.1. EM algorithm

Initialization The model is initialized by assigning the mean of annotations for each data instance as the estimate for $\mathbf{a}_*^m$. Given this, the initial parameters are estimated using update equations listed in the maximization step below.

E-step We compute the expected value of $\mathbf{a}_*^m$ with respect to the distribution $p(\mathbf{a}_*^m | \mathbf{a}_1^m \dots \mathbf{a}_K^m)$, which is assumed to be it's *soft* estimate for each data instance.

$$\mathbb{E}_{\mathbf{a}_*^m | \mathbf{a}_1^m \dots \mathbf{a}_K^m}[\mathbf{a}_*^m] = \Theta^T \mathbf{x}_m + \Sigma_{\mathbf{a}_*^m, \mathbf{a}_1^m \dots \mathbf{a}_K^m} \Sigma^{-1}_{\mathbf{a}_1^m \dots \mathbf{a}_K^m, \mathbf{a}_1^m \dots \mathbf{a}_K^m} (\mathbf{a}^m - \boldsymbol{\mu}^m)$$

$$\Sigma_{\mathbf{a}_*^m | \mathbf{a}_1^m \dots \mathbf{a}_K^m}[\mathbf{a}_*^m] = \Sigma_{\mathbf{a}_*^m, \mathbf{a}_*^m} - \Sigma_{\mathbf{a}_*^m, \mathbf{a}_1^m \dots \mathbf{a}_K^m} \Sigma^{-1}_{\mathbf{a}_1^m \dots \mathbf{a}_K^m, \mathbf{a}_1^m \dots \mathbf{a}_K^m} \Sigma_{\mathbf{a}_1^m \dots \mathbf{a}_K^m, \mathbf{a}_*^m}$$

M-step Given the soft estimate for $\mathbf{a}_*^m$, parameter estimates are computed by maximizing Equation 4. The update equations for this step are listed below.

$$\Theta = (\mathrm{X}^T \mathrm{X})^{-1} (\mathrm{X}^T \, \mathbb{E}[\mathrm{a}_*^m])$$

$$F_k = \left(\sum_{m=1}^{M_k} \mathbf{a}_K^m \, \mathbb{E}[(\mathbf{a}_*^m)^T] \right) \left(\sum_{m=1}^{M_k} \mathbb{E}[\mathbf{a}_*^m \, (\mathbf{a}_*^m)^T] \right)^{-1}$$

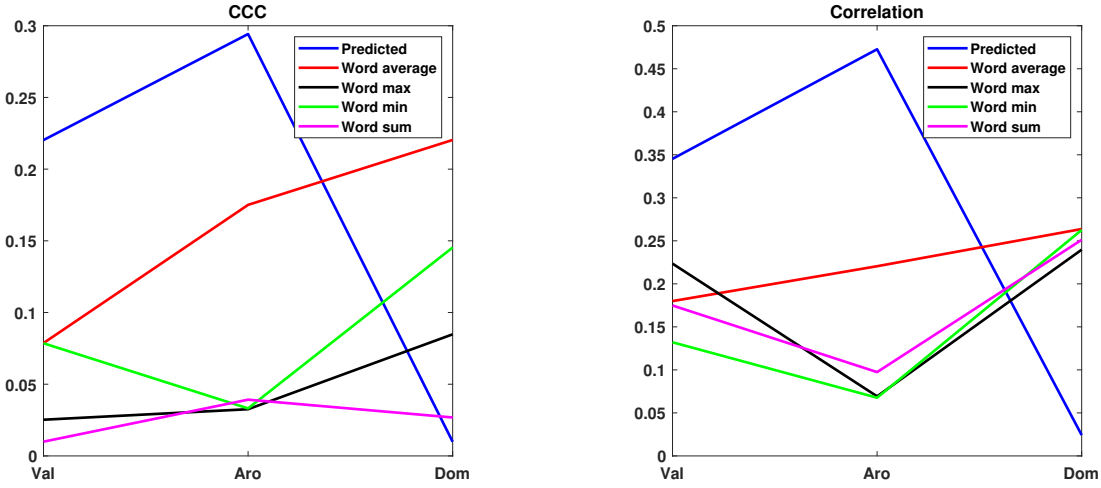
$$\sigma^2 = \frac{1}{md} \sum_{m=1}^{M} \Big(\mathbb{E}[(\mathbf{a}_*^m)^T \mathbf{a}_*^m] - 2tr\left(\Theta'^T \mathbf{x}_m \, \mathbb{E}[(\mathbf{a}_*^m)^T] \right) + tr(\mathbf{x}_m^T \Theta' \Theta'^T \mathbf{x}_m) \Big)$$

