The Language of Situational Empathy

KE ZHOU, Nokia Bell Labs & University of Nottingham, UK
LUCA MARIA AIELLO, Nokia Bell Labs & IT University of Copenhagen, UK/Denmark
SANJA ŠĆEPANOVIĆ, Nokia Bell Labs, UK
DANIELE QUERCIA, Nokia Bell Labs & King's College London, UK
SARA KONRATH, Indiana University & University of Notre Dame, USA

Empathy is the tendency to understand and share others' thoughts and feelings. Literature in psychology has shown through surveys potential beneficial implications of empathy. Prior psychology literature showed that a particular type of empathy called "situational empathy"—an immediate empathic response to a triggering situation (e.g., a distressing situation) —is reflected in the language people use in response to the situation. However, this has not so far been properly measured at scale. In this work, we collected 4k textual reactions (and corresponding situational empathy labels) to different stories. Driven by theoretical concepts, we developed computational models to predict situational empathy from text and, in so doing, we built and made available a list of empathy-related words. When applied to Reddit posts and movie transcripts, our models produced results that matched prior theoretical findings, offering evidence of external validity and suggesting its applicability to unstructured data. The capability of measuring proxies for empathy at scale might benefit a variety of areas such as social media, digital healthcare, and workplace well-being.

CCS Concepts: • **Human-centered computing**; • **Computer supported cooperative work**;

Additional Key Words and Phrases: empathy; computational model; lexicon

ACM Reference Format:
Ke Zhou, Luca Maria Aiello, Sanja Šćepanović, Daniele Quercia, and Sara Konrath. 2021. The Language of Situational Empathy. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 13 (April 2021), 19 pages. https://doi.org/10.1145/3449087

1 INTRODUCTION

Empathy is an important psychological process that facilitates human communication and interaction. There are a variety of definitions of empathy, some reflecting more stable traits [12], while others reflect in-the-moment responses [27]. The former is called *trait empathy*, which is a person's chronic disposition to provide empathic responses, whereas the latter is called *situational empathy*, which is an immediate empathic response of a person to a triggering situation. It is important to distinguish between these two types of empathy: *trait empathy* is a long-lasting personal characteristic, and, as such, it may determine the empathic response of the person in a given situation, i.e., his/her *situational empathy*. However, two persons with a similar level of *trait empathy* may still have different empathic responses in a particular situation. That is where

Authors' addresses: Ke Zhou, ke.zhou@nokia-bell-labs.com, Nokia Bell Labs & University of Nottingham, UK; Luca Maria Aiello, Nokia Bell Labs & IT University of Copenhagen, UK/Denmark, lajello@gmail.com; Sanja Šćepanović, Nokia Bell Labs, UK, sanja.scepanovic@nokia-bell-labs.com; Daniele Quercia, Nokia Bell Labs & King's College London, UK, daniele.quercia@nokia-bell-labs.com; Sara Konrath, Indiana University & University of Notre Dame, USA, skonrath@iu.edu.

2573-0142/2021/4-ART13 $15.00
https://doi.org/10.1145/3449087

situational empathy comes into play. Although people with higher trait empathy are more likely to feel compassion (situational empathy) when exposed to others in distress [36], situational empathy is also highly influenced by the context, the emotion, and the level of arousal induced by the situation [15]. Despite these variations, there is a general consensus that empathy consists of both *affective* and *cognitive* aspects. The affective aspect entails sharing somebody else's feelings or feeling compassion for them, while the cognitve aspect includes perspective taking, which involves trying to understand somebody else's internal states, including thoughts and feelings.

Empathy is associated with desirable social outcomes, including an increased ability of conflict resolution in groups [14], higher engagement and learning ability in the classroom (especially in culturally-diverse student cohorts [44]), and higher success rates in therapy and counseling sessions [19]. The most well-known desirable social outcome of empathy is increased prosocial behavior, such as giving and volunteering [3]. Its potential to foster positive social interactions makes empathy important for a range of social media applications: promoting empathy in online conversations could reduce polarization and increase community engagement [60], and could also help target interventions towards people in particularly vulnerable psychological conditions who are in need of support (e.g., young teens, people affected by mental illnesses) [35].

Despite its importance, there are complexities to consider when trying to measure empathy. This is especially true for *trait empathy* since it is an internal mental process that is difficult to gauge by direct observation. In addition, self-reporting of trait empathy might be biased because of issues of social desirability (e.g., trying to "look good") or a potential lack of self-understanding [32]. In contrast, proxies for *situational empathy* may be more easily measured, in part because it is often expressed through language [73]. Because this type of empathy entails *an immediate empathic response elicited by a specific situation*, it can be measured either by asking participants about their experiences immediately after they were exposed to a particular situation, or by various physiological measures such as the measurement of heart rate or skin conductance. However, mainly due to the lack of properly labeled data, very few research projects attempted to study situational empathy in relation to language, and have done so only in the specific setting of therapy sessions [25, 69, 70].

In this work, we leveraged recent advances in machine learning and natural language processing to measure situational empathy from potentially any text, from more structured to unstructured texts. Our focus was on answering the following research question: by mining linguistic and semantic characteristics of the language markers expressed by a person's textual responses, can we accurately predict that person's level of situational empathy? In so doing, we made three main contributions:

- We gathered a corpus of text manually-labeled by the level of situational empathy that it expresses (§3). The data includes three experimental contexts and, overall, contains 13k empathy assessments of about 4k textual messages, which is the largest dataset collected to date.
- We built multiple models to predict situational empathy (§4)[1]. To begin with, we develop a classifier based on features inspired by empathy principles discussed in the Psychology literature (§4.1). Then, we separately trained empathy classifiers on the different data sources coming from our three experimental contexts and determined which are the most discriminative words common to these contexts. In so doing, we created a list containing words that we found to be experimentally related to situational empathy (§4.2). We then tested to which extent that list is generalizable, that is, to what extent it is useful for predicting situational empathy across the three experimental contexts (§5).

[1] https://social-dynamics.net/LanguageEmpathy/

- To check for external validity, we formulated three hypotheses about how empathy unfolds in offline and online conversations, and tested these hypotheses by running a large-scale empathy classification of movie transcripts and Reddit posts (§6).

Our computational models and dictionaries for situational empathy can facilitate an in-depth understanding of empathy at scale and in the wild, enabling applications for promoting empathic communication.

2 RELATED WORK

Over the past decades, a consistent body of research on empathy has emerged, especially in psychology [58]. Psychologists distinguish between measurements of *situational empathy*, where empathy is understood through reactions in a specific situation, and measurements of *trait empathy* (sometimes referred to as *dispositional empathy*), where empathy is understood as a person's stable personality trait. Previous literature has extensively explored the notion of trait empathy. Trait empathy has been traditionally measured through questionnaires [48], either by relying on the reports of others (particularly in case of children) or, most often (in researching empathy in adults), by relying on the administration of various self-report questionnaires associated with specific empathy scales. When deployed online [37], those questionnaires enabled new ways to learn people's predisposition to empathy from their social media footprints, including their friending behavior, their preferences, and their linguistic style [40, 50, 57]. Trait empathy (or lack thereof) has been linked to a number of social processes that manifest online, including burnout [1], emotional contagion [21], and trolling [62].

