Recognizing Pathogenic Empathy in Social Media

Muhammad Abdul-Mageed,[1] **Anneke Buffone,**[2] **Hao Peng,**[3]
Salvatore Giorgi,[2] **Johannes Eichstaedt,**[2] **Lyle Ungar**[4]
[1]School of Library, Archival and Information Studies, University of British Columbia
[2]Department of Psychology, University of Pennsylvania
[3]School of Informatics and Computing, Indiana University Bloomington
[4] Computer and Information Science, University of Pennsylvania
muhammad.mageed@ubc.edu

Abstract

Empathy is an integral part of human social life, as people care about and for others who experience adversity. However, a specific "pathogenic" form of empathy, marked by automatic contagion of negative emotions, can lead to stress and burnout. This is particularly detrimental for individuals in caregiving professions who experience empathic states more frequently, because it can result in illness and high costs for health systems. Automatically recognizing pathogenic empathy from text is potentially valuable to identify at-risk individuals and monitor burnout risk in caregiving populations. We build a model to predict this type of empathy from social media language on a data set we collected of users' Facebook posts and their answers to a new questionnaire measuring empathy. We obtain promising results in identifying individuals' empathetic states from their social media (Pearson *r* = 0.252, p $<$ 0.003).

1 Introduction

Empathy is a fundamental motivational force in a species as interdependent as ours. Every day we are confronted with other people's need or suffering. The extent to which we respond to others' suffering can depend on the situation, but also on the person. Research in psychology shows that empathy affects health and well-being. On the one hand, empathy manifests behaviorally as helping behavior, which has been linked to improved health and even reduced mortality (Raposa, Laws, and Ansell 2015; Riess et al. 2012). On the other hand, particularly for those who are faced with human suffering professionally and therefore more frequently and intensely, the experience of empathic emotions has been found to be exhausting and overwhelming, and a potential source of stress and burnout (Manczak, DeLongis, and Chen 2016). According to (Buffone et al. in prep), momentary empathic states as well as more stable character dispositions for empathy can be either salutogenic (health promoting) or pathogenic (health demoting). We were able to build language-based models that can discern health-promoting from health-demoting empathic personality types. We focus our attention in the current work on characterizing and recognizing individuals high in pathogenic empathic disposition, as a first step in the computational treatment of this important psychological state and hence identifying vulnerable populations, and targeting interventions.

2 Background and Related Work

There has been only minimal computational work on empathy, with different definitions and operationalizations: (Litvak et al. 2016; Fung et al. 2016; Collins 2014) focus on 'empathetic concern' or the sharing others' emotions in a conversation, and (Xiao et al. 2015; Gibson et al. 2016; Xiao et al. 2016) work on modeling 'empathy' in psychotherapy sessions, defined as the ability of a therapist to attune to the emotions of their clients. Our work is different in that we focus on a specific type of empathy, i.e., pathogenic empathy (Buffone et al. in prep) and hence we ground our work in the psychological literature. In doing so we avoid collapsing over beneficial and detrimental forms of empathy to detect individuals for whom empathy can be a hazard to their health and who may benefit from developing more health-promoting forms of empathy. By conceptualizing pathogenic empathy as a trait-like psychological construct, our study is related to previous efforts to model personality (Schwartz et al. 2013) and mental health (Preotiuc-Pietro et al. 2015) from social data, as well as more state-like psychological experiences like mood (e.g., (Bollen, Mao, and Zeng 2011)) and emotion (e.g., (Mohammad and Kiritchenko 2015)).

The focus of our work is straightforward: Given the proliferation of social media and the abundance of user-generated psychological content available therein, we investigate the utility of mining Facebook data to detect a specific empathic personality trait. More specifically, we use a survey designed by psychologists to identify pathogenic empathy (Buffone et al. in prep). *Pathogenic empathy* is a state in which the potential helper automatically "catches" the emotions of the suffering other, and makes the suffering other's need his/her own, and thus emotionally and cognitively merges with the suffering other. This state is elicited when one imagines oneself in the other's shoes without maintaining a healthy level of self-other separation (Batson, Early, and Salvarani 1997). Emotionally, this state is thought to be marked by self-focused distress and anxiety alongside of empathic concern (Batson, Early, and Salvarani 1997). Pathogenic empathy has been shown to be experienced as physiologically and psychologically stressful (Batson, Early, and Salvarani

1997; Buffone 2015).

