Dreaddit: A Reddit Dataset for Stress Analysis in Social Media

Elsbeth Turcan, Kathleen McKeown
Columbia University
Department of Computer Science
{eturcan, kathy}@cs.columbia.edu

Abstract

Stress is a nigh-universal human experience, particularly in the online world. While stress can be a motivator, too much stress is associated with many negative health outcomes, making its identification useful across a range of domains. However, existing computational research typically only studies stress in domains such as speech, or in short genres such as Twitter. We present Dreaddit, a new text corpus of lengthy multi-domain social media data for the identification of stress. Our dataset consists of 190K posts from five different categories of Reddit communities; we additionally label 3.5K total segments taken from 3K posts using Amazon Mechanical Turk. We present preliminary supervised learning methods for identifying stress, both neural and traditional, and analyze the complexity and diversity of the data and characteristics of each category.

1 Introduction

In our online world, social media users tweet, post, and message an incredible number of times each day, and the interconnected, information-heavy nature of our lives makes stress more prominent and easily observable than ever before. With many platforms such as Twitter, Reddit, and Facebook, the scientific community has access to a massive amount of data to study the daily worries and stresses of people across the world.[1]

Stress is a nearly universal phenomenon, and we have some evidence of its prevalence and recent increase. For example, the American Psychological Association (APA) has performed annual studies assessing stress in the United States since 2007[2] which demonstrate widespread experiences of chronic stress. Stress is a subjective experience whose effects and even definition can vary from person to person; as a baseline, the APA defines stress as a reaction to extant and future demands and pressures,[3] which can be positive in moderation. Health and psychology researchers have extensively studied the connection between too much stress and physical and mental health (Lupien et al., 2009; Calcia et al., 2016).

In this work, we present a corpus of social media text for detecting the presence of stress. We hope this corpus will facilitate the development of models for this problem, which has diverse applications in areas such as diagnosing physical and mental illness, gauging public mood and worries in politics and economics, and tracking the effects of disasters. Our contributions are as follows:

- Dreaddit, a dataset of lengthy social media posts in five categories, each including stressful and non-stressful text and different ways of expressing stress, with a subset of the data annotated by human annotators;[4]
- Supervised models, both discrete and neural, for predicting stress, providing benchmarks to stimulate further work in the area; and
- Analysis of the content of our dataset and the performance of our models, which provides insight into the problem of stress detection.

In the remainder of this paper, we will review relevant work, describe our dataset and its annotation, provide some analysis of the data and stress detection problem, present and discuss results of some supervised models on our dataset, and finally conclude with our summary and future work.

2 Related Work

Because of the subjective nature of stress, relevant research tends to focus on physical sig-

[1] https://www.gse.harvard.edu/news/uk/17/12/social-media-and-teen-anxiety

[2] https://www.apa.org/news/press/releases/stress/index?tab=2

[3] https://www.apa.org/helpcenter/stress-kinds

[4] Our dataset will be made available at http://www.cs.columbia.edu/~eturcan/data/dreaddit.zip.

Proceedings of the 10th International Workshop on Health Text Mining and Information Analysis (LOUHI 2019), pages 97–107
Hong Kong, November 3, 2019.
https://doi.org/10.18653/v1/D19-62

nals, such as cortisol levels in saliva (Allen et al., 2014), electroencephalogram (EEG) readings (Al-Shargie et al., 2016), or speech data (Zuo et al., 2012). This work captures important aspects of the human reaction to stress, but has the disadvantage that hardware or physical presence is required. However, because of the aforementioned proliferation of stress on social media, we believe that stress can be observed and studied purely from text.

Other threads of research have also made this observation and generally use microblog data (e.g., Twitter). The most similar work to ours includes Winata et al. (2018), who use Long Short-Term Memory Networks (LSTMs) to detect stress in speech and Twitter data; Guntuku et al. (2018), who examine the Facebook and Twitter posts of users who score highly on a diagnostic stress questionnaire; and Lin et al. (2017), who detect stress on microblogging websites using a Convolutional Neural Network (CNN) and factor graph model with a suite of discrete features. Our work is unique in that it uses data from Reddit, which is both typically longer and not typically as conducive to distant labeling as microblogs (which are labeled in the above work with hashtags or pattern matching, such as "I feel stressed"). The length of our posts will ultimately enable research into the causes of stress and will allow us to identify more implicit indicators. We also limit ourselves to text data and metadata (e.g., posting time, number of replies), whereas Winata et al. (2018) also train on speech data and Lin et al. (2017) include information from photos, neither of which is always available. Finally, we label individual parts of longer posts for acute stress using human annotators, while Guntuku et al. (2018) label users themselves for chronic stress with the users' voluntary answers to a psychological questionnaire.