On the other hand, traditionally, *situational empathy* was measured either by asking participants about their experiences immediately after they were exposed to a particular situation, by studying the "facial, gestural, and vocal indices of empathy-related responding", or by various physiological measures such as the measurement of heart rate or skin conductance [73]. However, the data are difficult to collect and are often of small scale. Fewer studies have focused on situational empathy and on computational models to detect it [71]. Most methods developed to date are based on small-scale conversation data collected in controlled environments. By analyzing discretized facial expressions, gaze, and speech features captured from video, Kumano *et al.* [38] attempted to differentiate states of empathy, unconcern, and antipathy in four-party meetings. Xiao *et al.* [69, 70] used *n*-grams and acoustic features (e.g., pitch, energy, jitter) to detect empathy from therapists' conversations. Moving beyond simple *n*-grams, Lord *et al.* [41] applied LIWC (Linguistic Inquiry and Word Count) on counseling conversations and found that 11 LIWC categories were associated with high empathy sessions, and that the combination of those categories with *n*-grams further improved the ability of predicting empathy ratings [25]. More recently, Buechel et al. [7] developed deep-learning models for predicting both empathy and personal distress on people reacting to reading news stories. Despite its effectiveness, the computational model was not theory-driven and was hard to interpret, and it is not clear how the developed model generalizes to different contexts.

To sum up, most previous work has either looked at trait empathy or used simple models to capture language cues of empathy in specific domains, mainly in therapy sessions. We expanded on previous work by: *i)* focusing on expressions of situational empathy rather than on trait empathy, as the former can more easily be captured through language; and *ii)* investigating a more comprehensive set of linguistic models of empathy motivated by the existing literature in the social sciences. Previous work looked mostly at the association between trait empathy and restricted sets of linguistic features; no work so far has studied expressions of situational empathy in conversations on a large-scale (e.g., on social media data). Yet, it is important to do so to gain a better understanding of the language markers of empathy in day-to-day interactions. In this work,

we aim at providing a computational model that uses the most generalizable language features that can quantify situational empathy across different contexts.

3 COLLECTING TEXTUAL EXPRESSIONS OF EMPATHY

We gathered empathy annotations through three crowdsourcing studies (one via an online survey in a class research project and two via Amazon Mechanical Turk), one was carried out in previous work [7], and two were conducted as a part of this study. The data collection consisted of three main steps, common to all the three studies, and resulted in the generation of three datasets containing empathy annotations for: (a) *news stories*, (b) a *bus bullying* story, and (c) *vent posts* on Reddit. These three scenarios fit the purpose of our study, given that they all reflect events that trigger emotions. We selected these three particular scenarios as they cover different situations that people can respond to. The bus bullying story includes one evolving situation with a visual stimulus (video); the news stories cover various examples of human suffering, focusing mainly on pitiful and sad events; and the vent posts on Reddit contain events with different sentiments, with not only distressing situations (e.g., death of a relative) but also cheerful events (e.g., graduation from college). The diverse set of scenarios enabled us to further investigate the generalizability of our developed models (§4). We ensured that all the participants were paid at least the minimum wage throughout the study. The three steps of our data collection unfolded as follows. First, participants completed a short *pre-study questionnaire* in which they self-reported their age and gender[2]. The age and gender composition for the three datasets is as follows:

> *News stories:* 403 Amazon Mechanical Turk workers (aged 34 years on average, 48% female).
> *Bus bullying:* 558 online survey study participants as part of a class research project, in which undergraduate students recruited their friends and relatives via snowball sample to complete an online survey (aged 29 years on average, 74% female).
> *Vent posts:* 1,204 Amazon Mechanical Turk workers (aged 32 years on average, 57% female).

Second, the participants were exposed to a *situation*, which differed across the three studies:

> *News stories:* 418 stories of human suffering [7], manually selected from popular news outlets[3]; each participant was shown a random selection of 5 news stories.
>
> *Bus bullying:* a shortened 2-minute video[4], which shows a 68-year old bus monitor named Karen Klein bullied by a group of middle schoolers.
>
> *Vent posts:* 10k post-reply pairs from the *r/vent* subreddit [5], an online forum in which users generally voice their opinions or feelings in search of social support. We selected the most popular vent posts by the number of upvotes. For each post, we then selected the top 4 replies that did not contain hate speech or offensive language (detected by the HateSonar [10], a tool for automatic detection of abusive language). We ranked the top-level replies of each

[2]To preserve anonymity, age and gender are the only demographic information we collected.
[3]A wide range of news articles were selected based on the categories in terms of their intensity of suffering (major or minor), cause of suffering (political, human, nature or other), recipient of suffering (humans, animals, environment, or other) and scale of suffering (individual or collective).
[4]The bus bullying episode was caught on video, and received considerable attention from the media shortly after landing on Youtube in 2012: http://www.youtube.com/watch?v=E12R9fMMtos
[5]Reddit is a public discussion website structured in independent communities—called subreddits—dedicated to a broad range of topics [45]. Users can upload posts to a subreddit, write threads of replies to existing posts, and upvote or downvote posts and replies [4]: https://www.reddit.com/r/Vent/.

post by the number of their upvotes and selected the top 4 replies[6]. This procedure yielded a set of 881 posts and 3,524 replies.

Third, participants were instructed to *respond* to the situation and answer a *situational empathy questionnaire*, with eight numeric mood-related measures on a Likert scale from 1 ("not at all") to 5 ("extremely"). These eight measures ("moved", "sympathetic", "compassionate", "tender", "soft-hearted", "caring", "kind", and "warm") reflect people's *situational empathy* [3]. To collect the *textual response* for these situations as well, we asked participants how they felt after reading the situation, and asked them to write their personal responses in a free-text form.

News stories. We asked: "Now that you have read this article, please write a message 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."

Bus bullying: We asked: "What would you say to Karen if you could send her a note or message? Imagine that no one would ever know that you personally sent the note, but that she would actually read it. "

Vent posts: Instead of querying about the participant's mood/empathic reaction, the question in this case was: "Which feelings were likely experienced by the replier when writing his/her reply?". This *assessed* an empathic reaction by the participant on behalf of the Vent replier. We called it *mood assessment*, and we collected the eight mood-related measures similarly to how we collected the self-reported mood for News stories and Bus bullying. For each vent post-reply pair, we asked three crowd-sourcing participants and averaged their empathy/mood assessments. In this case, the participants were not asked to write any textual response; we consider the vent reply itself as the textual response.

Given this procedure, we ultimately obtained three datasets that consist of textual responses to the situations and eight situational empathy measures associated with those responses. A Cronbach's alpha test on the eight empathy-related mood measures yields a score of 0.81, indicating their strong relatedness [64]. As it is common practice in psychology when dealing with multiple survey items measuring the same underlying construct—and as previous work suggested [3]—we averaged the eight empathy measures and obtained an overall *situational empathy score* that reflects a person's momentary empathy level in relation to the specific situation they have been presented with.

The distributions of empathy scores follow normal distributions as assessed by the Shapiro-Wilk test. To obtain a binary and clear-cut distinction between empathic and non-empathic responses, we split the data from each of the crowdsourcing collections in three equally sized bins according to empathy scores, and kept only the top and bottom tertiles to get empathic and non-empathic responses, respectively. We worked on those binarized scores in the remainder of the study. Summary statistics of the datasets are reported in Table 1. Given how we binarized those empathy scores, the dataset was balanced. We can also observe in Table 1 that for textual responses, the average number of words in an empathic response did not deviate much from any response.