2.1 Setup and Motivation

Our goal is to recognize (in a machine learning sense) pathogenic empathy from social media language, thus finding a way to assess it in large populations and providing the basis for interventions ameliorating the stress, burnout, and exhaustion common in professional and lay caregivers. Toward that end, we obtain user scores on pathogenic empathy and authorization to access users' Facebook posts. We can thus build regression models to predict empathy from written text. To the best of our knowledge, the current paper is the first attempt to recognize the concept from large-scale social data. More concretely, it is also the first paper to be able to identify specifically the kind of empathy that is risky for health and well-being.

This work has applications related to both machine-enabled intervention and human-machine interaction. For intervention, creating machines that are able to detect pathogenic empathy can be used as a basis for identifying individuals (e.g., healthcare providers who might be at risk for burnout) in need of help. Early detection of pathogenic empathy in providers may help prevent burnout or reduce its severity e.g., via personalized interventions (Prot et al. 2014). Regarding human-machine interaction, creating machines that are able to 'empathize' (via e.g., language generation) with suffering individuals can make these individuals feel better. In addition, machines capable of recognizing empathy can help in the design of affective chatbots and automated therapists (Xiao et al. 2015). That is, machines may provide an important and beneficial kind of empathic support, the kind of empathy that is pathogenic for human providers, but not costly for a machine. Overall, we make the following contributions: (1) We introduce a novel dataset (as described in more detail in (Buffone et al. in prep)) to identify pathogenic empathy via user's questionnaire responses and Facebook posts, which enables us to detect a concept that is otherwise difficult to investigate, and (2) We build machine learning models to recognize the intensity of user pathogenic empathy.

3 Data

Our purpose is to be able to detect the extent to which a person is pathologically empathetic. For this *user-level* empathy, we introduce a dataset and survey on pathogenic empathy (Buffone et al. in prep). The survey assesses pathogenic empathy with questions which are based on a large volume of psychological research from various subdisciplines in psychology, e.g., (Batson, Early, and Salvarani 1997; Buffone 2015; Jordan, Amir, and Bloom 2016), to elicit each type of empathy. Some of these items were newly generated, while some were taken from a recent empathy scale with the researcher's permission (Jordan, Amir, and Bloom 2016). Participants responded to items # 1-5 in the survey on a 1-9 scale (1= not at all like me − 9= very much like me) and items # 6-8 (which are based on (Jordan, Amir, and Bloom 2016)) on a 1-7 scale (1= not at all like me − 7= very much like me). Each participant received a pathogenic empathy score based on the average of the standardized items provided. These scores formed our ground truth for the regression task. The survey included an integrated app grabbing participants' Facebook posts, likes, and basic profile information of consenting users (e.g. age, gender). The sample was recruited via Qualtrics.[1] Facebook users were only allowed to complete the survey if they had posted at least five times in the last 30 days, and had at least 100 lifetime posts. In total, we had data available for 2,405 users, among whom 71.43% (N=1718) are female, 28.27% (N=680) are male, and rest (%=0.0029, N=7) identified as "other." After filtering non-English posts, the data have a total of 1,835,884 Facebook posts, an average of 913 posts per user.

4 Predictive Models of Empathy

We view empathy as a continuous variable and model it with a regression setup. To account for any multicollinearity in the feature sets, we employ ridge regression with a wide range of α values between 0.1 and 100,000 and identify the best value with cross-validation. For all our experiments, we typically run with 10-fold cross-validation. We run a very extensive set of experiments, across a wide range of settings. We present each of these sets next.