Researchers have used Reddit data to examine a variety of mental health conditions such as depression (Choudhury et al., 2013) and other clinical diagnoses such as general anxiety (Cohan et al., 2018), but to our knowledge, our corpus is the first to focus on stress as a general experience, not only a clinical concept.

3 Dataset

3.1 Reddit Data

Reddit is a social media website where users post in topic-specific communities called subreddits, and other users comment and vote on these posts. The lengthy nature of these posts makes Reddit an ideal source of data for studying the nuances of phenomena like stress. To collect expressions of stress, we select categories of subreddits where members are likely to discuss stressful topics:

- **Interpersonal conflict**: abuse and social domains. Posters in the abuse subreddits are largely survivors of an abusive relationship or situation sharing stories and support, while posters in the social subreddit post about any difficulty in a relationship (often but not exclusively romantic) and seek advice for how to handle the situation.
- **Mental illness**: anxiety and Post-Traumatic Stress Disorder (PTSD) domains. Posters in these subreddits seek advice about coping with mental illness and its symptoms, share support and successes, seek diagnoses, and so on.
- **Financial need**: financial domain. Posters in the financial subreddits generally seek financial or material help from other posters.

We include ten subreddits in the five domains of abuse, social, anxiety, PTSD, and financial, as detailed in Table 1, and our analysis focuses on the domain level. Using the `PRAW` API,[5] we scrape all available posts on these subreddits between January 1, 2017 and November 19, 2018; in total, 187,444 posts. As we will describe in subsection 3.2, we assign binary stress labels to 3,553 segments of these posts to form a supervised and semi-supervised training set. An example segment is shown in Figure 1. Highlighted phrases are in-

> I have this feeling of dread about school right before I go to bed and I wake up with an upset stomach which lasts all day and nakes me feel like I'll throw up. This causes me to lose appetite and not wanting to drink water for fear of throwing up. I'm not sure where else to go with this, but I need help. If any of you have this, can you tell me how you deal with it? I'm tired of having this every day and feeling like I'll throw up.

Figure 1: An example of stress being expressed in social media from our dataset, from a post in r/anxiety (reproduced exactly as found). Some possible expressions of stress are highlighted.

[5] https://github.com/praw-dev/praw

Domain	Subreddit Name	Total Posts	Avg Tokens/Post	Labeled Segments
abuse	r/domesticviolence	1,529	365	388
	r/survivorsofabuse	1,372	444	315
	Total	**2,901**	**402**	**703**
anxiety	r/anxiety	58,130	193	650
	r/stress	1,078	107	78
	Total	**59,208**	**191**	**728**
financial	r/almosthomeless	547	261	99
	r/assistance	9,243	209	355
	r/food_pantry	343	187	43
	r/homeless	2,384	143	220
	Total	**12,517**	**198**	**717**
PTSD	r/ptsd	4,910	265	711
social	r/relationships	107,908	578	694
All		**187,444**	**420**	**3,553**

Table 1: **Data Statistics**. We include ten total subreddits from five domains in our dataset. Because some subreddits are more or less popular, the amount of data in each domain varies. We endeavor to label a comparable amount of data from each domain for training and testing.

dicators that the writer is stressed: the writer mentions common physical symptoms (nausea), explicitly names fear and dread, and uses language indicating helplessness and help-seeking behavior.

The average length of a post in our dataset is 420 tokens, much longer than most microblog data (e.g., Twitter's character limit as of this writing is 280 characters). While we label segments that are about 100 tokens long, we still have much additional data from the author on which to draw. We feel this is important because, while our goal in this paper is to predict stress, having longer posts will ultimately allow more detailed study of the causes and effects of stress.

In Table 2, we provide examples of labeled segments from the various domains in our dataset. The samples are fairly typical; the dataset contains mostly first-person narrative accounts of personal experiences and requests for assistance or advice. Our data displays a range of topics, language, and agreement levels among annotators, and we provide only a few examples. Lengthier examples are available in the appendix.

3.2 Data Annotation

We annotate a subset of the data using Amazon Mechanical Turk in order to begin exploring the characteristics of stress. We partition the posts into contiguous five-sentence chunks for labeling; we wish to annotate segments of the posts because we are ultimately interested in what parts of the post depict stress, but we find through manual inspection that some amount of context is important. Our posts, however, are quite long, and it would be difficult for annotators to read and annotate entire posts. This type of data will allow us in the future not only to *classify* the presence of stress, but also to *locate* its expressions in the text, even if they are diffused throughout the post.

