Modeling Empathy and Distress in Reaction to News Stories

Sven Buechel [*†3]   **Anneke Buffone** [*1]   **Barry Slaff** [1]   **Lyle Ungar** [1,2]   **João Sedoc** [1,2]

[1] Positive Psychology Center, University of Pennsylvania
[2] Computer & Information Science, University of Pennsylvania
[3] JULIE Lab, Friedrich-Schiller-Universität Jena

https://wwbp.org [1,2]   https://julielab.de [3]

Abstract

Computational detection and understanding of empathy is an important factor in advancing human-computer interaction. Yet to date, text-based empathy prediction has the following major limitations: It underestimates the psychological complexity of the phenomenon, adheres to a weak notion of ground truth where empathic states are ascribed by third parties, and lacks a shared corpus. In contrast, this contribution presents the first publicly available gold standard for empathy prediction. It is constructed using a novel annotation methodology which reliably captures empathy assessments by the writer of a statement using multi-item scales. This is also the first computational work distinguishing between multiple forms of empathy, empathic concern, and personal distress, as recognized throughout psychology. Finally, we present experimental results for three different predictive models, of which a CNN performs the best.

1 Introduction

Over two decades after the seminal work by Picard (1997) the quest of *Affective Computing*, to ease the interaction with computers by giving them a sense of how emotions shape our perception and behavior, is still far from being fulfilled. Undoubtedly, major progress has been made in NLP, with sentiment analysis being one of the most vivid and productive areas in recent years (Liu, 2015).

However, the vast majority of contributions has focused on *polarity prediction*, typically only distinguishing between positive and negative feeling or evaluation, usually in social media postings or product reviews (Rosenthal et al., 2017; Socher et al., 2013). Only very recently, researchers started exploring more sophisticated models of human emotion on a larger scale (Wang et al., 2016; Abdul-Mageed and Ungar, 2017; Mohammad and Bravo-Marquez, 2017a; Buechel and Hahn, 2017, 2018a,b). Yet such approaches, often rooted in psychological theory, also turned out to be more challenging in respect to annotation and modeling (Strapparava and Mihalcea, 2007).

Surprisingly, one of the most valuable affective phenomena for improving human-machine interaction has received surprisingly little attention: *Empathy*. Prior work focused mostly on *spoken dialogue*, commonly addressing conversational agents, psychological interventions, or call center applications (McQuiggan and Lester, 2007; Fung et al., 2016; Pérez-Rosas et al., 2017; Alam et al., 2017).

In contrast, to the best of our knowledge, only three contributions (Xiao et al., 2012; Gibson et al., 2015; Khanpour et al., 2017) previously addressed *text-based* empathy prediction[1] (see Section 4 for details). Yet, all of them are limited in three ways: (a) neither of their corpora are available leaving the NLP community without shared data, (b) empathy ratings were provided by others than the one actually experiencing it which qualifies only as a weak form of ground truth, and (c) their notion of empathy is quite basic, falling short of current and past theory.

---

[1] Psychological studies commonly distinguish between *state* and *trait* empathy. While the former construct describes the amount of empathy a person experiences as a direct result of encountering a given stimulus, the latter refers to how empathetic one is on average and across situations. This studies exclusively addresses *state empathy*. For a contribution addressing *trait empathy* from an NLP perspective, see Abdul-Mageed et al. (2017).

In this contribution we present the first publicly available gold standard for text-based empathy prediction. It is constructed using a novel annotation methodology which reliably captures empathy assessments via multi-item scales. The corpus as well as our work as a whole is also unique in being—to the best of our knowledge—the first computational approach differentiating *multiple types of empathy*, empathic concern and personal distress, a distinction well recognized throughout psychology and other disciplines.[2]

2   Corpus Design and Methodology

Background.   Most psychological theories of empathic states are focused on reactions to negative rather than positive events. Empathy for positive events remains less well understood and is thought to be regulated differently (Morelli et al., 2015). Thus we focus on empathetic reactions to need or suffering. Despite the fact that everyone has an immediate, implicit understanding of empathy, research has been vastly inconsistent in its definition and operationalization (Cuff et al., 2016). There is agreement, however, that there are multiple forms of empathy (see below). The by far most widely cited state empathy scale is Batson's Empathic Concern – Personal Distress Scale (Batson et al., 1987), henceforth *empathy* and *distress*.