Word N-Grams: To capture the local semantics (and syntax) of the data, we run experiments with n-grams of $N \leq 3$ and combinations of these. In each case, we experiment with different vocabulary sizes from the set $\{$*1K, 2K, 3K, 4K, 5K, 10K, 15K, 20K, 25K, 30K, 35K, 40K, 45K, 50K*$\}$. Figure 1 shows performance with this set of experiments, across the different vocabulary sizes. As Table 1 shows, the best unigram model is acquired with a vocabulary size $V = 10K$ words and reaches a Pearson correlation *(r)* = 0.240, which is a significant correlation (*p* < 0.014). As expected, higher up n-gram models require bigger vocabulary: The best bigram model is acquired with $V = 50K$ and is at a Pearson *(r)* = 0.240 (*p* < 0.004), and the best trigram model is acquired with $V = 30K$ with a lower Pearson (*r* = 0.192, *p* < 0.033). When we combine unigrams and bigrams, we acquire a better correlation than using each of these settings independently (*r* = 0.251, *p* < 0.003), at the cost of a big vocabulary size ($V = 50K$). The combination of unigrams, bigrams, and trigrams yields a lower correlation (*r* = 0.240, *p* < 0.008, $V = 45K$) than that acquired with the model combining unigrams and bigrams. We add the best performing text features from this set of experiments (i.e., unigrams+bigrams with $V = 50K$, *henceforth* `best_ngrams`) to all subsequent sets of experiments with the exception of word embeddings (since the goal of embeddings is to represent text words not as atomic symbols, but as vectors in a multidimensional space).

User Demographics: We apply features based on user gender and race as acquired from the surveys. For gender, we acquire 3 features (i.e., "male," "female," and "other") and for race, users chose one of 6 racial groups (i.e., "Asian," "Black," "Latino/Latina," "White," "Multiracial," and "Other") and were allowed to freely input other race affiliations in the form of free text. In our experiments, we exclusively apply features based on the 6 cate-

[1] www.qualtrics.com/Survey-Software

Setting	Pearson *r*	*p*
unigrams ($V = 10K$)	0.240	0.014
bigrams ($V = 50K$)	0.240	0.004
trigrams ($V = 30K$)	0.192	0.033
unigrams+bigrams ($V = 50K$)	0.251	0.003
unig+big+trig ($V = 45K$)	0.240	0.008
gender	0.142	0.093
gender+best_ngrams	**0.252**	**0.003**
race	0.056	0.384
race+best_ngrams	0.251	0.003
topics ($N = 1500$)	0.185	0.099
topics ($N = 1500$)+best_ngrams	0.248	0.004
EmoLex ($N = 50$)+best_ngrams	0.251	0.003
word embeddings (Wiki 300 dim)	0.191	0.044
word embeddings (FB 300 dim)	0.201	0.016

Table 1: Results with ridge regression in Pearson *r*.

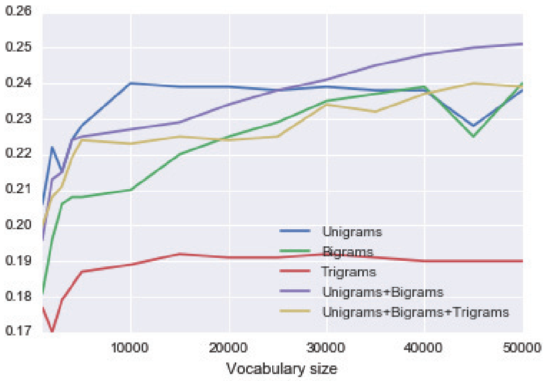

Figure 1: Performance of n-grams ≤ 3 by vocab. size.

gories users chose from. Female users had on average higher pathogenic empathy scores (i.e., μ = 0.060, σ = 0.667) than males (i.e., μ = -0.145, σ = 0.707), which suggests that women are more pathogenically empathetic than men. As Table 1 shows, when applied alone, gender features do not yield statistically significant correlation (Pearson *r* = 0.142, $p < 0.093$). Adding gender to best_ngrams (i.e., unigrams+bigrams with $V = 50K$), however, improves slightly over best_ngrams alone (with a Pearson *r* = 0.252, $p <$ 0.003). Interestingly, race features perform very poorly on the task, (Pearson *r* = 0.056), reflecting no significant correlation ($p < 0.384$) between racial groups and empathy levels (i.e., suggesting that people of all race empathize [or lack thereof] more or less to the same extents). Nor do race features improve correlation when added to best_ngrams.