$$\tau_k^2 = \frac{1}{m_k d} \sum_{m=1}^{M_k} \Big((\mathbf{a}_K^m)^T \mathbf{a}_K^m - 2tr\big(F_k'^T \mathbf{a}_K^m \, \mathbb{E}[(\mathbf{a}_*^m)^T] \big) + tr\big(F_k'^T F_k' \, \mathbb{E}[\mathbf{a}_*^m (\mathbf{a}_*^m)^T] \big) \Big)$$

Termination We terminate the algorithm when the change in model log-likelihood reduces to less than 0.001% from the previous iteration.

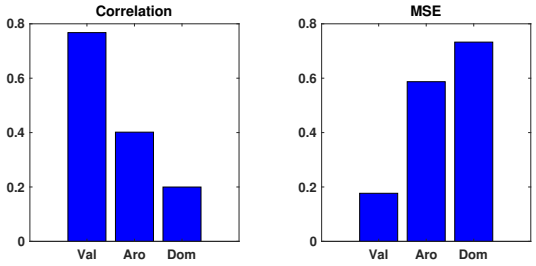
3. Data annotation

We performed two sets of experiments, collecting word and sentence level annotations on specific dimensions in each. In the first experiment (which we refer to as *VAD* from now), we collected annotations on the affective norms of Valence, Arousal and Dominance using Mturk for words sampled from [12]. This corpus was chosen because it provides expert ratings on Valence, Arousal and Dominance for nearly 14,000 English words. Annotators were asked to provide numeric ratings between 1 to



(a) *CCC* (b) *Pearson's Correlation*

Figure 2: *Performance of proposed and baseline models in predicting sentence level norms. Results show Concordance Correlation Coefficient (CCC) and Pearson Correlation values between the various estimates and the reference expert ratings on the EmoBank corpus. The estimates of the proposed model for Valence and Arousal are superior while those for Dominance are poor; subsequent analysis show poor human interannotator agreement for dominance ratings as a possible reason. See also Figure 3.*



(a) *Pearson's Correlation* (b) *Mean Squared Error*

Figure 3: *Performance of the best annotators for each dimension (but over all instances) in our dataset and annotator average when compared with expert ratings from the Emobank corpus*

5 (inclusive) for each dimension, on assignments consisting of a set of 20 words. In total, we collected 20 annotations each on a set of 200 words randomly sampled from [12]. Instructions for the annotation assignments included definitions along with examples for each of the dimensions being annotated. After manually filtering incomplete and noisy submissions, we retained only those annotators who provided ratings for at least 100 words in the subsequent sentence level annotation task, to ensure sufficient training data.

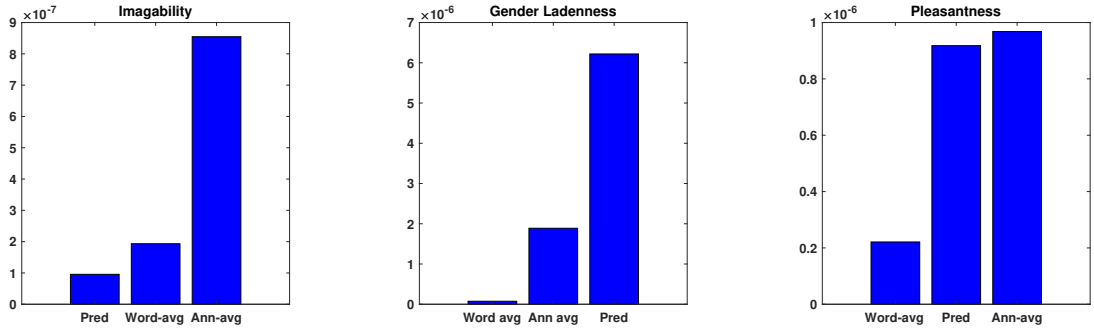
Sentence level annotations were collected on sentences from the Emobank corpus [13], which includes expert ratings on valence, arousal and dominance for 10000 English sentences. 21 different annotators from the word level annotation task described above were invited to provide labels for 100 sentences randomly sampled from this corpus. The assignments were presented in a similar fashion as word level annotations, with each assignment including 10 sentences and the workers providing numeric ratings for valence, arousal and dominance for each sentence. We use the annotator specific parameters F_k estimated at the word level to predict the norms at sentence level using the approach described in the next section.