Distress is a self-focused, negative affective state that occurs when one feels upset due to witnessing an entity's suffering or need, potentially via "catching" the suffering target's negative emotions. Empathy is a warm, tender, and compassionate feeling for a suffering target. It is other-focused, retains self-other separation, and is marked by relatively more positive affect (Batson and Shaw, 1991; Goetz et al., 2010; Mikulincer and Shaver, 2010; Sober and Wilson, 1997).

Selection of News Stories.   Two research interns (psychology undergraduates) collected a total of 418 articles from popular online news platforms, selected to likely evoke empathic reactions, after being briefed on the goal and background of this study. These articles were then used to elicit empathic responses in participants.

Acquiring Text and Ratings.   The corpus acquisition was set up as a crowdsourcing task on `MTurk.com` pointing to a `Qualtrics.com` questionnaire. The participants completed back-

ground measures on demographics and personality, and then proceeded to the main part of the survey where they read a random selection of five of the news articles. After reading each of the articles, participants were asked to rate their level of empathy and distress before describing their thoughts and feelings about it in writing.

In contrast to previous work, this set-up allowed us to acquire empathy scores of the actual *writer* of a text, instead of having to rely on an external evaluation by third parties (often student assistants with background in computer science). Arguably, our proposed annotation methodology yields more appropriate gold data, yet also leads to more variance in the relationship between linguistic features and empathic state ratings. That is because each rating reflects a single individual's feelings rather than a more stable average assessment by multiple raters. To account for this, we use *multi-item scales* as is common practice in psychology. I.e., participants give ratings for multiple items measuring the same construct (e.g., empathy) which are then averaged to obtain more reliable results. As far as we know, this is the first time that multi-item scales are used in sentiment analysis.[3]

In our case, participants used Batson's Empathic Concern – Personal Distress Scale (see above), i.e, rating 6 items for empathy (e.g., *warm, tender, moved*) and 8 items for distress (e.g., *troubled, disturbed, alarmed*) using a 7-point scale for each of those (see Appendix for details). After rating their empathy, participants were asked to share their feelings about the article as they would with a friend in a private message or with a group of friends as a social media post in 300 to 800 characters. Our final gold standard consists of these *messages* combined with the numeric ratings for empathy and distress.

In sum, 403 participants completed the survey. Median completion time was 32 minutes and each participant received 4 USD as compensation.

Post-Processing.   Each message was manually reviewed by the authors. Responses which deviated from the task description (e.g., mere copying from the articles at display) were removed (31 responses, 155 messages), leading to a total 1860 messages in our final corpus. Gold ratings for empathy and distress were derived by averaging the respective items of the two multi-item scales.

[3] Here, we use *sentiment* as an umbrella term subsuming semantic orientation, emotion, as well as highly related concepts such as empathy.

	E	D	Message
(1)	4.8	3.1	*I'm sorry to hear that about Dakota's parents. Even when you are adult it must be hard to see your parents splitting up. No one wants that to happen and it's unfortunate that her parents couldn't work it out. I hope they are able to still remain civil around the kids and family. Just because it didn't work romantically doesn't mean it won't work at all.*
(2)	4.0	5.5	*Here's an article about crazed person who murdered two unfortunate women overseas. Life is crazy. I can't imagine what the families are going through. Having to go to or being forced into sex work is bad enough, but for it to end like this is just sad. It feels like there's no place safe in this world to be a woman sometimes.*
(3)	1.0	1.3	*I just read an article about some chowder-head who used a hammer and a pick ax to destroy Donald Trump's star on the Hollywood walk of fame. Wow, what a great protest. You sure showed him. Good job. Lol, can you believe this garbage? Who has such a hollow and pathetic life that they don't have anything better to do with their time than commit petty vandalism because they dislike some politician? What a dingus.*

Table 1: Illustrative examples from our newly created gold standard with ratings for empathy (**E**) and distress (**D**).
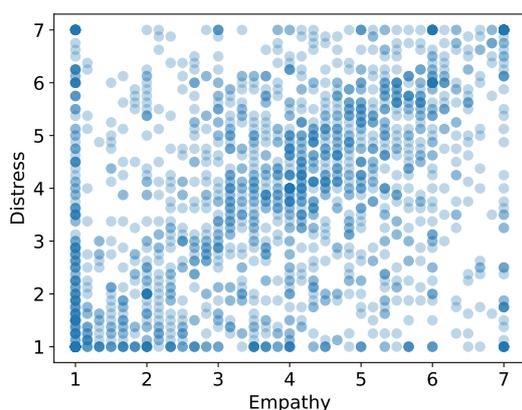

Figure 1: Scatter plot of the bivariate distribution of empathy and distress ratings.

