Classifying Minority Stress Disclosure on Social Media with Bidirectional Long Short-Term Memory

Cory J. Cascalheira[1,2], Shah Muhammad Hamdi[1], Jillian R. Scheer[2], Koustuv Saha[3], Soukaina Filali Boubrahimi[4], Munmun De Choudhury[5]

New Mexico State University[1], Syracuse University[2], Microsoft Research[3], Utah State University[4],
Georgia Institute of Technology[5]

cjcascalheira@gmail.com, shamdi1@nmsu.edu, jrscheer@syr.edu, koustuv.saha@gmail.com,
soukaina.boubrahimi@usu.edu, munmun.choudhury@cc.gatech.edu

Abstract

Because of their stigmatized social status, sexual and gender minority (SGM; e.g., gay, transgender) people experience minority stress (i.e., identity-based stress arising from adverse social conditions). Given that minority stress is the leading framework for understanding health inequity among SGM people, researchers and clinicians need accurate methods to detect minority stress. Since social media fulfills important developmental, affiliative, and coping functions for SGM people, social media may be an ecologically valid channel for detecting minority stress. In this paper, we propose a bidirectional long short-term memory (BI-LSTM) network for classifying minority stress disclosed on Reddit. Our experiments on a dataset of 12,645 Reddit posts resulted in an average accuracy of 65%.

Introduction

Establishing the construct validity of psychological phenomena with computational, data-driven approaches is a federal funding priority (National Institutes of Health 2021). To pursue this priority, researchers have used machine learning (ML) and deep learning (DL) to classify mental health disorders, such as depression (Aldarwish and Ahmad 2017), from social media posts. Despite these advances, computational support for the construct validity of *minority stress* (i.e., identity-based stress arising from adverse social conditions; Meyer 2003) is lacking.

Minority Stress and Social Media

Because of their stigmatized social status, sexual and gender minority (SGM; e.g., gay, transgender) people experience minority stress (Meyer 2003). Minority stress is associated with poorer health outcomes (e.g., Hatzenbuehler and Pachankis 2016). Consequently, minority stress is the leading framework for understanding health inequity among SGM people relative to the general population (Institute of Medicine of the National Academies 2011). Prominent SGM-tailored clinical interventions aim to ameliorate minority stress or to bolster coping (e.g., Pachankis, McConocha, et al., 2020). To study and treat minority stress, researchers and clinicians must accurately detect it. However, like most constructs in psychological science (Sassenberg and Ditrich 2019), the dominant method of detecting minority stress is via survey-based self-report, which has numerous methodological limitations (e.g., retrospective reporting; Heppner et al. 2016). While physiological and observational techniques address these limitations (Heppner et al. 2016), these methods can be hard to implement (e.g., participant must visit lab) if SGM participants are not out or live in high-stigma areas. Consequently, hard-to-reach SGM samples often are solicited from social media (Lunn et al. 2019).

Social media *as a data source* may be an ecologically valid way to detect minority stress because social media often fulfills important developmental, affiliative, and coping functions for SGM people (Formby 2017; Tropiano 2014; McInroy and Craig 2020; Woznicki et al. 2021). Compared to 58% of the general public, 80% of SGM adults have used a social media website (Pew Research Center 2013). Many SGM people solicit social support on social media (McInroy and Craig 2020). Seeking social support is an effective strategy to cope with minority stress (e.g., Toomey et al., 2018). To obtain social support online, social media users are motivated to self-disclose (Luo and Hancock 2020), and research shows that SGM people disclose minority stress on social media (Saha et al. 2019).

Consequently, using social media to detect minority stress disclosures may circumvent survey-based limitations, meet the logistical demands of research with SGM people, and increase understanding of minority stress disclosure within the virtual environment. Large social media data sets require sophisticated pattern mining techniques. Thus, deep learning methods, for their end-to-end learning ability from large datasets, have the potential to complement and to extend minority stress theory and existing SGM research. However, no study has tested the potential of a particular type of DL—the recurrent neural network (RNN)—to classify minority stress disclosures on social media. The main contribution of this proof-of-concept paper is the application of a bidirectional long short-term memory (BI-LSTM) network to classify minority stress on a social media dataset.

