EMPATHBERT: A BERT-based Framework for Demographic-aware Empathy Prediction

Bhanu Prakash Reddy Guda
Adobe Research
guda@adobe.com

Aparna Garimella
Adobe Research
garimell@adobe.com

Niyati Chhaya
Adobe Research
nchhaya@adobe.com

Abstract

Affect preferences vary with user demographics, and tapping into demographic information provides important cues about the users' language preferences. In this paper, we utilize the user demographics, and propose EMPATHBERT, a demographic-aware framework for empathy prediction based on BERT. Through several comparative experiments, we show that EMPATHBERT surpasses traditional machine learning and deep learning models, and illustrate the importance of user demographics to predict empathy and distress in user responses to stimulative news articles. We also highlight the importance of affect information in the responses by developing affect-aware models to predict user demographic attributes.

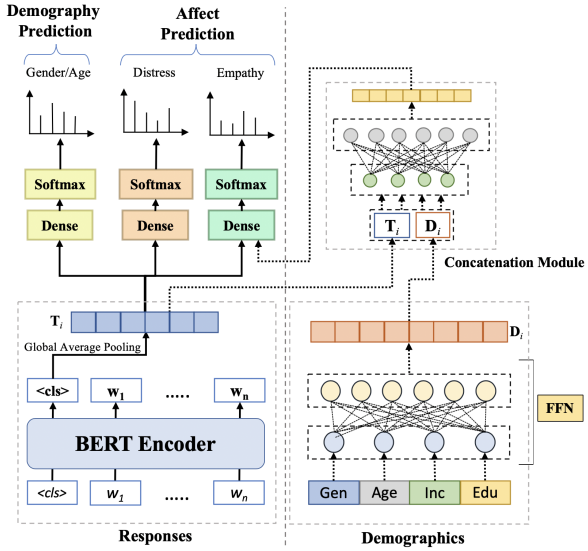

Figure 1: EMPATHBERT architecture.

1 Introduction

Modeling complex human reactions and affect from text has been a challenging research area with innovations focusing on sentiment and emotion understanding (Picard, 1997; Li and Liu, 2015; Rosenthal et al., 2017; Socher et al., 2011, 2013). The study of non-trivial human reactions has been limited. These methods, often rooted in psychological theories, have turned out to be more complex in terms of annotation and modeling (Strapparava and Mihalcea, 2007). A critical affective phenomena, *empathy*, has received surprisingly less attention.

Empathy assesses feelings of sympathy towards *others*, and Distress measures anxiety and discomfort oriented towards *self* (Davis, 1980). Empathy has been positively associated to a number of well-being activities, such as volunteering (Batson et al., 1987), charity (Pavey et al., 2012), and longevity (Poulin et al., 2013), and in consumer marketing, advertising and customer interfaces (Wang et al., 2016; Escalas and Stern, 2003). Works on empathy in text have focused on spoken dialogue, addressing conversational agents, psychological interventions, or call center transcripts (McQuiggan and Lester, 2007; Fung et al., 2016; Pérez-Rosas et al., 2017; Alam et al., 2018; Demasi et al., 2019). Buechel et al. (2018) collected an empathy-distress dataset by leveraging users' reactions to textual stimulus content. Sedoc et al. (2019) constructed an empathy lexicon by obtaining word ratings from document-level ratings from this dataset. Xiao et al. (2012); Gibson et al. (2015); Khanpour et al. (2017) presented predictive models for empathy in the healthcare domain. However, we believe none of the above works focus on (a) predicting empathy from textual reactions, and (b) studying the impact of demographics on the expression of empathy.

Language preferences vary with user demographics (Tresselt and Mayzner, 1964; Eckert and McConnell-Ginet, 2013; Garimella et al., 2016; Lin et al., 2018; Loveys et al., 2018), and this has led to studies leveraging the user demographic information to obtain better language representations and classification models for various NLP

tasks (Volkova et al., 2013; Bamman et al., 2014; Hovy, 2015; Garimella et al., 2017). Owing to the recent success of large language models such as BERT (Devlin et al., 2018) and GPT-2 (Radford et al., 2019) in improving the performances of several downstream tasks, we propose a BERT-based demographic-aware framework for empathy (distress) prediction, and through several comparative experiments, show that it surpasses existing baselines and demographic-agnostic approaches.