3 Corpus Analysis

For a first impression of the language of our new gold standard, we provide illustrative examples in Table 1. The participant in Example (1) displays higher empathy than distress, (2) displays higher distress than empathy, and (3) shows neither empathic state, but employs sarcasm, colloquialisms and social-media-style acronyms to express lack of emotional response to the article. As can be seen, the language of our corpus is diverse and authentic, featuring many phenomena of natural language which render its computational understanding difficult, thus constituting a sound but challenging gold standard for empathy prediction.

Token Counts. We tokenized the 1860 messages using NLTK tools (Bird, 2006). In total, our corpus amounts to 173, 686 tokens. Individual message length varies between 52 and 198 tokens, the median being 84. See Appendix for details.

Rating Distribution. Figure 1 displays the bivariate distribution of empathy and distress rat-

ings. As can be seen both target variables have a clear linear dependence, yet show only a moderate Pearson correlation of $r=.451$, similar to what was found in prior research (Batson et al., 1987, 1997). This finding supports that the two scales capture distinct affective phenomena and underscores the importance of our decision to describe empathic states in terms of *multiple* target variables, constituting a clear advancement over previous work. Both kinds of ratings show good coverage over the full range of the scales.

Reliability of Ratings. Since each message is annotated by only one rater, its author, typical measures of inter-rater agreement are not applicable. Instead, we compute *split-half reliability* (SHR), a standard approach in psychology (Cronbach, 1947) which also becomes increasingly popular in sentiment analysis (Mohammad and Bravo-Marquez, 2017a; Buechel and Hahn, 2018a). SHR is computed by splitting the ratings for the individual scale items (e.g., *warm*, *tender*, etc. for empathy) of all participants randomly into two groups, averaging the individual item ratings for each group and participant, and then measuring the correlation between both groups. This process is repeated 100 times with random splits, before again averaging the results. Doing so for empathy and distress, we find very high[4] SHR values of $r=.875$ and $.924$, respectively.

4 Modeling Empathy and Distress

In this section, we provide experimental results for modeling empathy and distress ratings based on the participants' messages (see Section 2). We examine three different types of models, varying in

---

[4] For a comparison against previously reported SHR values for different emotional categories, see Mohammad and Bravo-Marquez (2017b).

design complexity. Distinct models were trained for empathy and distress prediction.

First, ten percent of our newly created gold standard were randomly sampled to be used in development experiments. Then, the main experiment was conducted using 10-fold cross-validation (CV), providing each model with identical train-test splits to increase reliability. The dev set was excluded for the CV experiment.

Model performance is measured in terms of Pearson correlation r between predicted values and the human gold ratings. Thus, we phrase the prediction of empathy and distress as regression problems.

The input to our models is based on word embeddings, namely the publicly available Fast-Text embeddings which were trained on Common Crawl ($\approx$600B tokens) (Bojanowski et al., 2017; Mikolov et al., 2018).

Ridge. Our first approach is Ridge regression, an ℓ^2-regularized version of linear regression. The centroid of the word embeddings of the words in a message is used as features (embedding centroid). The regularization coefficient α is automatically chosen from $\{1, .5, .1, ..., .0001\}$ during training.

FFN. Our second approach is a Feed-Forward Net with two hidden layers (256 and 128 units, respectively) with ReLU activation. Again, the embedding centroid is used as features.

CNN. The last approach is a Convolutional Neural Net.[5] We use a single convolutional layer with filter sizes 1 to 3, each with 100 output channels, followed by an average pooling layer and a dense layer of 128 units. ReLUs were used for the convolutional and again for the dense layer.

Both deep learning models were trained using the Adam optimizer (Kingma and Ba, 2015) with a fixed learning rate of 10^{-3} and a batch size of 32. We trained for a maximum of 200 epochs yet applied early stopping if the performance on the validation set did not improve for 20 consecutive epochs. We applied dropout with probabilities of .2, .5 and .5 on input, dense and pooling layers, respectively. Moreover ℓ^2 regularization of .001 was applied to the weights of conv and dense layers. Word embeddings were not updated.