Data Source

We used a existing dataset. Saha et al. (2019) manually labeled 350 (2.77%) of the 12,645 Reddit posts in the data set. The team extracted 659 features (e.g., sentiment, hate speech) of minority stress from these manually labeled examples, then used the multilayer perceptron (MLP) algorithm as a classifier to label the remaining Reddit posts. The MLP classifier identified 4,419 (35%) Reddit posts as exhibiting minority stress (Saha et al. 2019); thus, the dataset was unbalanced. We used this dataset to examine the performance of BI-LSTM. The dataset was split into training (80%), validation (10%), and test (10%) using stratification. Binary labels for the presence (1) or absence (0) of minority stress were used.

Neural Network Selection and Architecture

To build upon Saha et al. (2019), we investigated a particular type of RNN, the BI-LSTM. Compared to feedforward neural networks (FFNNs), such as the MLP classifier, RNNs warrant investigation for several reasons.

First, RNNs are specialized for sequential data (Goodfellow, Bengio, and Courville 2016), such as textual Reddit posts. Unlike the traditional FFNN, RNNs learn representations given the temporal relationship of the data points.

Second, similar to FFNNs, RNNs learn parameters via backpropagation. However, when backpropagation is applied to sequential data, the gradients tend to vanish (i.e., shrink exponentially) or explode (i.e., grow exponentially), thereby introducing error and impairing the model's capability to learn parameters (Hochreiter and Schmidhuber 1997). Essentially, the output gate learns when "to trap" the error and prevent it from impacting the model, whereas the input gate learns when "to release" the error (Hochreiter and Schmidhuber 1997, 7). LSTM (and, by extension, BI-LSTM) learns when to keep or to forget information that improves parameter estimation even when the sequences are substantially long (Goodfellow, Bengio, and Courville 2016). Therefore, unlike FFNNs, LSTM models efficiently handle long-term dependencies (Goodfellow, Bengio, and Courville 2016). LSTM's ability to remember information over a long sequence period is beneficial for our work because Reddit posts can be several paragraphs long. Finally, LSTM-based models have shown excellent performance in classification tasks for related psychological constructs (Bisht et al. 2020), achieving better accuracy over MLP (Süt and Şenocak 2007).

In this work, we used a BI-LSTM architecture (Cheng 2020) and implemented it in PyTorch (Paszke et al. 2019). See Figure 1 for the network's architecture. First, each Reddit post was transformed into an object, which held the length of each Reddit post and the tokens (i.e., individual words). Reddit posts were transformed into tokens using spaCy (Honnibal and Montani 2021).

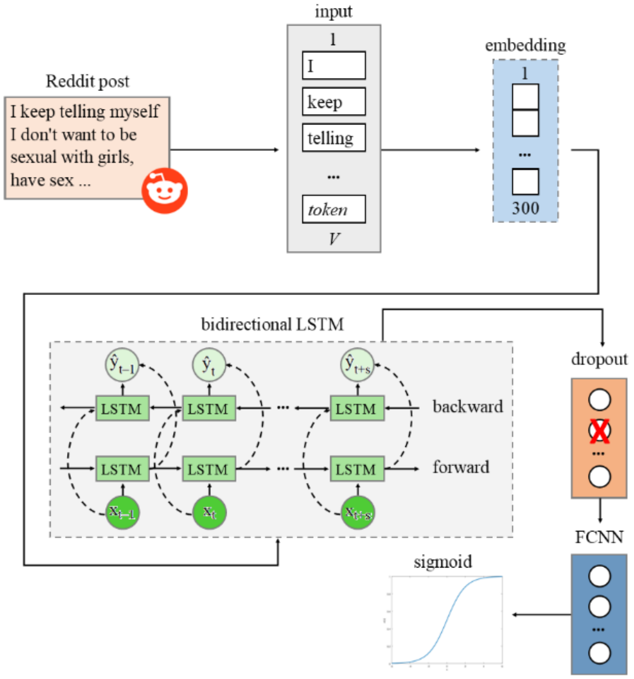

Figure 1. BI-LSTM network architecture.