There are ethical considerations related to our study, especially some concerning the bus bullying data collection: having the participants watching a video of bullying could be traumatic. This bus bullying study received IRB approval from the university. All researchers and members of the class received training on the ethical conduct of research and signed a confidentiality pledge. Participants were all over the age of 18, gave their consent to be in the study, were anonymous and

[6]This is selected based on the distribution analysis of number of replies per post and reply lengths.

Data	Bus Bullying	News Stories	Vent Posts
Situation	Karen's Video	Textual News	Vent Post-reply Pairs
Empathy Responses	Self-reported Mood	Self-reported Mood	Mood Assessment
Textual Responses	Participant Message	Participant Message	Vent Reply
Num of Participants	558	403	1,204
Num of Situations	1	418	3,524
Num of Responses	558	1,860	10,572
Avg. # Words	63	84	58
Avg. # Words for Empathic Responses	58	86	63
Avg. # Words for Non-Empathic Responses	67	81	54

Table 1. Statistics on our three empathy datasets.

not identifiable, were aware of the study content, and could drop out at any point without penalty. The bullying video, although sensitive, had been shown on the news and is similar to other types of content featured in news media.

4 PREDICTING EMPATHY FROM TEXT

By using the three datasets collected (§3), we first constructed theory-driven computational models to predict empathy scores from textual responses (§4.1). We then generated a dictionary that captures language markers of empathy that are predictive across domains (situations) (§4.2).

4.1 Theory-driven classifiers

To detect the presence of situational empathy in textual responses, we adopted a binary classification approach using logistic regression. We trained the models with three families of independent features (§4.1.1–§4.1.3), which we tested independently and in combination. We did so to predict the dependent variable of situational empathy. Since the interpretation of regression coefficients of those features is sensitive to the scale of the inputs, we follow [24] and divide each numeric predictor by two times their standard deviation. The resulting coefficients are then standardized and directly comparable: such scaling allows the coefficients of numeric predictors to be interpreted in the same way as with binary predictors. Those binary predictors can remain unscaled because their coefficients can already be interpreted directly. For example, a binary predictor with equal probabilities has mean 0.5 and standard deviation 0.5. The coefficients for the binary predictors correspond to a comparison of 0 to 1 (i.e., a 1-unit difference on this transformed scale corresponds to a difference of 1.0 on the original predictor), or two standard deviations. We describe the models next.

4.1.1 Demographic features (control). Some demographics attributes are associated with empathy. Women tend to report more understanding of others' thoughts and feelings [18, 49], and there is a negative association between age and empathy [49, 61]. Therefore, we used age and gender as control variables.

4.1.2 Vocabulary features (Phrases model). As most previous work used bag-of-words approaches to model empathy [71], we considered a model that uses as features the frequencies of the 10k most frequent uni-grams, bi-grams, and tri-grams found in the training data. These short phrases account for the linguistic context in which empathic (or non-empathic) expressions occur.

4.1.3 Linguistic style features. Moving beyond prior literature, we aim to establish the relationship between empathy and textual stylistic properties that previous social science studies linked to empathy:

- *Degree of interdependent thinking.* People with independent self-construals define themselves independently of others, while people with interdependent self-construals, in contrast, define themselves interdependently with their close relationships and social groups. Such self-views play a mediating role in the expression of empathy [26, 42, 68]. Gardner et al. [22] operationalized such self-views through pronoun use; they found that people with independent thinking use more first-person singular, while people with interdependent thinking use more first-person plural pronouns (*"I" Value Freedom, but "We" Value Relationships*). To partially model the degree of interdependent thinking, we simply counted the occurrences of these two types of first-person pronouns.
- *Integrative Complexity.* Integrative Complexity (IC) is a psychometric concept that measures the ability of a person to recognize multiple perspectives and connect them, thus identifying paths for conflict resolution. IC reflects *perspective taking*, a crucial cognitive aspect of empathy [11]. To measure it, we used a computational model developed by Robertson *et al.* ([59]), which scores the level of IC of any text on a scale from 1 to 7 based on its semantic and syntactic structure.
- *Controversy.* When the process of integrating different perspectives fails, controversy arises. Controversy often results in contrast, which reduces the likelihood of empathic responses [14], but it can also originate ethical reflection processes that foster empathy [66]. Prior work compiled a lexicon of controversial words from news articles [46]. To measure controversy, we calculated the fraction of controversial words in the text.
- *Categorical-Dynamic Index.* Categorical language reflects the analytic nature of thinking, as opposed to dynamical language that represents personal narratives that are more conducive of empathic responses. The Categorical-Dynamic Index (CDI) is a measure that combines eight LIWC categories (article, preposition, personal pronoun, impersonal pronoun, auxiliary verb, conjunction, adverb and negation) to capture these two dimensions [53].

We refer to all the above: demographics, phrases, and linguistic features as *theory-driven* features for predicting empathy, given that they are linked to empathy in prior literature. Yet to our knowledge, this is the first time that their effectiveness in predicting situational empathy has been explored.

4.2 Empathy lexicons

The performance of classification models on data from unseen sources depends on the quality and abundance of training data and on such data's representativeness with respect to a more general context. As verbal expressions of empathy might manifest in a wide variety of forms and contexts, a drop in performance is expected when applying a prediction model on a new type of data (as we shall detail in §5). To minimize the cost of collecting domain-specific training data and to keep the generality of the classification high, approaches based on dictionaries or hard-coded general rules have been proposed in the past and successfully used for text classification tasks [30, 54]. Therefore, we experimented with two strategies to create dictionary-based approaches to empathy classification.

4.2.1 Lexicon 1: Empathy synonyms. We put together a simple lexicon that includes the word "empathy" and all its synonyms taken from four English dictionaries: Oxford, Collins, Merriam-Webster, and Google. This is a simple way to compile a list of concepts—curated by dictionary

creators—that are semantically related to empathy. In total, we gathered a union of 75 words related to empathy.[7]

4.2.2 Lexicon 2: Terms by interpolation. One could leverage the knowledge that a classifier learns from datasets of a different nature as a way to bootstrap the creation of a general dictionary. In particular, we are inspired by the approach of multi-source domain adaptation [43] whose objective is to select a feature space in which training data in multiple domains are semantically close, while keeping good performances on the each individual domain.

In our scenario, the logistic regression model trained on n-grams only (§4.1) learned which n-grams are associated with empathic and non-empathic responses. Given two out of the three datasets we collected, we trained such n-gram model on dataset D_1 and D_2 separately, and we used a simple linear interpolation to combine the two coefficient vectors of the two logistic regression models for different n-grams. Effectively, for the case of two vectors, this is equivalent to averaging the coefficients from the two models:

$$w_{interpolation} = \frac{w_{D_1} + w_{D_2}}{2} \tag{1}$$

where w_{D_i} refers to the coefficient vector (the vector of coefficient betas for each n-gram of the logistic regression n-grams model) trained on a specific dataset D_i. Essentially, for each n-gram, we average its coefficient beta of logistic regression models trained on the different datasets. Given the interpolated coefficients of all the n-grams according to $w_{interpolation}$, we then extracted the top 200 n-grams based on the absolute values of the coefficients, i.e., those n-grams that are most correlated (either positively or negatively) with empathy within this interpolated model. We used this interpolated representation as a new lexicon of terms, and we applied the interpolated logistic regression model with this lexicon to the remaining dataset D_3 to test its generalizability. It might be possible to further extend this dictionary by inferring connotations between n-grams within the neural embedding space [20]. However, this data-driven approach requires labeling a large textual dataset with empathy-related annotations, which is costly and beyond this work's scope.