The results are provided in Table 2. As can be seen, all of our models achieve satisfying performance figures ranging between r=.379 and .444,

	Empathy	Distress	Mean
Ridge	.385	.410	.398
FFN	.379	.401	.390
CNN	**.404***	**.444***	**.424***

Table 2: Model performance for predicting empathy and distress in Pearson's r; with row-wise mean; best result per column in bold, significant ($p < .05$) improvement over other models marked with '*'.

given the assumed difficulty of the task (see Section 3). On average over the two target variables, the CNN performs best, followed by Ridge and the FFN. While the CNN significantly outperforms the other models in every case, the differences between Ridge and the FFN are not statistically significant for either empathy or distress.[6] The improvements of the CNN over the other two approaches are much more pronounced for distress than for empathy. Since only the CNN is able to capture semantic effects from composition and word order, our data suggest that these phenomena are more important for predicting distress, whereas lexical features alone already perform quite well for empathy.

Discussion. In comparison to closely related tasks such as emotion prediction (Mohammad and Bravo-Marquez, 2017a) our performance figures for empathy and distress prediction are generally lower. However, given the small amount of previous work for the problem at hand, we argue that our results are actually quite strong. This becomes obvious, again, in comparison with emotion analysis where early work achieved correlation values around r=.3 at most (Strapparava and Mihalcea, 2007). Yet state-of-the-art performance literally doubled over the last decade (Beck, 2017), in part due to much larger training sets.

Comparison to the limited body of previous work in text-based empathy prediction is difficult for a number of reasons, e.g., differences in domain, evaluation metric, as well as methodology and linguistic level of annotation. Khanpour et al. (2017) annotate and model empathy in online health communities on the *sentence*-level, whereas the instances in our corpus are much longer and comprise multiple sentences. In contrast to our work, they treat empathy prediction as a classification problem. Their best performing model, a CNN-LSTM, achieves an F-score of .78. Gibson

---

[5] Recurrent models did not perform well during development due to high sequence length.

[6] We use a two-tailed t-test for paired samples based on the results of the individual CV runs; $p < .05$.

et al. (2015) predict therapists' empathy in motivational interviews. Each therapy session transcript received one numeric score. Thus, each prediction is based on much more language data than our individual messages comprise. Their best model achieves a Spearman rank correlation of .61 using n-gram and psycholinguistic features.

Our contribution goes beyond both of these studies by, first, enriching empathy prediction with personal distress and, second, by annotating and modeling the empathic state actually felt by the writer, instead of relying on external assessments.

5 Conclusion

This contribution was the first to attempt empathy prediction in terms of *multiple* target variables, empathic concern and personal distress. We proposed a novel annotation methodology capturing empathic states actually felt by the author of a statement, instead of relying on third-party assessments. To ensure high reliability in this single-rating setting, we employ multi-item scales in line with best practices in psychology. Hereby we create the first publicly available gold standard for empathy prediction in written language, our survey being set-up and supervised by an expert psychologist. Our analysis shows that the data set excels with high rating reliability and an authentic and diverse language, rich of challenging phenomena such as sarcasm. We provide experimental results for three different predictive models, our CNN turning out superior.

Acknowledgments

References

Muhammad Abdul-Mageed, Anneke Buffone, Hao Peng, Johannes C Eichstaedt, and Lyle H Ungar. 2017. Recognizing pathogenic empathy in social media. In *ICWSM 2017 — Proceedings of the 11th International Conference on Web and Social Media*, pages 448–451, Montreal, Canada, May 15–18, 2017.

Muhammad Abdul-Mageed and Lyle H. Ungar. 2017. EmoNet: Fine-grained emotion detection with gated recurrent neural networks. In *ACL 2017 — Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume 1, long papers, pages 718–728, Vancouver, British Columbia, Canada, July 30 – August 4, 2017.

Firoj Alam, Morena Danieli, and Giuseppe Riccardi. 2017. Annotating and modeling empathy in spoken conversations. *Computer Speech & Language*, pages 40–61.

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.

C Daniel Batson, Karen Sager, Eric Garst, Misook Kang, Kostia Rubchinsky, and Karen Dawson. 1997. Is empathy-induced helping due to self–other merging? *Journal of personality and social psychology*, 73(3):495.

C Daniel Batson and Laura L Shaw. 1991. Evidence for altruism: Toward a pluralism of prosocial motives. *Psychological inquiry*, 2(2):107–122.

Daniel Beck. 2017. Modelling representation noise in emotion analysis using gaussian processes. In *IJC-NLP 2017 — Proceedings of the 8th International Joint Conference on Natural Language Processing*, volume 2, short papers, pages 140–145, Taipei, Taiwan, November 27 – December 1, 2017.

Steven Bird. 2006. NLTK: The natural language toolkit. In *COLING-ACL 2006 — Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, volume 4, interactive presentation sessions, pages 69–72, Sydney, Australia, July 17–21, 2006.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomáš Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5(1):135–146.