Second, a word embedding layer was constructed. The embedding layer is a $|V| \times D$ matrix where V is the vocabulary (i.e., number of tokens in the Reddit post) and D is the dimensionality of the embeddings ($D = 300$; Paszke et al. 2019). Word embeddings are able to represent the syntactic and semantic information attributed to words (Lai et al. 2016). Capturing linguistic meaning allows the words in the Reddit post to serve as better input units for BI-LSTM.

Next, a one-layer BI-LSTM was created with an input tensor of size (N, L, H_{in}), where N is the batch size, L is the sequence length, and H_{in} is the size of embedding dimension (Paszke et al. 2019); 128 features were used in the hidden state. BI-LSTM expands the LSTM cell's ability to consider *context* in learning representation (Goodfellow, Bengio, and Courville 2016). Context, in text data, indicates how words are influenced by neighboring words. Unlike traditional LSTM, which can only assess the previous context (i.e., the sequence of words in the past), BI-LSTM takes advantages of the future context (i.e., the next sequence of words) as well. That is, BI-LSTM processes text in both directions, updating hidden layer activations by moving from the start of the input sentence (i.e., the forward sequence) as well as from the end of the input sentence (i.e., the backward sequence; Graves, Mohamed, and Hinton 2013). The BI-LSTM output was concatenated and fed into a dropout regularization layer with probability 0.5.

Finally, the concatenated and regularized representation from the BI-LSTM layer was passed to a FCNN (fully connected neural network) with a sigmoid activation function to obtain the probability that a Reddit post evinces minority stress. The FCNN layer has size 256 (i.e., 2 * 128, the number of features in the hidden state).

During training and evaluation, cross-entropy loss and the Adam optimizer were used. Cross-entropy loss is appropriate for this binary classification problem. The Adam optimizer adapts the learning rate throughout training (Goodfellow, Bengio, and Courville 2016) and the following values were initialized: learning rate = 1e-3, β_1 = 0.9, β_2 = 0.999, ε = 1e-8.

Classification Results

Hand-coded hyperparameter tuning was used. Only the batch size, learning rate (i.e., alpha), and number of epochs were tuned. Six sets of hyperparameters were examined and are presented in Table 1. Figure 2 shows the training and validation loss. Results indicated that accuracy and the F1 scores were greatest for hyperparameter set 1 due to a longer training time. As shown in hyperparameter Set 3, increasing the alpha reduced the accuracy and F1 scores. Compared to Set 2, increasing the batch size in hyperparameter Set 4 did not influence the accuracy. We reduced the epochs in attempt to address the overfitting evident in hyperparameter Set 1, but reducing epochs to 3 (Set 5) resulted in inferior accuracy. Thus, although the accuracy and F1 scores were not superior in Set 2, we used these hyperparameters in Set 6 to evaluate our model.

During evaluation, the one-layer BI-LSTM model correctly classified 65.34% of the Reddit posts with F1 scores falling from training (1 = 0.11, 0 = 0.78). As shown in Figure 3, the model generated many false negatives.

Set	Hyperparameters	Acc.	F1, 1	F1, 0
1	batch (4), alpha (0.001), #epochs (10)	0.89	0.83	0.92
2	batch (4), alpha (0.001), #epochs (5)	0.82	0.73	0.87
3	batch (4), alpha (0.01), #epochs (5)	0.64	0.41	0.74
4	batch (10), alpha (0.001), #epochs (5)	0.82	0.72	0.87
5	batch (10), alpha (0.001), #epochs (3)	0.76	0.60	0.82
6	batch (4), alpha (0.001), #epochs (5)	0.81	0.71	0.86

Table1. Training and validation loss, unbalanced dataset.