An overview of the dictionary generation approach and its application for empathy prediction is shown in Figure 1.

4.2.3 Processing lexicon: Sparse vs. Dense vs. Count. Given a new lexicon, we turned it into features for logistic regression in two ways.

- *Sparse representation.* We constructed an occurrence-count vector of all the lexicon terms, which corresponds to a sparse and high-dimensional representation of the target text in the semantic space of the lexicon.
- *Dense representation (embeddings).* Rather that using simple word counts, we followed the Distributed Dictionary Representations (DDR) approach [23] and mapped the words into a dense embedding space. Specifically, we averaged the 300-dimensional embedding vectors for each lexicon word using GloVe embeddings [55] trained on the Common Crawl corpus[8].

The above two approaches used a machine learning algorithm on top of our lexicon for the final prediction. To further validate the effectiveness of the lexicon, without using machine learning, we also propose another approach called *lexicon count*: by simply counting the occurrences of empathic and non-empathic n-grams in the lexicon, we deem a text to be empathic if it contains a larger number of empathic expressions than non-empathic n-grams. This *lexicon count* approach

[7] We have also experimented with the strategy of intersection of words from four dictionaries. However, this resulted in a few words and the performance is worse than the union strategy. We adopt the union of empathy synonyms for the rest of the paper.

[8] https://nlp.stanford.edu/projects/glove/

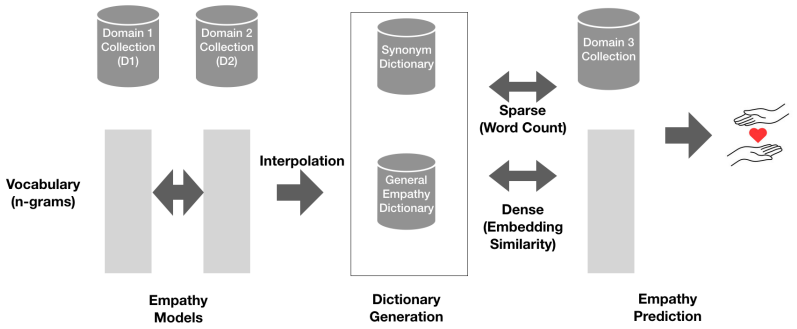

Fig. 1. An overview of the approach for extracting and evaluating the interpolation-based dictionary of empathy-related terms. By interpolating logistic regression models' vocabulary feature coefficients, we generate the lexicon based on *terms by interpolation*, in addition to the empathy synonyms based lexicon curated by dictionary creators. Exploiting one lexicon, we can represent any other texts using sparse or dense representations and utilize machine learning to predict their empathy. The lexicon facilitate the machine learning algorithms to focus on only those empathy-related n-grams that are potentially more general, which may alleviate the problem of over-fitting.

has been extensively adopted in prior work [63]. We apply the above three approaches for both lexicons: empathy synonyms and terms by interpolation.

To sum up, we create two types of lexicons: *empathy synonyms* curated by four existing dictionaries; and *terms by interpolation* by simple linear interpolation of trained logistic regression models' feature coefficients. Exploiting a given lexicon, we can utilize machine learning to quantify empathy of a text through two ways: *sparse representation* by counting occurrences of lexicon terms; and *dense representation* by using distributed dictionary embedding representation. Without machine learning, we also report the *lexicon count* approach that simply counts the occurrences of empathic and non-empathic phrases in the lexicon, and deem a text to be empathic if it contains more empathic expressions.

5 EXPERIMENTAL RESULTS

First, we compared the performance of different families of features in classifying empathy from text (§5.1). We then assessed the ability of different classification models to generalize across datasets (§5.2).

5.1 Classification

On each dataset independently, we performed 5-fold cross validation and recorded the average error rate across folds. Given the balance between the two classes, the error of the random baseline is 0.5. Figure 2 shows the results for the different feature families independently and in combination with the demographic control factors.

When the empathy scores were self-reported by the person who wrote the text (Bus Bullying and News Stories datasets), demographic control variables alone resulted in relatively low error rates compared to other feature sets. This is in line with the literature: demographic factors are good proxies of trait empathy which, naturally, is a strong determinant of the situational empathy that those people exhibited in their written responses [49]. On these two datasets, textual features did not outperform the control variables and only the theory-driven features yielded an improvement consistent across datasets when used in combination with the controls. It is worth noting that in

Fig. 2. Average error rates on 5-fold cross validation for empathy classification on the three datasets, and using different combination of features.

the "Bus Bullying" dataset, adding more variables might even deteriorate the performance (as in the case of "phrases+control", compared to "phrases" or "control"). This can be explained mainly by the fact that the "Bus Bullying" dataset is quite small (see the relatively large error bar in Figure 2) and it is more likely to be overfitting (note the phrase features are sparse).

The prediction of the empathy assessments in Vent Posts produced a different trend. In this case, the situational empathy assessed on the textual response (vent replies) has little to do with the demographic characteristics of the person who assigns the score, which brings the error of the controls-only predictor closer to the random guess. Text-based features performed considerably better than in the other two datasets, with the theory-driven classifier leading the rank with an error rate as little as 0.26. This is expected, not least because training data for "Vent Posts" is one order of magnitude larger; in addition, for "Bus Bullying" and "New Stories", there might be potential discrepancy between the mood reports, i.e., the moods expressed in the textual responses might deviate from the moods self-reported by the respondent. This makes it even more difficult to predict self-reported mood for both "Bus Bullying" and "News Stories" datasets. However, this is not the case for "Vent Posts", since the mood assessments were directly applied on the textual responses (i.e., vent replies).

To explore in-depth the weight of different features in the prediction, we inspected the logistic regression feature coefficients for three types of models (Table 2). For the phrases model, we only report the top 10 n-grams with positive and negative coefficients. When focusing on the control variables only, more empathic responses were given by women and by older people. Integrative Complexity, LIWC variables (e.g., first-person pronouns "I" and "we", which partially reflect degrees of independent thinking), and sentiment (either positive or negative) were the most predictive theory-driven features. As for content features, messages that convey confidence and complexity of thinking, while acknowledging the emotions expressed in the story, were perceived as most empathic. In the model trained on n-grams, "warm-hearted" expressions (e.g., "sorry", "beautiful") were associated with empathic messages, whereas expressions of blame and anger were associated with non-empathic text.