LDA Topics: Do the topics users discuss interact with empathy expression? In other words, is it the case that if someone posts about topics involving human suffering or difficulties in family relationships, for example, the person would be expressing some level of empathy? To investigate this question, we train LDA (Blei, Ng, and Jordan 2003) models on the *MyPersonality* Facebook dataset (Kosinski, Stillwell, and Graepel 2013) mentioned earlier (a dataset of about 1.2 million Facebook posts belonging to 152,845 users, comprising about 14.5 million words). We employ LDA-topic-based features with 500, 1,000, 1,500, and 2,000 topics. As Table 1 shows, 1,500 topics acquire the best performance, (*r* = 0.185, $p < 0.099$). These results are still less than those acquired with unigrams alone. We then add LDA features to best_ngrams, but this causes a slight drop in the performance of the latter (from a Pearson *r* = 0.251 to an *r* = 0.248). This suggests that n-gram features are superior to LDA-based features in recognizing pathogenic empathy.

Emotion and Sentiment Features: To test the utility of capturing pathogenic empathy with emotion lexica, we employ different emotion-based features using the NRC lexicon, EmoLex (Mohammad and Turney 2013). EmoLex is manually labeled via crowdsourcing and has a total of 14,182 entries tagged for the 8 emotion categories *anger, anticipation, disgust, fear, happiness, sadness, surprise,* and *trust* as well as *positive* and *negative* valence. We apply 10 frequency-thresholded binary features corresponding to these 10 NRC categories where a feature will fire if the entries corresponding to its category are >= a given threshold. We add these features to the best text features, best_ngrams. We use thresholds from the set $\{$*10, 25, 50, 75, 100, 125, 150*$\}$. As Table 1 shows, the best result is acquired with a threshold = 50 and is at a Pearson *r* = 0.251, $p < 0.003$. As such, these features do not improve over best_ngrams. This may be surprising at first sight. However, it remains an open question to identify ways in which emotional states like those the lexicon represents interact with a trait like pathogenic empathy (e.g., *vis-a-vis* language expression). In addition, EmoLex may not have ideal coverage of social data (of the nature of Facebook posts).

Word Embeddings: We train and employ Wikipedia and Facebook word embedding features using the word2vec tool (Mikolov et al. 2013). Similar to previous research (e.g., (Zhang, Zhao, and LeCun 2015)), for each data point, we average the word vectors to acquire a vector with a bag of means. For Wikipedia, we use a 2016 dump (i.e., > 200 milion tokens) and for Facebook we use the *MyPersonality* dataset (comprising < 10 million tokens). In each case, we use a context window of ±10 words surrounding a focus word and keep a vocabulary of the most frequent 400,000 tokens. We find that the Facebook model outperforms the Wikipedia model (with a Pearson *r* = 0.201 for the first, and *r* = 0.191 for the second). Again, both of these results are below what we acquire with unigrams of $V = 10K$ (which are at a Pearson r = 0.240). Since we find the Facebook model to perform better than the Wikipedia model, confirming the superiority of in-domain unsupervised pre-training data for word embeddings, we further train models on Facebook data with lower (200 and 250) and higher (400 and 500) dimensions. However, none of these models improves over the one with 300 dimensions. Even though these word embedding models are outperformed by the simple unigram model, it is notable that a model with a number of dimensions as low as 300 (like the Facebook model reported here) acquires statistically significant correlations ($p < 0.016$) on the task. In general, results acquired with word embeddings here are in line with previous work on sentiment, e.g., (Zhang, Zhao, and LeCun 2015), another social meaning task, where embeddings perform below bag-of-word representations (unlike their proven effectiveness on many syntactic and sequence modeling tasks).

5 Conclusion

In this paper, we investigated the utility of building models for recognizing pathogenic empathy from social media language: We introduced a survey-based method grounded in psychology for acquiring ground-truth pathogenic empathy scores of individuals whose Facebook data we label with these scores, thus enabling the measurement of this important psychological construct at the user level. We then built promising models for detecting pathogenic empathy. Among other applications, we believe our models will be useful for testing hypotheses about the correlation of pathogenic empathy with outcomes such as burnout and stress in people for whom we have social media but no questionnaire data, and eventually for designing chatbots which exhibit different forms of empathy.

References

Batson, C. D.; Early, S.; and Salvarani, G. 1997. Perspective taking: Imagining how another feels versus imaging how you would feel. *Personality and social psychology bulletin* 23(7):751–758.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *the Journal of machine Learning research* 3:993–1022.