Sven Buechel and Udo Hahn. 2017. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *EACL 2017 — Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 2, short papers, pages 578–585, Valencia, Spain, April 3–7, 2017.

Sven Buechel and Udo Hahn. 2018a. Emotion representation mapping for automatic lexicon construction (mostly) performs on human level. In *COLING 2018 — Proceedings of the 27th International Conference on Computational Linguistics*, volume 1, technical papers, pages 2892–2904, Santa Fe, New Mexico, USA, August 20–26, 2018.

Sven Buechel and Udo Hahn. 2018b. Word emotion induction for multiple languages as a deep multi-task learning problem. In *NAACL-HLT 2018 — Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies, volume 1, long papers, pages 1907–1918, New Orleans, Louisiana, USA, June 1–6, 2018.

Lee J Cronbach. 1947. Test reliability: Its meaning and determination. *Psychometrika*, 12(1):1–16.

Benjamin MP Cuff, Sarah J Brown, Laura Taylor, and Douglas J Howat. 2016. Empathy: a review of the concept. *Emotion Review*, 8(2):144–153.

Mark H Davis. 1980. *Interpersonal Reactivity Index*. Edwin Mellen Press.

ED Diener, Robert A Emmons, Randy J Larsen, and Sharon Griffin. 1985. The satisfaction with life scale. *Journal of Personality Assessment*, 49(1):71–75.

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 *NAACL 2016 — Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics*, volume 3, demonstrations, pages 87–91, San Diego, California, USA, June 12–17, 2016.

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 *Interspeech 2015 — Proceedings of the 16th Annual Conference of the International Speech Communication Association*, pages 1947–1951, Dresden, Germany, September 6–10, 2015.

Jennifer L Goetz, Dacher Keltner, and Emiliana Simon-Thomas. 2010. Compassion: an evolutionary analysis and empirical review. *Psychological Bulletin*, 136(3):351.

Samuel D Gosling, Peter J Rentfrow, and William B Swann Jr. 2003. A very brief measure of the big-five personality domains. *Journal of Research in Personality*, 37(6):504–528.

Hamed Khanpour, Cornelia Caragea, and Prakhar Biyani. 2017. Identifying empathetic messages in online health communities. In *IJCNLP 2017 — Proceedings of the 8th International Joint Conference on Natural Language Processing*, volume 2, short papers, pages 246–251, Taipei, Taiwan, November 27 – December 1, 2017.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR 2015 — Proceedings of the 3rd International Conference on Learning Representations*, pages 1–15, San Diego, California, USA, May 7–9, 2015.

Bing Liu. 2015. *Sentiment Analysis: Mining Opinions, Sentiments and Emotions*. Cambridge University Press.

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.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *LREC 2018 — Proceedings of the 11th International Conference on Language Resources and Evaluation*, pages 52–55, Miyazaki, Japan, May 7-12, 2018.

M. Mikulincer and P. R. Shaver, editors. 2010. *Prosocial motives, emotions, and behavior: The better angels of our nature*. American Psychological Association.

Saif Mohammad and Felipe Bravo-Marquez. 2017a. WASSA-2017 shared task on emotion intensity. In *WASSA 2017 — Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis @ EMNLP*, pages 34–49, Copenhagen, Denmark, September 8, 2017.

Saif M. Mohammad and Felipe Bravo-Marquez. 2017b. Emotion intensities in tweets. In **SEM 2017 — Proceedings of the 6th Joint Conference on Lexical and Computational Semantics*, pages 65–77, Vancouver, British Columbia, Canada, August 3–4, 2017.

Sylvia A Morelli, Matthew D Lieberman, and Jamil Zaki. 2015. The emerging study of positive empathy. *Social and Personality Psychology Compass*, 9(2):57–68.

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. 2017. Understanding and predicting empathic behavior in counseling therapy. In *ACL 2017 — Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume 1, long papers, pages 1426–1435, Vancouver, British Columbia, Canada, July 30 – August 4, 2017.

R. W. Picard. 1997. *Affective Computing*. MIT Press.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 Task 4: Sentiment analysis in Twitter. In *SemEval 2017 — Proceedings of the 11th International Workshop on Semantic Evaluation @ ACL*, pages 502–518, Vancouver, British Columbia, Canada, August 3–4, 2017.