This paper makes three main contributions. **(1)** We present EMPATHBERT, a demographic-aware empathy (distress) prediction framework, using BERT-based models infused with demographic information. **(2)** Through comparisons against several baseline and demographic-agnostic approaches, we illustrate the importance of user demographics in end-to-end modeling and predicting empathy (distress). **(3)** Conversely, we show that empathy (distress) also contributes to demographic attribute prediction, by developing affect-aware models for demographic attribute prediction, backed by empirical comparison with baselines and generic models. To the best of our knowledge, ours is the first computational effort addressing empathy (distress) through the lens of demographic biases, a phenomenon well-understood in psychology.

2 Dataset

We use the empathy-distress dataset introduced by Buechel et al. (2018). It consists of 418 news articles from popular news platforms, and responses to them from 403 annotators (5 articles each), resulting in a total of 2,015 responses. Filtering the responses that deviated from the task description led to 1,860 responses (empathy: 916, distress: 905) , with a total token count of 173,686 (min: 52, max: 198, median: 84). The number of responses per article ranges from 1 to 7, with an average of 4.46 responses per article. We report some example responses from the dataset in Table 1[1]. We focus on the responses only, and use the empathy (distress) tags associated with these responses. We group the data into binary classes for age (C_0: < 35, C_1: ≥ 35), income (C_0: $\leq \$50,000$, C_1: $> \$50,000$), and education ($C_0$: no degree, C_1: bachelor's or above), to mitigate class imbalances.[2] The resulting dataset is balanced for all dimensions, with a maximum deviation of 5.5% (age) among classes.

[1]Please refer to Buechel et al. (2018) for further details on the dataset.

[2]We do not study race; it has even heavier class-imbalance.

3 EMPATHBERT

In this section, we describe our approach for demographic-aware empathy (distress) prediction from text. Figure 1 shows the proposed architecture. Our model takes as input a response (a sequence of words $w_1, w_2, \ldots, w_n$) and demographic information of the corresponding annotator. We represent the response using BERT, a bidirectional Transformer-based (Vaswani et al., 2017) language model. We use the final 768-dimensional hidden vector corresponding to the [CLS] token as the aggregate sequence representation. We employ cross-domain pre-training (Sun et al., 2019), fine-tuning, and multi-task fine-tuning (Liu et al., 2019) techniques to customize BERT for our tasks.

Cross-domain Pre-training (PT). We use the pre-trained BERT language model trained on the English Wikipedia and Book Corpus (Zhu et al., 2015) datasets for masked word and next sentence prediction, and perform further pre-training on demographic-specific datasets to introduce demographic-specific language preferences. This enables slanting the BERT model towards a specific demographic group. For this, we use a corpus different from the empathy dataset in two scenarios. (1) ALL: train the BERT model on all of the external corpus, and (2) DEMOGRAPHIC-SPECIFIC: train only on the demographic-specific samples from the external corpus.

Fine-tuning Only (tBERT). BERT-based fine-tuning has had significant success, due to the ease in implementation and performance gains reported for various NLP tasks (Huang et al., 2019; Liu and Lapata, 2019). We fine-tune BERT for sequence classification by adding a classification layer, where the input is response represented by the hidden vector of the [CLS] token, and output is the prediction for empathy (distress). We train on generic data and demographic-specific portions, and compare the performances to study the demographic effect on empathy (distress) prediction.