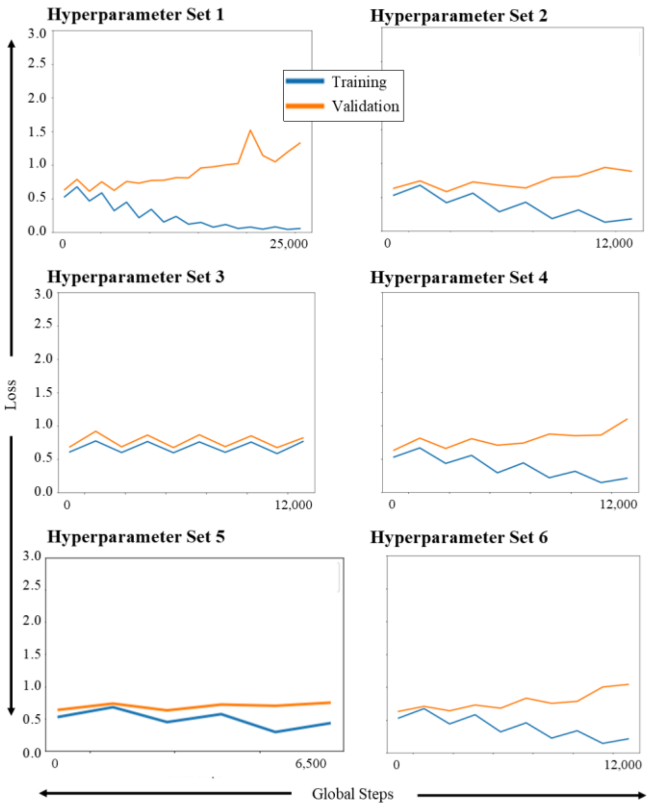

Figure 2. Training and validation loss, unbalanced dataset.

Since unbalanced datasets can affect accuracy (Goodfellow, Bengio, and Courville 2016), we trained the BI-LSTM model again using hyperparameter Set 6 with a balanced dataset. We randomly sampled negative cases of minority stress disclosure to match the number of positive cases (n = 4,419) to establish a 1:1 ratio (N = 8,838).

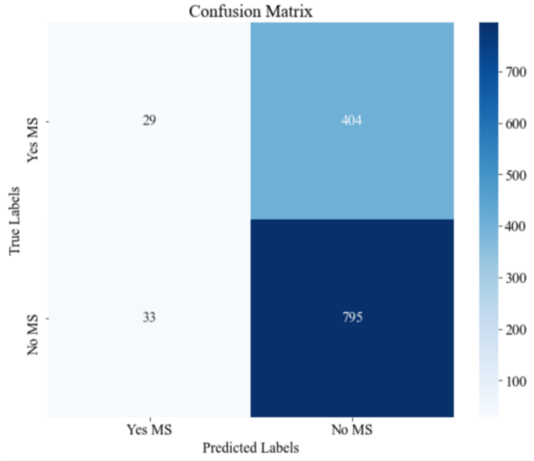

Figure 3. Classification performance, unbalanced dataset.

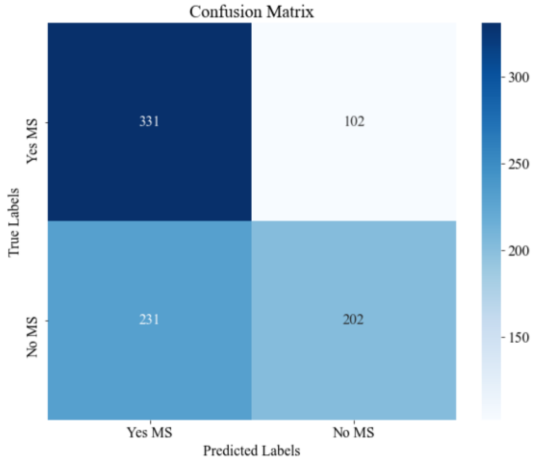

Figure 4. Classification performance, balanced dataset.

Although the accuracy fell to 61.55% with the balanced dataset, BI-LSTM detected more true positive and true negative cases of minority stress (see Figure 4).

Discussion

This proof-of-concept paper provided initial evidence in support of using sequence models (i.e., BI-LSTM) in classifying minority stress disclosures on social media. Although BI-LSTM (65%) achieved a lower accuracy than MLP (75%; Saha et al. 2019), several limitations must be considered to contextualize the results. Importantly, this paper opens the door to using deep neural networks for studying minority stress which, as a theory-driven concept, is novel given that no features were engineered by experts in this study.