H Andrew Schwartz, Salvatore Giorgi, Maarten Sap, Patrick Crutchley, Lyle Ungar, and Johannes Eichstaedt. 2017. DLATK: Differential language analysis toolkit. In *EMNLP 2017 — Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, volume 2, system demonstrations, pages 55–60, Copenhagen, Denmark, September 7–11, 2017.

Elliott Sober and David Sloan Wilson. 1997. Unto others: The evolution of altruism. *Harvard University*.

Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP 2013 — Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October 18–21, 2013.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval 2007 Task 14: Affective text. In *SemEval 2007 — Proceedings of the 4th International Workshop on Semantic Evaluations @ ACL 2007*, pages 70–74, Prague, Czech Republic, June 23–24, 2007.

Jin Wang, Liang-Chih Yu, K. Robert Lai, and Xuejie Zhang. 2016. Dimensional sentiment analysis using a regional CNN-LSTM model. In *ACL 2016 — Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 2, short papers, pages 225–230, Berlin, Germany, August 7–12, 2016.

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 *APSIPA 2012 — Proceedings of the 2012 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1–4, Hollywood, California, USA, December 3–6, 2012.

A Supplemental Material

Details on Stimulus and Instructions

Before being used in our survey, the selected news articles were categorized by the research interns who gathered them in terms of their intensity of suffering (major or minor), cause of suffering (political, human, nature or other), patient of suffering (humans, animals, environment, or other) and scale of suffering (individual or mass). Research interns also provided a short list of key words for each article. This additional information was gathered to examine the influence of these factors on empathy elicitation and modeling performance in later studies.

At the beginning of the survey participants completed background items covering general demographics (including age, gender, and ethnicity), the most commonly used *trait* empathy scale, the Interpersonal Reactivity Index (Davis, 1980), a brief assessment of the Big 5 personality traits (Gosling et al., 2003), life satisfaction (Diener et al., 1985), as well as a brief measure of generalized trust.

After reading each of the articles, participants rated their level of empathic concern and personal distress using multi-item scales. **Figure 2**

shows a cropped screenshot of the survey hosted on `Qualtrics.com`. The first six items (*warm, tender, sympathetic, softhearted, moved*, and *compassionate*) refer to empathy. The last eight items (*worried, upset, troubled, perturbed, grieved, disturbed, alarmed*, and *distressed*) refer to distress.

Figure 2: Multi-item scales for empathic concern and personal distress.

After completing the rating items, participants were instructed to describe their reactions in writing as follows: *Now that you have read this article, please write a message to a friend or friends about your feelings and thoughts regarding the article you just read. This could be a private message to a friend or something you would post on social media. Please do not identify your intended friend(s) — just write your thoughts about the article as if you were communicating with them. Please use between 300 and 800 characters.*

Further Corpus Analyses

The word clouds in **Figure 3** and **Figure 4** show 1-grams of our corpus which correlate significantly (Benjamini-Hochberg corrected $p < .05$) with high empathy and high distress ratings, respectively. In the word clouds, larger size indicates higher correlation and the color scale, gray-blue-red, indicates word frequency, dark red being most prevalent. The Differential Language Analysis Toolkit (Schwartz et al., 2017) was utilized for this analysis. As can be seen, the word clouds display high face-validity, giving further evidence for the soundness of our acquisition methodology.

Figure 3: Word cloud of high empathy 1-grams.

Figure 4: Word cloud of high distress 1-grams.

Figure 5 displays the distribution of the message length of our corpus in tokens. As can be seen the majority of messages contain between 60 and 100 tokens. Yet outliers go up to almost 200. The introduction of a character cap for the writing task proved successful in comparison to a pilot study where this measure has not been in place. In the latter case, the maximum number of tokens was nearly twice as high due to even stronger outliers.
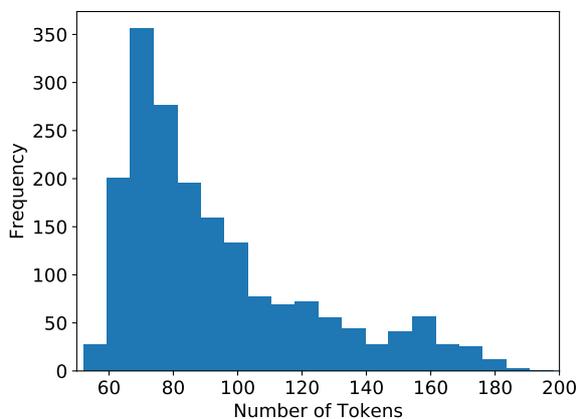

Figure 5: Histogram of message length in our corpus.