Multi-task Fine-tuning (tBERT-MT). We fine-tune BERT in multi-task learning (MTL) setup for classification, similar to (Liu et al., 2019), where the tasks under consideration are empathy (distress) classification and demographic attribute prediction. Both the tasks have shared BERT layers, while the classification heads containing the final dense and

TEXT	DEMOGRAPHIC ATTRIBUTES	SCORE
A 6.4 magnitude on the Richter scale earthquake has shaken up the whole capital of Santiago, Chile. Chile is very propense to earthquakes and natural disasters. We have heard of an earthquake that scaled out to be 8.8 and destroys over 200 thousand homes in Chile. I feel very bad for the people who died. and send out my compassion to the family of the 55 dead in this earthquake.	Female, Age $\geq$ 35, Education $<$ Bachelors, Income $\leq$ \$50,000	0.84
This is just crazy, you have to feel for the mother, but at the same time what kind of apartment has that many violations and is still not punished. They need to sue them and anybody involved with this. I can't believe that in today's society that tragedies like this are tolerated. Somebody needs to go to jail for the death of this little girl and the injuries that her mother suffered. I can't imagine what the mother is going through and she probably blames herself. Things like this should just not happen.	Male, Age $\geq$ 35, Education $\geq$ Bachelors, Income $\leq$ \$50,000	0.82

Table 1: Qualitative examples of high empathy (above) and high distress (below) with scores on empathy and distress dimensions as predicted by our **tBERT-C (fnn)** model.

softmax layers are specific to each task. We replace the final dense and softmax layers in tBERT setup with multiple classification heads based on the number of tasks. We experiment with (1) *Alternative training*: In each epoch, we cyclically train only one classification head, freezing the parameters of the remaining heads; and (2) *Parallel training*: In each epoch, we train the model end-to-end on the joint loss from all the classification heads.

Explicit Demographic Knowledge. PT, tBERT and tBERT-MT intrinsically infuse demographic information. We also incorporate this explicitly by concatenating a *demographic vector* $\vec{d}$ to the output of the global average pooling layer (Lin et al., 2013) from tBERT or tBERT-MT (concatenation in Figure 1) in two ways. (1) **tBERT-[MT]-C**: $\vec{d}$ is a d-dimensional one-hot encoding vector (d: number of demographics). (2) **tBERT-[MT]-C (fnn)**: $\vec{d}$ is the output of a feedforward neural network (FNN), the input for which is a one-hot encoding vector. Three dense layers are stacked before the task-specific heads, and this model is trained end-to-end for empathy (distress) prediction. In tBERT-MT where one of the tasks heads predicts a demographic attribute, the corresponding binary value in $\vec{d}$ is removed. To assess the contribution of specific attributes, we also propose to concatenate a 1-bit encoding (**tBERT-[MT]-C (attribute)**) for each given attribute.

4 Experiments

We model empathy (distress) prediction as a binary classification task. To study the efficacy of empathy (distress) to predict demography attributes, we also conduct experiments for empathy (distress)-aware demographic attribute prediction. Such a prediction can be used for further demographic removal from text to mitigate adversarial attacks and protect privacy of users (Elazar and Goldberg, 2018).

Implementation Details (1) Cross-domain Pre-training: We use the Blog Authorship Corpus[4] (Schler et al., 2006), which consists of 681,288 blogposts and self-provided demographic attributes, gender, age, industry, and astrological sign of the corresponding 19,320 bloggers to further pre-train BERT. Out of these we use the gender attribute to pre-train for male-specific and female-specific pre-training experiments. We train the model on the *Masked Language Model task* (Taylor, 1953) for 10 epochs using a learning rate of 3e-5. (2) Finetuning: We train the model end-to-end (110M parameters) using binary cross-entropy loss and decoupled weight decay Adam optimizer (Loshchilov and Hutter, 2017), in batches of 32. The best performance is observed when the maximum input sequence length is set to 150, learning rate to 3e-5, and number of epochs to 3. (3) Explicit Demographic Attributes: We use gender, age, education and income attributes corresponding to each annotator in the empathy dataset. The d-dimensional vector size 4 resulting in a 16-d FFN output.

Evaluation metrics. We use five-fold cross validation (five random shuffled restarts) with 80-20

[3] Statistical significance using McNemar's Test (McNemar, 1947) with * $p < 0.05$, † $p < 0.01$, ‡ $p < 0.001$.