5.2 Cross-domain adaptation

Strikingly, the most predictive n-grams of situational empathy have little overlap across datasets (§5.2), which is a possible expression of the limited ability of the classifiers to adapt to unseen data sources. To measure this limitation quantitatively, we set up a cross-domain adaptation experiment

Feature Importance

Theory + Controls					
Bus Bullying		**News Stories**		**Vent Posts**	
IC	1.21	liwc_IWe	0.80	liwc_posemo	1.15
male	-1.16	male	-0.76	liwc_negemo	1.10
age	0.38	liwc_negemo	0.64	IC	0.95
liwc_IWe	0.23	CDI	0.53	CDI	-0.78
sentiment	-0.17	liwc_posemo	0.40	liwc_IWe	0.59
controversy	0.14	control_age	0.27	controversy	0.33
CDI	-0.14	sentiment	-0.25	sentiment	-0.26
liwc_negemo	0.13	controversy	0.13	age	-0.24
liwc_posemo	0.07	IC	0.11	male	-0.24
Phrases + Controls					
Bus Bullying		**News Stories**		**Vent Posts**	
hope	0.62	family	1.02	**sad**	0.46
mean	0.50	**sad**	0.95	rest	0.41
person	0.42	dying	0.75	**sorry**	0.39
best	0.40	**help**	0.71	mom	0.37
valuable.person	0.33	especially	0.67	know	0.37
sorry	0.33	imagine	0.66	sick	0.36
treated.like	0.32	able	0.64	**don.let**	0.36
upset	0.30	**terrible**	0.62	bad	0.36
beautiful	0.30	things.like	0.59	peace	0.35
ignored	0.29	**sorry**	0.59	awful	0.34
great.person	-0.27	stories	-0.54	sure	-0.27
immature	-0.29	couple	-0.55	god.saying	-0.28
kids.treated	-0.30	cost	-0.57	job	-0.29
wonderful	-0.33	certainly	-0.58	family	-0.29
bus	-0.33	little	-0.60	means	-0.29
need	-0.35	opinion	-0.64	money	-0.30
amazing	-0.36	trump	-0.68	f*ck	-0.35
make	-0.38	comes	-0.69	laugh	-0.36
sure	-0.43	guess	-0.83	friend	-0.37
male (control)	-1.00	don't	-1.10	tell	-0.38
Controls Only					
Bus Bullying		**News Stories**		**Vent Posts**	
age	0.40	age	0.62	age	0.07
male	-1.12	male	-0.40	male	-0.60

Table 2. Logistic regression coefficients of the most discriminative features for empathy classification across three models and for the three datasets considered. All the regression coefficients are standardized by following the procedure described in [24] and are all found to be statistically significant. Positive and negative coefficients respectively represent those predictors that contribute to empathic and non-empathic responses. The bold phrases form part of the lexicon obtained by interpolation of the three datasets.

in which we trained a model on two data collections and tested it on the third one. We kept the class balance both in training and testing with undersampling.

In this experiment, we compared our two dictionary-based approaches with our proposed theory-driven model, the phrases model, and the Convolutional Neural Network (CNN) model proposed in prior work [7]. The CNN approach utilized a publicly available FastText word embeddings as input and adopted three layers (ReLU convolutional, average pooling and ReLU dense layer). This CNN approach performed the best in [7]. The results are presented in Table 3.

	Test data		
	News	Bus Bullying	Vent
Theory	0.433	0.489	0.423
Phrases	0.401	0.395	0.404
CNN [7]	0.424	0.537	0.475
Empathy Synonyms (sparse)	0.475	0.485	0.478
Empathy Synonyms (dense)	0.416	0.440	0.392
Terms by interpolation (sparse)	**0.383**	0.401	0.349
Terms by interpolation (dense)	0.385	**0.389**	**0.343**
Empathy Synonyms (count)	0.479	0.508	0.482
Terms by interpolation (count)	0.412	0.443	0.428

Table 3. Cross-domain error classification of different models. Values refer to the error obtained when testing on the specified data source and training on the remaining two. Top section: logistic regression classifiers trained on theory-informed features and on n-grams (phrases), and CNN classifier based on Fast Text word embedding features [7]. Middle section: logistic regression classifiers trained by phrases extracted from the dictionary-based approaches. Bottom section: classifier based on simply counting empathic and non-empathic phrases in the dictionary without using machine learning algorithms (lexicon count). Best results are highlighted in bold.

As expected, the performance of both the theory-driven and the phrase-based models dropped compared to the cross-validation results obtained by training and testing on a single dataset (Figure 2). The CNN model from previous work fared worst on the Bus Bullying and Vent stories, and performed worse than the dictionaries on the News. The simple dictionary made of empathy synonyms achieved generally a poor performance, yet being significantly better than a random guess (the baseline).

It is interesting to observe that the dictionary extracted from the interpolation approach yielded the lowest error in cross-domain classification, especially when using a dense, embedding-based representation of words. This is in line with previous studies that demonstrated the effectiveness of semantic similarity encoded in low-dimensional embedding spaces [23]. This demonstrates the generalizability of using *terms by interpolation* for empathy classification, showing that it can work across different domains, outperforming those classifiers trained to specific domains.

Lastly, we demonstrate that although not as competitive as using embedding-based representation on the dictionary, classifying text based on simply counting the empathic and non-empathic phrases in our interpolation based lexicon can result in a moderate performance that is significantly better than the baseline. Several examples of the lexicon phrases can be found in Table 2 (bold features). For example, in Table 2, we can observe that one of the lexicon phrase "sorry" is shown to be one of the top phrases that positively contribute to predicting empathic responses across all the three datasets. Despite of the minor differences observed on the feature coefficients across datasets, this example demonstrates the potential generalizability of our lexicon on classifying empathic textual responses.

To sum up, if building an empathy classifier in a new domain is required, for which you have labelled data and demographics, it is recommended to utilize both demographics information (§4.1.1) and those theory-driven features (§4.1.3) to build such machine learning classifier. When demographics information is not available, the phrases approach (§4.1.2) is the best approach to train such classifier. However, the performances of such trained classifiers deteriorate significantly when we try to predict in a new domain (i.e., different situations for empathy responses).

When labelled data is unavailable or a classifier that would work well across different situations is required, it is preferably to use our lexicon created by interpolation on a dense, embedding-based representation (§4.2). This would provide a more generalizable empathy classifier that achieves the competitive performance across domains.

6 EVALUATION OF EXTERNAL VALIDITY

To test the external validity of our model and to show its applicability on large scale text data from the Web, we collected around 3.2M messages from two data sources (§6.1), formulated hypotheses about the expected trends of empathy in such datasets, and verified them by estimating the level of situational empathy they express (§6.2). This is the first study that measures situational empathy on a large scale.

6.1 Data sources

6.1.1 Movie scripts. The Cornell Movie-Dialogs Corpus is one of the most comprehensive open collections of movie scripts, containing 304,713 utterances exchanged between 10,292 pairs of characters from 617 movies. Movies have a great influence over childrens' behavior [33], and this is why the plots of kids movies are often intentionally crafted to promote altruism and empathy [67]. Examples of such movies range from classics, such as"To Kill a Mockingbird", to modern kid-friendly favorites like "Inside Out" and "Zootopia". We selected 23 education movies[9] (for a total of 28k lines) and as many randomly-selected movies for comparison (for a total of 17k lines).

6.1.2 Reddit. Using the Reddit API, we collected all the posts and comments made between January 2018 and June 2019, in four subreddits: r/vent, the forum we used in the evaluation; r/depression, a community dedicated to gather support around stories of mental health; r/technology and r/science, two communities focused on knowledge exchange. In total, we collected 15k posts and 37k comments from r/vent, 40k posts and 143k comments from r/science, and 141k posts and 1.7M comments from r/technology.

6.2 Hypothesis validation

We set out to test three hypotheses:

H1: Movies targeted to younger audiences, especially children, tend to promote messages of altruism and empathy [65, 67]. We therefore hypothesized that the scripts of movies for kids exhibit a higher level of empathy compared to a random selection of movies. We focused on the content of the messages conveyed by the characters in the movie rather than on the post-impression of people reviewing the movie, which are biased by several confounders that are hard to control for.

H2: Witnessing people in distress triggers empathic responses, which are more likely to be verbalized if the context is perceived as safe [52]. r/depression is among the most popular social support communities in Reddit. It has been shown in prior study [13] that there is abundance of social supports and empathy in such forum. We hypothesized that people posting in r/depression exhibit higher level of empathy compared to discussion spaces aimed at sharing technical knowledge like r/technology and r/science.