Limitations and Future Research

To build upon this initial study, the following limitations should be addressed. First, ad-hoc hyperparameter tuning was performed. A more rigorous method would be the use of random search for hyperparameter selection. Second, and relatedly, only three hyperparameters were tuned. Future work should consider tuning additional hyperparameters, such as increasing the number of BI-LSTM layers. Third, although the BI-LSTM model performed adequately with custom word embeddings, future work should consider phrase or sentence embeddings. Indeed, minority stress is a nuanced psychological construct that is difficult to communicate without using multiple words. As the results indicate, single-word embeddings appeared unable to capture the semantic and syntactic meaning of minority stress. Fourth, if single-word embeddings are pursued, then instead of custom word embeddings, future work should consider pretrained embeddings, such as Word2Vec, FastText, and GloVe. Finally, we should examine other DL techniques for text classification, such as convolutional LSTM (Zhou et al. 2015) or bidirectional encoder representations from transformers (BERT; Devlin et al. 2019).

Broader Perspectives and Ethics

Despite these limitations, detecting minority stress on social media with sequence models warrants further research. If sequence models can accurately classify minority stress disclosure, then their deployment as services could (a) identify SGM Internet users most at risk for adverse consequences; (b) link SGM people to professional care (e.g., personalized ads to SGM-affirming therapists); and (c) generate brief, automated interventions (e.g., chatbots to affirm disclosure, screen for comorbid risks, and link to resources). Without addressing the above limitations, definitive conclusions about the potential of sequence models to classify minority stress on social media remain elusive.

Several ethical considerations are worth mentioning. First, SGM people may use social media for relative anonymity. If algorithms identify posts as evincing minority stress, then SGM users may be targeted by malicious others. Relatedly, if an SGM user discloses minority stress that is based on or political injustice, then it is possible that the user could be identified by the perpetrator. We minimized these risks by using Reddit, which is relatively more anonymous than services such as Facebook. Second, although SGM people consent to third parties accessing their information by signing up for Reddit, it is neither clear whether SGM users would consent to the computational classification of their content nor whether they would appreciate the algorithmic interventions proposed above. As Saha et al. (2019) concluded, it is imperative that minority stress classification from social media data does no harm and benefits the SGM community.

References

Aldarwish, M. M., and H. F. Ahmad. 2017. "Predicting Depression Levels Using Social Media Posts." In , 277–80.

Bisht, A., A. Singh, H. S. Bhadauria, J. Virmani, and Kriti. 2020. "Detection of Hate Speech and Offensive Language in Twitter Data Using LSTM Model." In *Recent Trends in Image and Signal Processing in Computer Vision*, edited by S. Jain and S. Paul, 243–64. Singapore: Springer Singapore.

Cheng, R. 2020. "LSTM Text Classification Using PyTorch." Towards Data Science. 2020. Accessed December 05, 2021. https://towardsdatascience.com/lstm-text-classification-using-pytorch-2c6c657f8fc0.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of NAACL-HLT*, 4171–86. Association for Computational Linguistics.

Formby, E. 2017. *Exploring LGBT Spaces and Communities: Contrasting Identities, Belongings and Wellbeing*. Routledge.

Goodfellow, I., Y. Bengio, and A. Courville. 2016. *Deep Learning*. The MIT Press.

Graves, A., A. Mohamed, and G. Hinton. 2013. "Speech Recognition with Deep Recurrent Neural Networks." In *Proceedings of the ICASSP*, 6645-49. IEEE.

Hatzenbuehler, M. L., and J. E. Pachankis. 2016. "Stigma and Minority Stress as Social Determinants of Health among Lesbian, Gay, Bisexual, and Transgender Youth: Research Evidence and Clinical Implications." *Pediatric Clinics* 63 (6): 985–97.

Heppner, P. P., B. E. Wampold, J. Owen, M. N. Thompson, and K. T. Wang. 2016. *Research Design in Counseling*. 4th ed. Boston, MA: Cengage Learning.

Hochreiter, S., and J. Schmidhuber. 1997. "Long Short-Term Memory." *Neural Computation* 9 (8): 1735–80.

Honnibal, M., and I. Montani. 2021. *SpaCy: Industrial-Strength Natural Language Processing* (version 3.2). Accessed December 05, 2021. https://spacy.io/.

Institute of Medicine of the National Academies. 2011. *The Health of Lesbian, Gay, Bisexual, and Transgender People: Building a Foundation for Better Understanding*. Washington, DC: The National Academics Press.

Lai, S., K. Liu, S. He, and J. Zhao. 2016. "How to Generate a Good Word Embedding." *IEEE Intelligent Systems* 31 (6): 5–14.