[4] https://u.cs.biu.ac.il/~koppel/BlogCorpus.htm

	Method→	PT			tBERT			PT + tBERT		
	Test→	M	F	A_s	M	F	A_s	M	F	A_s
Empathy	Male	**50.02**†	52.42	62.70	**64.73**‡	60.37	62.22	**61.82**†	57.95	58.65
	Female	49.07	**53.12***	48.28	63.70	**64.56**‡	63.32	58.16	**61.77**‡	58.51
	All_s	49.74	52.91	49.64	63.08	62.19	63.00	57.24	58.63	56.30
Distress	Male	**51.21***	52.26	52.41	**64.44**‡	61.56	62.11	**61.92**†	57.60	59.63
	Female	50.71	**52.77***	51.57	61.52	**63.16**‡	60.51	57.35	**59.30**†	60.19
	All_s	49.43	51.42	50.53	63.18	62.77	62.88	59.78	58.77	59.57

Table 2: Accuracies using gender-specific training for empathy (distress) prediction. **M**ale, **F**emale, $A_s$ll denote the respective data subsets. A_s is a sampled dataset with approximately equal number of samples from M and F subsets, hence comparable in size.[3]

	Dem→	Age			Income			Education		
	Test→	C_0	C_1	A_s	C_0	C_1	A_s	C_0	C_1	A_s
Empathy	$Class_0$	**62.79***	62.59	61.44	**62.05**‡	61.82	62.66	59.44	61.18	58.81
	$Class_1$	59.27	**64.95***	60.05	58.40	**64.96**‡	60.41	**62.34**‡	**63.40**‡	61.40
	All_s	59.73	62.05	60.26	60.62	63.21	60.92	60.81	63.03	59.38
Distress	$Class_0$	**62.80**‡	62.32	61.01	**62.46**‡	61.43	60.86	**62.65**‡	62.04	62.30
	$Class_1$	57.20	**68.08**‡	60.68	60.23	**66.59**†	62.39	60.45	**66.85**†	63.31
	All_s	60.89	65.08	61.16	59.92	61.54	60.88	61.80	63.13	62.06

Table 3: Demographic-specific training accuracies for empathy (distress) prediction.

train-test proportions, and report the F1 and accuracy (Ac) averaged across the 5 runs on the test set.

Baselines. We compare our model against the Random Forest (RF) model with Glove embeddings (Pennington et al., 2014) for text and demographic attributes (excluding the prediction attribute) as one-hot vectors as features. We also report performance against deep learning baselines, CNN (Kim, 2014), biLSTM, and biLSTM with Attention (Yang et al., 2016) and the pre-trained BERT without further training.

4.1 Results

Table 2 shows the accuracies using BERT for pre-training (PT), fine-tuning (tBERT), and both (PT + tBERT) for gender-specific empathy (distress) prediction. On the M and F test sets, models trained on the same demographic subset (M or F) outperform those trained on the opposite subset or A_s. The acccuracies of plain BERT are 48.37, 49.49, and 50.42 on A_s, M, and F test sets respectively for empathy prediction. tBERT outperforms all other variants. The results support our hypothesis that empathy is dependent on and influenced by the gender associated with the author. We note similar patterns for age, income, and education (Table 3).

Table 4 shows results for empathy (distress) prediction using tBERT-[MT]-[C (fnn/attribute)] variants trained on the full dataset. In the notation, we replace [MT] with the heads on which the multi-tasking is performed. For example tBERT-MT-(E+D)-G-C implies fine-tuned BERT with empathy prediction, distress prediction, and gender prediction multi-tasking heads with demographic information concatenated to the text representation directly before classification.[5] We report performances on demographic-wise test sets (A, M, F).