H3: Stories characterized by distress and sadness tend to trigger stronger and more frequent empathic responses than joyful stories [56]. We hypothesized that stories in r/vent characterized predominantly by distress receive more empathic comments than those conveying joy.

[9] https://www.commonsensemedia.org/lists/movies-that-inspire-empathy

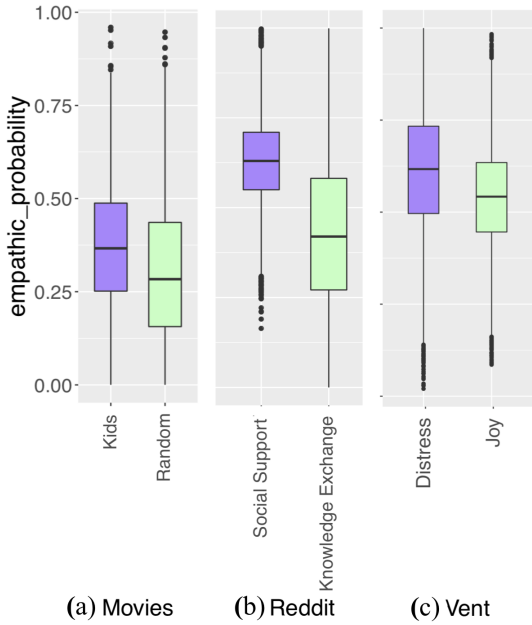

Fig. 3. Distribution of the likelihood p_e of a text to be empathic, across different data collections and partitions: (a) movie transcripts of kids movies vs. a random selection of movies; (b) messages from the support subreddit r/depression vs. messages from the knowledge exchange subreddits r/science and r/technology; (c) messages posted on r/vent characterized by joy vs. those with distress. The distributions of pairs of groups are all statistically different according to a two-tailed t-test ($p < 0.01$).

We ran our each text m through our empathy classifier that were trained on theory-driven and interpolated dictionary features, and obtained the probability $p_e(m)$ of the text containing an expression of empathy, according to the logistic regression.

To verify H1, we calculated the distribution of p_e for kids movies and for the random selection of movies separately. Results support our hypothesis (Figure 3a): even though the overall likelihood of empathy is rather low across all movies, movie scripts for children show significantly more markers of empathy than other type of pictures (median p_e 0.36 against 0.23, with 13% difference).

Similarly, we tested H2 by comparing the distribution of p_e calculated from the posts and comments in r/depression with that from post and comments in r/technology and r/science. Results again met our expectation, as the typical depression-related message was roughly 25% more empathic compared to knowledge-related posts (Figure 3b).

Last, to validate H3, we first computed the emotions of all messages from r/vent using Emolex (joy) [47] and LIWC (anxiety and sadness) [63] to quantify joy and distress respectively. We then partitioned the dataset into two sets of texts with events of joy and distress. We then computed the distribution of p_e in the two groups and found, as expected, that distress-related posts are 11% more likely to be empathic than the ones with joy (Figure 3c).

Note that for all the above results shown, we conducted statistical two-tailed t-test to confirm whether the means of the two groups are significantly different. We indeed found all the differences observed on the three hypothesis are statistically different ($p < 0.01$). This further validates our empathy classifier and demonstrates its applicability to large-scale text data on the Web.

7 CONCLUSION

Driven by theoretical work in psychology, we developed computational models to score text according to their situational empathy—the empathic response expressed in relation to a specific situation. Our models capture both empathic concern in reaction to negative events, and empathy triggered by positive stories. By looking at the commonalities between empathic expressions in multiple domains, we created a vocabulary of empathy-related words that generalize across domains to a good extent. To support the applicability of our method, we ran the first large-scale study of empathy on digital data and successfully matched the results with theoretical expectations.

Our results bring forth several theoretical and practical implications, and have limitations that future work might address.

7.1 Implications

From the theoretical standpoint, we have shown that situational empathy is expressed verbally through distinctive textual markers characterized by a combination of expression of sentiment, complexity of thinking, and demographic characteristics. In the future, automatic classification of empathy could enhance the descriptive power of textual analytics like sentiment analysis. Text characterized by positive sentiment may contain superficial exchanges that lack empathy, or deeper and supportive messages that convey empathy. Similarly, negative sentiment text could be the result of conflict and lack of support but, when it expresses empathy, it may actually reflect an attempt to overcome sad or hurtful events.

From a practical perspective, we created a generalizable tool to capture empathy at scale that is applicable to many types of text. To aid this process, we made available the empathy dictionaries and our crowdsourced data, and we encourage researchers to experiment with it. Our user studies and corresponding guidelines could also help researchers to collect empathy data in other situations. This new capability opens up the way to a number of new application domains.

Empathy analytics on social media. Our classifiers could contribute to creating new text analytics tools for large-scale social media data. In particular, we believe that the analysis of empathy expressed in social networking sites could help to unearth pockets of social ill-being associated with the decline of empathy [35] as well as promoting empathic content that could reduce conflict and bridge community disconnects [60].

Supporting digital healthcare. Online applications have gained an increasingly important role among the service offered by healthcare providers [2]. For example, when dealing with mental health issues, counseling and psychotherapy are becoming increasingly mediated by digital tools [9]. Online interactions between patient and therapist could be augmented by automatic triggers that foster empathic responses, which in the long run result in higher success rates of the therapy process [41].

Fostering success of communities with a purpose. In goal-oriented communities such as a working group or a class of students, empathy fosters positive engagement, triggers virtuous circles of gratitude, and results into better outcomes and higher chances of collective success [6, 44]. In a future where social interactions in bounded environments (e.g., office, schools) could be recorded in a privacy-preserving fashion, empathy analytics could be used to act upon situations in which lack of empathic exchanges is detected.

7.2 Limitations

Despite our effort to leverage multiple data sources to create a context-independent dictionary of empathy-related words, our datasets suffer from a number of biases. The vast majority of messages

we used for training and testing are written by US residents and are labeled by annotators residing in English-speaking countries, which might generate a biased representation of the expressions that are considered empathic. In addition, the context (e.g., public *vs.* private channels) in which users were instructed to write the textual response might influence the degree of emotions they expressed. Previous research studies found that, in social media, people shared more intense and negative emotions in private messages than in public channels [5], whereas, in online support groups, people revealed more negative self-disclosure in public channels [72]. The impact of audience channel on empathy is an open research question and requires further investigation. When obtaining annotations, we attempted to cover both private and public channels (private channels with Bus Bullying, and with part of News stories; public channels with the other part of News stories, and with Vent posts). Overall, the labeled data we collected is the largest to date, yet it is relatively small when it comes to training machine learning algorithms. Therefore, larger data collections with reduced socio-demographic and cultural biases are in order.

In our theory-driven model, we considered a number of linguistic style features that capture concepts related to empathy. This set is not meant to be exhaustive though, and future work could explore additional features that may further improve classification performance. Also, other deep-learning approaches should be explored. By exploiting a pre-trained BERT model [16], and performing empathy classifications (§5.1), we found that compared to theory-driven classifiers, BERT could slightly improve the classification errors for two datasets: news stories (4% improvement) and vent posts (7% improvement), but not on Bus bullying dataset (3% degradation). Given that the performance improvement of using BERT was small and not consistent across datasets, it was preferable to adopt the theory-driven models that are more interpretable without compromising accuracy.