Bollen, J.; Mao, H.; and Zeng, X. 2011. Twitter mood predicts the stock market. *Journal of Computational Science* 2(1):1–8.

Buffone, A. E. K.; Giorgi, S.; Jordan, M.; Eichstaedt, J.; Kern, M.; Carpenter, J.; Abdul-Mageed, M.; Ungar, L.; and Seligman, Martin, E. P. in prep. Big data insights into different forms of empathy: pathogenic versus salutogenic empathy and their associations with health, stress, and prosocial behavior.

Buffone, A. E. K. 2015. *Perspective taking and the biopsychosocial model of challenge and threat: Effects of imagine-other and imagine-self perspective taking on active goal pursuit*. Ph.D. Dissertation, STATE UNIVERSITY OF NEW YORK AT BUFFALO.

Collins, F. M. 2014. The relationship between social media and empathy.

Fung, P.; Bertero, D.; Wan, Y.; Dey, A.; Chan, R. H. Y.; Siddique, F. B.; Yang, Y.; Wu, C.-S.; and Lin, R. 2016. Towards empathetic human-robot interactions. *arXiv preprint arXiv:1605.04072*.

Gibson, J.; Can, D.; Xiao, B.; Imel, Z. E.; Atkins, D. C.; Georgiou, P.; and Narayanan, S. 2016. A deep learning approach to modeling empathy in addiction counseling. *Commitment* 111:21.

Jordan, M. R.; Amir, D.; and Bloom, P. 2016. Are empathy and concern psychologically distinct? *Emotion* 16(8):1107.

Kosinski, M.; Stillwell, D.; and Graepel, T. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences* 110(15):5802–5805.

Litvak, M.; Otterbacher, J.; Ang, C. S.; and Atkins, D. 2016. Social and linguistic behavior and its correlation to trait empathy. *PEOPLES 2016* 128.

Manczak, E. M.; DeLongis, A.; and Chen, E. 2016. Does empathy have a cost? diverging psychological and physiological effects within families.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Mohammad, S. M., and Kiritchenko, S. 2015. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence* 31(2):301–326.

Mohammad, S. M., and Turney, P. D. 2013. Crowdsourcing a word-emotion association lexicon. 29(3):436–465.

Preotiuc-Pietro, D.; Eichstaedt, J.; Park, G.; Sap, M.; Smith, L.; Tobolsky, V.; Schwartz, H. A.; and Ungar, L. 2015. The role of personality, age and gender in tweeting about mental illnesses. *NAACL HLT 2015* 21.

Prot, S.; Gentile, D. A.; Anderson, C. A.; Suzuki, K.; Swing, E.; Lim, K. M.; Horiuchi, Y.; Jelic, M.; Krahé, B.; Liuqing, W.; et al. 2014. Long-term relations among prosocial-media use, empathy, and prosocial behavior. *Psychological science* 25(2):358–368.

Raposa, E. B.; Laws, H. B.; and Ansell, E. B. 2015. Prosocial behavior mitigates the negative effects of stress in everyday life. *Clinical Psychological Science* 2167702615611073.

Riess, H.; Kelley, J. M.; Bailey, R. W.; Dunn, E. J.; and Phillips, M. 2012. Empathy training for resident physicians: a randomized controlled trial of a neuroscience-informed curriculum. *Journal of general internal medicine* 27(10):1280–1286.

Schwartz, H. A.; Eichstaedt, J. C.; Kern, M. L.; Dziurzynski, L.; Ramones, S. M.; Agrawal, M.; Shah, A.; Kosinski, M.; Stillwell, D.; Seligman, M. E.; et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one* 8(9):e73791.

Xiao, B.; Imel, Z. E.; Georgiou, P. G.; Atkins, D. C.; and Narayanan, S. S. 2015. " rate my therapist": Automated detection of empathy in drug and alcohol counseling via speech and language processing. *PloS one* 10(12):e0143055.

Xiao, B.; Huang, C.; Imel, Z. E.; Atkins, D. C.; Georgiou, P.; and Narayanan, S. S. 2016. A technology prototype system for rating therapist empathy from audio recordings in addiction counseling. *PeerJ Computer Science* 2:e59.

Zhang, X.; Zhao, J.; and LeCun, Y. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, 649–657.