Insights: (1) **tBERT** variants with a single training objective outperform all baselines. (2) Performance of **tBERT-MT** varies with the affect dimension. Empathy prediction shows marginal loss in performance with explicit concatenation (**tBERT-C**) and further loss in the multitask setup. (3) For distress, introduction of gender as the demographic attribute shows an observable improvement across different test sets. (4) A similar trend is observed for age. Table 5 shows performance of age and gender prediction with empathy (distress)-aware models on affect-wise test sets (Empathy (Em) and Distress (Dist)). Empathy-aware gender prediction models show consistent improvement over baselines, with **tBERT (G)** reporting the best score when tested on the complete dataset and empathy-specific test set. **tBERT (A)** helps improve the accuracies for age prediction by atleast 5% over baselines for the com-

[5] In the models where a demographic attribute prediction is involved, we remove that attribute from the demographic vector.

	Affect⟶	Empathy						Distress					
	Test Set⟶	All		Male		Female		All		Male		Female	
	Approach↓	F_1	**Ac**	F_1	**Ac**	F_1	**Ac**	F_1	**Ac**	F_1	**Ac**	F_1	**Ac**
Trad.ML	RF-text	57.5	59.9	58.4	60.3	56.4	57.5	58.4	61.0	58.9	60.9	57.9	61.3
	RF-dem	58.99	59.12	58.59	58.64	59.77	59.77	58.06	58.03	60.61	60.70	59.77	59.29
	RF-text+dem	58.5	60.7	57.9	59.7	57.1	59.7	58.4	60.5	59.6	59.9	58.1	61.2
DL Models	CNN	59.5	61.3	60.7	62.1	58.2	60.5	58.8	63.9	57.8	62.5	59.9	63.5
	biLSTM	53.3	55.4	55.4	57.1	50.8	53.5	57.1	59.3	54.3	56.9	60.2	62.1
	biLSTM-Attention	60.8	62.6	60.0	62.0	61.7	63.3	59.9	62.7	59.8	62.3	59.8	63.1
	BERT	65.6	49.0	65.3	48.8	65.9	49.2	66.1	49.5	65.8	49.3	66.3	49.6
Proposed Methods	Aff-biLSTM-text+dem	61.9	63.0	63.0	63.6	60.8	62.3	62.9	64.2	62.9	64.6	62.9	63.7
	tBERT (E)	**67.1**†	**67.8**†	**68.7***	**69.4***	**65.4***	**66.1***	–	–	–	–	–	–
	tBERT (D)	–	–	–	–	–	–	67.6	67.0	69.3	68.6	65.7	65.1
	tBERT-MT-(E+D)	65.2	66.2	66.7	67.6	63.6	64.6	**69.2**‡	**68.5**‡	71.2	70.4	**66.7**†	**66.3**†
	tBERT-MT-G	(E) 63.9	64.9	65.0	66.7	62.8	62.8	(D) 67.0	67.5	70.8	70.3	62.3	64.3
	tBERT-MT-(E+D)-G	64.5	64.7	65.7	66.3	63.1	63.0	68.1	67.7	**71.3***	**70.5***	64.3	64.5
	tBERT-MT-A	(E) 61.8	63.5	65.3	66.7	58.1	60.0	(D) 65.0	65.1	67.6	67.5	62.2	62.5
	tBERT-MT-(E+D)-A	64.1	65.2	65.8	66.8	62.3	63.5	66.0	66.2	69.3	68.8	62.1	63.1
	tBERT-C (fnn)	(E) 66.4	67.4	67.6	68.6	65.0	66.0	67.4	67.4	69.4	69.2	65.0	65.3
	tBERT-C	66.0	66.4	66.8	67.0	65.0	65.8	68.2	67.7	69.9	69.5	66.2	65.6
	tBERT-C (gender)	64.3	66.8	65.0	67.7	63.5	65.9	66.8	66.9	68.6	68.7	64.7	64.9
	tBERT-MT-G-C	(E) 63.8	66.0	65.4	67.6	62.2	64.2	(D) 65.9	67.0	68.6	68.9	62.5	64.9
	tBERT-MT-(E+D)-G-C	(E) 62.2	64.0	64.5	65.7	59.7	62.2	(D) 64.6	66.1	67.5	68.3	61.3	63.6

Table 4: Demographic-aware empathy (distress) prediction. For tBERT-MT, the multitask attributes are specified in the method name i.e. gender (-G), age (-A) along with empathy (E) or distress (D) along side the accuracies. F_1: F1 score; Ac: Accuracy.