We focused on situational empathy and controlled for trait empathy—a person's intrinsic predisposition to provide empathic responses—only based on two demographic indicators (age and gender). In the future, we envision a more systematic analysis of the relationship between trait empathy and situational empathy, starting from assessing the ability of our models to predict trait empathy from text.

All our datasets were collected through crowdsourcing rather than through traditional lab studies. Crowdsourcing enables data collection at scale [17]. Compared to traditional lab studies, which normally recruit dozens of students, crowdsourcing offers fast access to a relatively large and diverse set of research participants. On the other hand, it is harder to control experimental conditions. For example, experimental tasks that require rigorous programming (e.g., the measurement of reaction time) are hard to do online. In addition, given the nature of the crowdsourcing experiments, participants' attention to the experimental task - despite being important - is hard to measure. As a result, crowdsourcing generally requires various mechanisms to ensure that the data being collected is of quality [28]. However, if those quality-control checks are in place, crowdsourcing turns out to be an effective way of gathering large-scale responses [8].

Our results are not intended to amplify existing social stereotypes about gender, age, and empathy. Previous research has indeed found that trait empathy is higher among women and older people [51]. Yet most of the research on such differences relies on individuals' self-reported agreement on dispositional measures, which is subject to several issues (e.g., lack of self-knowledge, socially desirable responding). Concerning gender, research finds that gender differences in empathy are smaller in objective measures (e.g., empathic accuracy, emotion reading) and are more likely when participants are aware of gender-role expectations [31, 34]. Both of these suggest that men and women score differently in empathy as a way of fulfilling gender role expectations. Still, since the current study relies on self-reported emotional responses, including gender in our model is

justifiable. With respect to age, research also finds that age differences depend upon the measure used and individuals' motivation [29, 39].

REFERENCES

[1] Muhammad M Abdul-Mageed, Anneke Buffone, Hao Peng, Johannes Eichstaedt, and Lyle Ungar. 2017. Recognizing pathogenic empathy in social media. In *Eleventh International AAAI Conference on Web and Social Media*.

[2] Ritu Agarwal, Guodong Gao, Catherine DesRoches, and Ashish K Jha. 2010. Research commentary—The digital transformation of healthcare: Current status and the road ahead. *Information Systems Research* 21, 4 (2010), 796–809.

[3] Charles Daniel Batson. 2011. *Altruism in humans*. Oxford University Press, USA.

[4] Jason Baumgartner. 2015. *I have every publicly available Reddit comment for research. 1.7 billion comments 250 GB compressed. Any interest in this?* https://redd.it/3bxlg7

[5] Natalya N Bazarova, Yoon Hyung Choi, Victoria Schwanda Sosik, Dan Cosley, and Janis Whitlock. 2015. Social sharing of emotions on Facebook: Channel differences, satisfaction, and replies. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*. 154–164.

[6] Giacomo Bono, Robert A Emmons, and Michael E McCullough. 2004. Gratitude in practice and the practice of gratitude. *Positive psychology in practice* (2004), 464–481.

[7] Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and Joao Sedoc. 2018. Modeling Empathy and Distress in Reaction to News Stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 4758–4765.

[8] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. 2016. Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data? (2016).

[9] Ronan Cummins, Michael P Ewbank, Alan Martin, Valentin Tablan, Ana Catarino, and Andrew D Blackwell. 2019. TIM: A Tool for Gaining Insights into Psychotherapy. In *The World Wide Web Conference*. ACM, 3503–3506.

[10] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*.

[11] Mark H Davis. 1983. Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of personality and social psychology* 44, 1 (1983), 113.

[12] Mark H Davis et al. 1980. A multidimensional approach to individual differences in empathy. (1980).

[13] Munmun De Choudhury and Sushovan De. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Eighth international AAAI conference on weblogs and social media*.

[14] Minet De Wied, Susan JT Branje, and Wim HJ Meeus. 2007. Empathy and conflict resolution in friendship relations among adolescents. *Aggressive Behavior: Official Journal of the International Society for Research on Aggression* 33, 1 (2007), 48–55.

[15] Jean Ed Decety and William Ed Ickes. 2009. *The social neuroscience of empathy*. MIT Press.

[16] 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* (2018).

[17] Jennifer Edgar, Joe Murphy, and Michael Keating. 2016. Comparing traditional and crowdsourcing methods for pretesting survey questions. *Sage Open* 6, 4 (2016), 2158244016671770.

[18] Nancy Eisenberg and Randy Lennon. 1983. Sex differences in empathy and related capacities. *Psychological bulletin* 94, 1 (1983), 100.

[19] Robert Elliott, Arthur C Bohart, Jeanne C Watson, and Leslie S Greenberg. 2011. Empathy. *Psychotherapy* 48, 1 (2011), 43.

[20] Ethan Fast, Binbin Chen, and Michael S Bernstein. 2016. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 4647–4657.

[21] Emilio Ferrara and Zeyao Yang. 2015. Measuring emotional contagion in social media. *PloS one* 10, 11 (2015), e0142390.

[22] Wendi L Gardner, Shira Gabriel, and Angela Y Lee. 1999. "I" value freedom, but "we" value relationships: Self-construal priming mirrors cultural differences in judgment. *Psychological science* 10, 4 (1999), 321–326.

[23] Justin Garten, Joe Hoover, Kate M Johnson, Reihane Boghrati, Carol Iskiwitch, and Morteza Dehghani. 2018. Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis. *Behavior research methods* 50, 1 (2018), 344–361.

[24] Andrew Gelman. 2008. Scaling regression inputs by dividing by two standard deviations. *Statistics in medicine* 27, 15 (2008), 2865–2873.

[25] 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*.

[26] Shihui Han and Glyn Humphreys. 2016. Self-construal: A cultural framework for brain function. *Current Opinion in Psychology* 8 (2016), 10–14.

[27] Martin L Hoffman. 2001. *Empathy and moral development: Implications for caring and justice*. Cambridge University Press.

[28] Pei-Yun Hsueh, Prem Melville, and Vikas Sindhwani. 2009. Data quality from crowdsourcing: a study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 workshop on active learning for natural language processing*. 27–35.

[29] Isabell Hühnel, Mara Fölster, Katja Werheid, and Ursula Hess. 2014. Empathic reactions of younger and older adults: No age related decline in affective responding. *Journal of Experimental Social Psychology* 50 (2014), 136–143.

[30] Clayton J Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*.

[31] William Ickes, Paul R Gesn, and Tiffany Graham. 2000. Gender differences in empathic accuracy: Differential ability or differential motivation? *Personal Relationships* 7, 1 (2000), 95–109.

[32] Nicole Kämpfe, Jan Penzhorn, Julia Schikora, Julia Dünzl, and Jane Schneidenbach. 2009. Empathy and social desirability: a comparison of delinquent and non-delinquent participants using direct and indirect measures. *Psychology, Crime & Law* 15, 1 (2009), 1–17.

[33] Marsha Kinder. 1991. *Playing with power in movies, television, and video games: from Muppet Babies to Teenage Mutant Ninja Turtles*. Univ of California Press.

[34] Kristi JK Klein and Sara D Hodges. 2001. Gender differences, motivation, and empathic accuracy: When it pays to understand. *Personality and Social Psychology Bulletin* 27, 6 (2001), 720–730.

[35] Sara Konrath. 2013. The empathy paradox: Increasing disconnection in the age of increasing connection. In *Handbook of research on technoself: Identity in a technological society*. IGI Global, 204–228.