Lunn, M. R., M. R. Capriotti, A. Flentje, K. Bibbins-Domingo, M. J. Pletcher, A. J. Triano, C. Sooksaman, J. Frazier, and J. Obedin-Maliver. 2019. "Using Mobile Technology to Engage Sexual and Gender Minorities in Clinical Research." *PLOS ONE* 14 (5): e0216282.

Luo, M., and J. T. Hancock. 2020. "Self-Disclosure and Social Media: Motivations, Mechanisms and Psychological Well-Being." *Privacy and Disclosure, Online and in Social Interactions* 31 (February): 110–15.

McInroy, L. B., and S. L. Craig. 2020. "'It's like a Safe Haven Fantasy World': Online Fandom Communities and the Identity Development Activities of Sexual and Gender Minority Youth." *Psychology of Popular Media* 9 (2): 236–46.

Meyer, I. H. 2003. "Prejudice, Social Stress, and Mental Health in Lesbian, Gay, and Bisexual Populations: Conceptual Issues and Research Evidence." *Psychological Bulletin* 129: 674–97.

National Institutes of Health. 2021. "Computational Approaches for Validating Dimensional Constructs of Relevance to Psychopathology (R01 Clinical Trial Optional)." Funding Opportunity Announcement, RFA-MH-19-242. July 7, 2021. https://grants.nih.gov/grants/guide/pa-files/PAR-21-263.html.

Pachankis, J. E., E. M. McConocha, K. A. Clark, K. Wang, K. Behari, B. K. Fetzner, C. D. Brisbin, J. R. Scheer, and K. Lehavot. 2020. "A Transdiagnostic Minority Stress Intervention for Gender Diverse Sexual Minority Women's Depression, Anxiety, and Unhealthy Alcohol Use: A Randomized Controlled Trial." *Journal of Consulting and Clinical Psychology* 88 (7): 613–30.

Paszke, A., S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, et al. 2019. "PyTorch: An Imperative Style, High-Performance Deep Learning Library." In *Advances in Neural Information Processing Systems 32*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. D. Alché-Buc, E. Fox, and R. Garnett, 8024–35. Curran Associates, Inc. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Pew Research Center. 2013. "A Survey of LGBT Americans: Attitudes, Experiences and Values in Changing Times." Accessed December 05, 2021. https://www.pewresearch.org/social-trends/2013/06/13/a-survey-of-lgbt-americans/.

Saha, K., S. C. Kim, M. D. Reddy, A. J. Carter, E. Sharma, O. L. Haimson, and M. De Choudhury. 2019. "The Language of LGBTQ+ Minority Stress Experiences on Social Media." *Proceedings of the ACM on Human-Computer Interaction* 3 (CSCW).

Sassenberg, K., and L. Ditrich. 2019. "Research in Social Psychology Changed between 2011 and 2016: Larger Sample Sizes, More Self-Report Measures, and More Online Studies." *Advances in Methods and Practices in Psychological Science* 2 (2): 107–14.

Süt, N., and M. Şenocak. 2007. "Assessment of the Performances of Multilayer Perceptron Neural Networks in Comparison with Recurrent Neural Networks and Two Statistical Methods for Diagnosing Coronary Artery Disease." *Expert Systems* 24 (3): 131–42.

Toomey, R. B., C. Ryan, R. M. Diaz, and S. T. Russell. 2018. "Coping with Sexual Orientation–Related Minority Stress." *Journal of Homosexuality* 65 (4): 484–500.

Tropiano, S. 2014. "A Safe and Supportive Environment: LGBTQ Youth and Social Media." In *Queer Youth and Media Cultures*, edited by C. Pullen, 46–62. Palgrave Macmillan.

Woznicki, N., A. S. Arriaga, N. A. Caporale-Berkowitz, and M. C. Parent. 2021. "Parasocial Relationships and Depression among LGBQ Emerging Adults Living with Their Parents during COVID-19: The Potential for Online Support." *Psychology of Sexual Orientation and Gender Diversity* 8 (2): 228–37.

Zhou, C., C. Sun, Z. Liu, and F. C. M. Lau. 2015. "A C-LSTM Neural Network for Text Classification." *ArXiv* abs/1511.08630: 1–10.