Demography ⟶	Gender			Age		
Dataset ⟶	All	Em	Dist	All	Em	Dist
RF-text	59.8	60.8	58.7	56.5	55.7	57.3
RF-text-E/D	58.0	59.1	56.9	56.6	54.2	59.1
Aff-biLSTM(att)-text	59.2	60.1	58.2	56.2	57.7	54.6
Aff-biLSTM(att)-text-E/D	58.9	60.2	57.4	56.9	57.3	56.6
BERT	47.5	47.3	47.7	40.5	41.3	39.6
tBERT	(G) 64.2‡	65.2	63.4	(A) 62.7*	63.2‡	63.8‡
tBERT-MT-E	62.0	61.5	63.3	60.1	61.1	61.9
tBERT-MT-D	61.6	61.7	63.9	60.6	60.8	61.7
tBERT-MT-(E+D)	63.1	62.9	65.1*	61.6	59.8	63.7

Table 5: F1 values of affect-aware demography prediction.

plete (All) test set. For the empathy-specific test set, best results are observed with MTL (**tBERT-MT-(E+D)**). We infer that while having affect-aware demographic prediction models do improve performance over fine-tuned models, they may also lead to a marginally negative impact. The overall inference from above experiments is that demographic-aware models aid affect predictions but the reverse relationship is much weaker. End-to-end training across a variety of train sets and demographic attributes establishes that the variance observed in language preferences and expressions has an impact on the manner of expressing empathy and distress in reactions.

5 Conclusion

We proposed a novel demographic-aware empathy prediction framework based on fine-tuning and multi-tasking using BERT, showed that it surpasses existing methods, and illustrated the impact of demography in modeling subjective phenomena such as empathy and distress. Our framework is generalizable, and we extended it to empathy-aware demography prediction, and showed that empathy also improves demographic prediction. We believe this is a significant checkpoint towards developing models for empathy (distress), and tapping into demographic information while doing so.

References

Firoj Alam, Morena Danieli, and Giuseppe Riccardi. 2018. Annotating and modeling empathy in spoken conversations. *Computer Speech & Language*, 50:40–61.

David Bamman, Chris Dyer, and Noah A. Smith. 2014. Distributed representations of geographically situated language. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 828–834, Baltimore, Maryland. Association for Computational Linguistics.

C Daniel Batson, Jim Fultz, and Patricia A Schoenrade. 1987. Distress and empathy: Two qualitatively distinct vicarious emotions with different motivational consequences. *Journal of personality*, 55(1):19–39.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4758–4765, Brussels, Belgium. Association for Computational Linguistics.

Mark H Davis. 1980. Interpersonal reactivity index (iri). *A multidimensional approach to individual differences in empathy. JSAS Catalog of Selected Documents in Psychology*, 10:85.

Orianna Demasi, Marti A. Hearst, and Benjamin Recht. 2019. Towards augmenting crisis counselor training by improving message retrieval. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 1–11, Minneapolis, Minnesota. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Penelope Eckert and Sally McConnell-Ginet. 2013. *Language and gender*. Cambridge University Press.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21, Brussels, Belgium. Association for Computational Linguistics.

Jennifer Edson Escalas and Barbara B Stern. 2003. Sympathy and empathy: Emotional responses to advertising dramas. *Journal of Consumer Research*, 29(4):566–578.

Pascale Fung, Anik Dey, Farhad Bin Siddique, Ruixi Lin, Yang Yang, Yan Wan, and Ho Yin Ricky Chan. 2016. Zara the Supergirl: An empathetic personality recognition system. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 87–91, San Diego, California. Association for Computational Linguistics.

Aparna Garimella, Carmen Banea, and Rada Mihalcea. 2017. Demographic-aware word associations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2285–2295, Copenhagen, Denmark. Association for Computational Linguistics.