[36] Sara Konrath, Brian P Meier, and Brad J Bushman. 2018. Development and validation of the single item trait empathy scale (SITES). *Journal of research in personality* 73 (2018), 111–122.

[37] Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences* 110, 15 (2013), 5802–5805.

[38] Shiro Kumano, Kazuhiro Otsuka, Dan Mikami, and Junji Yamato. 2011. Analyzing empathetic interactions based on the probabilistic modeling of the co-occurrence patterns of facial expressions in group meetings. In *Face and Gesture 2011*. IEEE, 43–50.

[39] Serena Lecce, Irene Ceccato, Federica Bianco, Alessia Rosi, Sara Bottiroli, and Elena Cavallini. 2017. Theory of Mind and social relationships in older adults: the role of social motivation. *Aging & Mental Health* 21, 3 (2017), 253–258.

[40] Marina Litvak, Jahna Otterbacher, Chee Siang Ang, and David Atkins. 2016. Social and linguistic behavior and its correlation to trait empathy. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*. 128–137.

[41] Sarah Peregrine Lord, Elisa Sheng, Zac E Imel, John Baer, and David C Atkins. 2015. More than reflections: empathy in motivational interviewing includes language style synchrony between therapist and client. *Behavior therapy* 46, 3 (2015), 296–303.

[42] Christine Ma-Kellams and Jim Blascovich. 2012. Inferring the emotions of friends versus strangers: The role of culture and self-construal. *Personality and Social Psychology Bulletin* 38, 7 (2012), 933–945.

[43] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. 2009. Domain adaptation with multiple sources. In *Advances in neural information processing systems*. 1041–1048.

[44] Gretchen McAllister and Jacqueline Jordan Irvine. 2002. The role of empathy in teaching culturally diverse students: A qualitative study of teachers? beliefs. *Journal of teacher education* 53, 5 (2002), 433–443.

[45] Alexey N Medvedev, Renaud Lambiotte, and Jean-Charles Delvenne. 2017. The anatomy of Reddit: An overview of academic research. In *Dynamics on and of Complex Networks*. Springer, 183–204.

[46] Yelena Mejova, Amy X Zhang, Nicholas Diakopoulos, and Carlos Castillo. 2014. Controversy and sentiment in online news. *arXiv preprint arXiv:1409.8152* (2014).

[47] Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence* 29, 3 (2013), 436–465.

[48] Sylvia A Morelli, Desmond C Ong, Rucha Makati, Matthew O Jackson, and Jamil Zaki. 2017. Empathy and well-being correlate with centrality in different social networks. *Proceedings of the National Academy of Sciences* 114, 37 (2017), 9843–9847.

[49] Ed O'brien, Sara H Konrath, Daniel Grühn, and Anna Linda Hagen. 2012. Empathic concern and perspective taking: Linear and quadratic effects of age across the adult life span. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences* 68, 2 (2012), 168–175.

[50] Jahna Otterbacher, Chee Siang Ang, Marina Litvak, and David Atkins. 2017. Show me you care: Trait empathy, linguistic style, and mimicry on Facebook. *ACM Transactions on Internet Technology (TOIT)* 17, 1 (2017), 6.

[51] Ed O'Brien, Sara H Konrath, Daniel Grühn, and Anna Linda Hagen. 2013. Empathic concern and perspective taking: Linear and quadratic effects of age across the adult life span. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences* 68, 2 (2013), 168–175.
[52] Kyung Hye Park, Dong-hee Kim, Seok Kyoung Kim, Young Hoon Yi, Jae Hoon Jeong, Jiun Chae, Jiyeon Hwang, and HyeRin Roh. 2015. The relationships between empathy, stress and social support among medical students. *International journal of medical education* 6 (2015), 103.
[53] James W Pennebaker, Cindy K Chung, Joey Frazee, Gary M Lavergne, and David I Beaver. 2014. When small words foretell academic success: The case of college admissions essays. *PloS one* 9, 12 (2014), e115844.
[54] James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 71, 2001 (2001), 2001.
[55] 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)*. 1532–1543.
[56] Daniella Perry, Talma Hendler, and Simone G Shamay-Tsoory. 2011. Can we share the joy of others? Empathic neural responses to distress vs joy. *Social cognitive and affective neuroscience* 7, 8 (2011), 909–916.
[57] Marco Polignano, Pierpaolo Basile, Gaetano Rossiello, Marco de Gemmis, and Giovanni Semeraro. 2017. Learning inclination to empathy from social media footprints. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*. ACM, 383–384.
[58] Stephanie D Preston and Frans BM De Waal. 2002. Empathy: Its ultimate and proximate bases. *Behavioral and brain sciences* 25, 1 (2002), 1–20.
[59] Alexander Robertson, Luca Maria Aiello, and Daniele Quercia. 2019. The Language of Dialogue Is Complex. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13. 428–439.
[60] Michelle Rodino-Colocino. 2018. Me too,# MeToo: Countering cruelty with empathy. *Communication and Critical/Cultural Studies* 15, 1 (2018), 96–100.
[61] Scott Schieman and Karen Van Gundy. 2000. The personal and social links between age and self-reported empathy. *Social Psychology Quarterly* (2000), 152–174.
[62] Natalie Sest and Evita March. 2017. Constructing the cyber-troll: Psychopathy, sadism, and empathy. *Personality and Individual Differences* 119 (2017), 69–72.
[63] Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology* 29, 1 (2010), 24–54.
[64] Mohsen Tavakol and Reg Dennick. 2011. Making sense of Cronbach's alpha. *International journal of medical education* 2 (2011), 53.
[65] Ashton D Trice and Hunter W Greer. 2019. *The Psychology of Moviegoing: Choosing, Viewing and Being Influenced by Films*. McFarland.
[66] Linus Vanlaere, Trees Coucke, and Chris Gastmans. 2010. Experiential learning of empathy in a care-ethics lab. *Nursing Ethics* 17, 3 (2010), 325–336.
[67] Barbara J Wilson. 2008. Media and children's aggression, fear, and altruism. *The future of children* 18, 1 (2008), 87–118.
[68] Karl-Andrew Woltin, Vincent Y Yzerbyt, and Olivier Corneille. 2011. On reducing an empathy gap: The impact of self-construal and order of judgment. *British Journal of Social Psychology* 50, 3 (2011), 553–562.
[69] Bo Xiao, Daniel Bone, Maarten Van Segbroeck, Zac E Imel, David C Atkins, Panayiotis G Georgiou, and Shrikanth S Narayanan. 2014. Modeling therapist empathy through prosody in drug addiction counseling. In *Fifteenth Annual Conference of the International Speech Communication Association*.
[70] 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*. IEEE, 1–4.
[71] Bo Xiao, Zac E Imel, Panayiotis Georgiou, David C Atkins, and Shrikanth S Narayanan. 2016. Computational analysis and simulation of empathic behaviors: A survey of empathy modeling with behavioral signal processing framework. *Current psychiatry reports* 18, 5 (2016), 49.
[72] Diyi Yang, Zheng Yao, Joseph Seering, and Robert Kraut. 2019. The channel matters: Self-disclosure, reciprocity and social support in online cancer support groups. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–15.
[73] Qing Zhou, Carlos Valiente, and Nancy Eisenberg. 2003. Empathy and its measurement. (2003).

Received June 2020; revised October 2020; accepted December 2020