In our second experiment (which we refer to as *IGP* from now), we collected word and sentence level annotations on three new psycholinguistic normative dimensions: imageability, which measures the degree of the stimulus' proclivity to create a mental picture; genderladenness, which measures the degree of masculine or feminine association evoked by the stimulus; and pleasantness, which measures the degree of pleasant feelings associated with the stimulus. We used the same words and sentences used in our previous experiment for annotations on valence, arousal and dominance. Since we do not have expert ratings for pleasantness, imageability and genderladenness, we use training error as a proxy, where we train linear regression models on labels predicted by the different schemes and compare their MSE to evaluate label performance, since low MSE can be indicative of higher correlations between the labels and features.

4. Experiments

Given annotator parameters F_k estimated at the word level, we use partial annotator ratings at the sentence level to predict the remaining norms. For example, in the *VAD* experiment, while predicting sentence level scores of valence, we use the sentence level annotator ratings on arousal and dominance along with the word level parameter matrix F_k^{word}. The use of partial annotations enables us to predict sentence level norms on challenging psycholinguistic dimensions using ratings on dimensions which maybe easier to annotate.

$$\begin{bmatrix} \cdot \\ a_1^{m,d'} \\ \vdots \\ a_K^{m,d'} \\ \cdot \end{bmatrix} = \begin{bmatrix} \cdot \\ F_1^{\text{word}}[d',:] \\ \vdots \\ F_K^{\text{word}}[d',:] \\ \cdot \end{bmatrix} \begin{bmatrix} \\ \mathbf{a}_*^m \\ \\ \end{bmatrix} \quad (5)$$



(a) *Imageability* (b) *Genderladenness* (c) *Pleasantness*

Figure 4: *MSE of predicted and baseline models in predicting Imageability, Genderladenness and Pleasantness*

where, $d' \in \{1 \dots D\}; d' \neq d$ is the dimension to predict. $\mathbf{a}_*^m$ is estimated using linear regression.

In both our experiments, we make use of the IID Gaussian noise assumption in Equation 2, which reduces the task of predicting the sentence level norm to a linear regression problem shown in Equation 5. Rows of the matrix F_k^{word} are treated as features of the regression model with vector $\mathbf{a}_*^m$ as the regression parameter. Given sentence level partial annotations $\mathbf{a}_k^{m,\setminus d}$ (vector $\mathbf{a}_k^m$ with $a_k^{m,d}$ removed), and matrix F_k^{word}, the regression parameter vector $\mathbf{a}_*^m$ can be estimated using normal equations or gradient descent. For each dimension within a given experiment, we use Equation 5 to estimate the sentence level normatives. The features x_m used in both our experiments were 300 dimensional GloVe embeddings [14] at word level, which were aggregated using simple averaging at sentence level.

In the *VAD* experiment, we compare the predicted dimensions with expert ratings from the Emobank corpus, which acts as our reference to evaluate model performance. For baselines, we compute different aggregations of word level normative scores after filtering out non-content words as is common in literature [6]. Word level scores for the norms were computed using the approach described in [5]. We used unweighted average, maximum, minimum and sum of the word level norms as the baseline aggregation functions.

In the *IGP* experiment, we train linear regression models using predictions from the proposed model and directly compare the training set error with baselines. Low training error implies *higher learnability* (due to better correlations with the features) of the predicted signal and serves as a crude proxy for quality. For baselines, we use training error from labels obtained by simple averaging of word level normative scores, and sentence level average of annotations.

We use Concordance Correlation Coefficient (ρ_c) [15] and the Pearson's correlation coefficient (ρ) as evaluation metrics. ρ_c measures any departures from the *concordance line* (line passing through the origin at 45° angle). Hence it is sensitive to rotations or rescaling in the predicted values of $\mathbf{a}_*^m$. Given two samples x and y, the sample concordance coefficient $\hat{\rho_c}$ is defined as shown below.

$$\hat{\rho_c} = \frac{2 s_{xy}}{s_x^2 + s_y^2 + (\bar{x} - \bar{y})^2} \tag{6}$$

where $\bar{x}$ and $\bar{y}$ are sample means, s_x and s_y are sample standard deviations, and s_{xy} is the sample covariance.