Aparna Garimella, Rada Mihalcea, and James Pennebaker. 2016. Identifying cross-cultural differences in word usage. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 674–683, Osaka, Japan. The COLING 2016 Organizing Committee.

James Gibson, Nikolaos Malandrakis, Francisco Romero, David C Atkins, and Shrikanth S Narayanan. 2015. Predicting therapist empathy in motivational interviews using language features inspired by psycholinguistic norms. In *Sixteenth annual conference of the international speech communication association*.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 752–762, Beijing, China. Association for Computational Linguistics.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. Glossbert: Bert for word sense disambiguation with gloss knowledge. *arXiv preprint arXiv:1908.07245*.

Hamed Khanpour, Cornelia Caragea, and Prakhar Biyani. 2017. Identifying empathetic messages in online health communities. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 246–251, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Chen Li and Yang Liu. 2015. Improving named entity recognition in tweets via detecting non-standard words. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 929–938, Beijing, China. Association for Computational Linguistics.

Bill Yuchen Lin, Frank F. Xu, Kenny Zhu, and Seungwon Hwang. 2018. Mining cross-cultural differences and similarities in social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 709–719, Melbourne, Australia. Association for Computational Linguistics.

Min Lin, Qiang Chen, and Shuicheng Yan. 2013. Network in network. *arXiv preprint arXiv:1312.4400*.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. MoEL: Mixture of empathetic listeners. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language*

Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 121–132, Hong Kong, China. Association for Computational Linguistics.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504*.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Kate Loveys, Jonathan Torrez, Alex Fine, Glen Moriarty, and Glen Coppersmith. 2018. Cross-cultural differences in language markers of depression online. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 78–87, New Orleans, LA. Association for Computational Linguistics.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Scott W McQuiggan and James C Lester. 2007. Modeling and evaluating empathy in embodied companion agents. *International Journal of Human-Computer Studies*, 65(4):348–360.

Louisa Pavey, Tobias Greitemeyer, and Paul Sparks. 2012. "i help because i want to, not because you tell me to" empathy increases autonomously motivated helping. *Personality and Social Psychology Bulletin*, 38(5):681–689.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. 2017. Understanding and predicting empathic behavior in counseling therapy. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1426–1435, Vancouver, Canada. Association for Computational Linguistics.

Rosalind W. Picard. 1997. *Affective Computing*. MIT Press, Cambridge, MA, USA.

Michael J Poulin, Stephanie L Brown, Amanda J Dillard, and Dylan M Smith. 2013. Giving to others and the association between stress and mortality. *American journal of public health*, 103(9):1649–1655.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Hannah Rashkin, Eric Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada. Association for Computational Linguistics.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 6, pages 199–205.

João Sedoc, Sven Buechel, Yehonathan Nachmany, Anneke Buffone, and Lyle Ungar. 2019. Learning word ratings for empathy and distress from document-level user responses. *arXiv preprint arXiv:1912.01079*.

Richard Socher, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng, and Christopher D. Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 151–161, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74, Prague, Czech Republic. Association for Computational Linguistics.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *China National Conference on Chinese Computational Linguistics*, pages 194–206. Springer.

Wilson L Taylor. 1953. "cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433.

Margaret E Tresselt and Mark S Mayzner. 1964. The kent-rosanoff word association: Word association norms as a function of age. *Psychonomic Science*, 1(1-12):65–66.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1815–1827, Seattle, Washington, USA. Association for Computational Linguistics.

Chen Wang, Rui Juliet Zhu, and Todd C Handy. 2016. Experiencing haptic roughness promotes empathy. *Journal of Consumer Psychology*, 26(3):350–362.

Bo Xiao, Dogan Can, Panayiotis G Georgiou, David Atkins, and Shrikanth S Narayanan. 2012. Analyzing the language of therapist empathy in motivational interview based psychotherapy. In *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1–4. IEEE.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J. Smola, and Eduard H. Hovy. 2016. Hierarchical attention networks for document classification. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, page 1480–1489.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.