5. Results

5.1. *VAD*

Figure 2 shows the performance of the proposed model along with the different baselines. As seen from the figure, the proposed model outperforms the baselines in predicting valence and arousal in both evaluation metrics, suggesting the efficacy of the approach. Using partial ratings at sentence level along with matrix F_k which captures relationships between the dimensions, the proposed approach seems to outperform the baseline word aggregation schemes in these two dimensions. Performance in ρ_c appears to be lower than ρ, suggesting the presence of a rotation in the predicted values. This can be attributed to the unidentifiability commonly observed in matrix factorization models such as the annotation fusion model of [10]. Common solutions to address this involve assuming a suitable prior on the parameter F_k, which may lead to better estimates of ρ_c.

Model performance on dominance, on the other hand, is considerably low in both metrics. To further investigate the reason for this, we examined the performance of the best possible annotator for each dimension in this experiment and compare their predictions with the expert ratings from the Emobank corpus in Figure 3. Evidently, for dominance, we notice very low correlation and high MSE between our best annotators and the experts, suggesting a high disagreement for this dimension. This may have been due to a possibly differing definition and/or interpretation of dominance between the two sets of annotators.

5.2. *IGP*

In our second experiment, we use model training error as a proxy for evaluating prediction quality since we do not have expert ratings. Figure 4 shows the training error for the proposed model when compared with two baselines. The proposed model shows lowest training error in predicting imageability while the performance is relatively worse in genderladenness and pleasantness, suggesting relatively stronger dependency of imageability on the other dimensions.

6. Conclusion

We presented a novel computational approach to estimate sentence level psycholinguistic norms using joint multidimensional annotation fusion. We evaluate our approach by predicting sentence level normatives on various dimensions in two different experiments, and showed improvements in specific cases. Future work includes evaluating the model on more abstract dimensions such as concreteness.

7. References

[1] A. Paivio, J. C. Yuille, and S. A. Madigan, "Concreteness, imagery, and meaningfulness values for 925 nouns." *Journal of experimental psychology*, vol. 76, no. 1p2, p. 1, 1968.

[2] S. Tanaka, A. Jatowt, M. P. Kato, and K. Tanaka, "Estimating content concreteness for finding comprehensible documents," in *Proceedings of the sixth ACM international conference on Web search and data mining*. ACM, 2013, pp. 475–484.

[3] F. Å. Nielsen, "A new anew: Evaluation of a word list for sentiment analysis in microblogs," *arXiv preprint arXiv:1103.2903*, 2011.

[4] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore, "Using linguistic cues for the automatic recognition of personality in conversation and text," *Journal of artificial intelligence research*, vol. 30, pp. 457–500, 2007.

[5] N. Malandrakis and S. S. Narayanan, "Therapy language analysis using automatically generated psycholinguistic norms." in *INTERSPEECH*, 2015, pp. 1952–1956.

[6] A. Ramakrishna, V. R. Martínez, N. Malandrakis, K. Singla, and S. Narayanan, "Linguistic analysis of differences in portrayal of movie characters," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 1669–1678.

[7] J. Gibson, N. Malandrakis, F. Romero, D. C. Atkins, and S. S. Narayanan, "Predicting therapist empathy in motivational interviews using language features inspired by psycholinguistic norms," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

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

[9] A. Ramakrishna, R. Gupta, R. B. Grossman, and S. S. Narayanan, "An expectation maximization approach to joint modeling of multidimensional ratings derived from multiple annotators," in *Proceedings of Interspeech*, 2016, pp. 1555–1559.

[10] A. Ramakrishna, R. Gupta, and S. Narayanan, "Joint multidimensional model for global and time-series annotations," *arXiv preprint arXiv:2005.03117*, 2020.

[11] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[12] A. B. Warriner, V. Kuperman, and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 english lemmas," *Behavior research methods*, vol. 45, no. 4, pp. 1191–1207, 2013.

[13] S. Buechel and U. Hahn, "Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 2017, pp. 578–585.

[14] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543. [Online]. Available: http://www.aclweb.org/anthology/D14-1162

[15] I. Lawrence and K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, pp. 255–268, 1989.