We set up an annotation task in which English-speaking Mechanical Turk Workers are asked to label five randomly selected text segments (of five sentences each) after taking a qualification test; Workers are allowed to select "Stress", "Not Stress", or "Can't Tell" for each segment. In our instructions, we define stress as follows: "The Oxford English Dictionary defines stress as 'a state of mental or emotional strain or tension resulting from adverse or demanding circumstances'. This means that stress results from someone being uncertain that they can handle some threatening situation. We are interested in cases where that someone also feels negatively about it (sometimes we can find an event stressful, but also find it exciting and positive, like a first date or an interview).". We specifically ask Workers to decide whether the author is expressing both stress and a negative attitude about it, not whether the situation itself seems stressful. Our full instructions are available in the appendix.

We submit 4,000 segments, sampled equally from each domain and uniformly within domains,

Text	Domain	Label	Ann. Agreed
I only get it when I have a flashback or strong reaction to a trigger. I notice it sticks around even when I feel emotionally calm and can stick around for a long time after the trigger, like days or weeks. Its a new symptom I think. Also been having lots of nightmares again recently. Not sure what to do as Im not currently in therapy, but I am waiting to be seen at a mental health clinic.	PTSD	stress	6/7 (86%)
Regardless, that didn't last long, maybe half a year. I released that apartment, and most of my belongings (I kept a few boxes of my things from the military, personal effects, but little else). Looking back, there were some signs of emotional manipulation here, but it was subtle... and you know how it is, love is blind. We got engaged. It was quite the affair.	abuse	not stress	5/5 (100%)
Our dog Jett has been diagnosed with diabetes and is now in the hospital to stabilize his blood sugar. Luckily, he seems to be doing well and he will be home with us soon. Unfortunately, his bill is large enough that we just can't cover it on our own (especially with our poor financial situation). We're being evicted from our home soon and trying to find a place with this bill is just too much for us by ourselves. To help us pay the bill we've set up a GoFundMe.	financial	stress	3/5 (60%)

Table 2: **Data Examples.** Examples from our dataset with their domains, assigned labels, and number of annotators who agreed on the majority label (reproduced exactly as found, except that a link to the GoFundMe has been removed in the last example). Annotators labeled these five-sentence segments of larger posts.

to Mechanical Turk to be annotated by at least five Workers each and include in each batch one of 50 "check questions" which have been previously verified by two in-house annotators. After removing annotations which failed the check questions, and data points for which at least half of the annotators selected "Can't Tell", we are left with 3,553 labeled data points from 2,929 different posts. We take the annotators' majority vote as the label for each segment and record the percentage of annotators who agreed. The resulting dataset is nearly balanced, with 52.3% of the data (1,857 instances) labeled stressful.

Our agreement on all labeled data is $\kappa = 0.47$, using Fleiss's Kappa (Fleiss, 1971), considered "moderate agreement" by Landis and Koch (1977). We observe that annotators achieved perfect agreement on 39% of the data, and for another 32% the majority was 3/5 or less.[6] This suggests that our data displays significant variation in how stress is expressed, which we explore in the next section.

[6]It is possible for the majority to be less than 3/5 when more than 5 annotations were solicited.

4 Data Analysis

While all our data has the same genre and personal narrative style, we find distinctions among domains with which classification systems must contend in order to perform well, and distinctions between stressful and non-stressful data which may be useful when developing such systems. Posters in each subreddit express stress, but we expect that their different functions and stressors lead to differences in how they do so in each subreddit, domain, and broad category.

By domain. We examine the vocabulary patterns of each domain on our training data only, not including unlabeled data so that we may extend our analysis to the label level. First, we use the word categories from the Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2015), a lexicon-based tool that gives scores for psychologically relevant categories such as sadness or cognitive processes, as a proxy for topic prevalence and expression variety. We calculate both the percentage of tokens per domain which are included in a specific LIWC word list, and the percentage of words in a specific LIWC word list that appear

Domain	"Negemo" %	"Negemo" Coverage	"Social" %	"Anxiety" Coverage
Abuse	2.96%	39%	12.03%	58%
Anxiety	3.42%	37%	6.76%	62%
Financial	1.54%	31%	8.06%	42%
PTSD	3.29%	42%	7.95%	61%
Social	2.36%	38%	13.21%	59%
All	2.71%	62%	9.62%	81%

Table 3: **LIWC Analysis by Domain.** Results from our analysis using LIWC word lists. Each term in quotations refers to a specific word list curated by LIWC; percentage refers to the percent of words in the domain that are included in that word list, and coverage refers to the percent of words in that word list which appear in the domain.

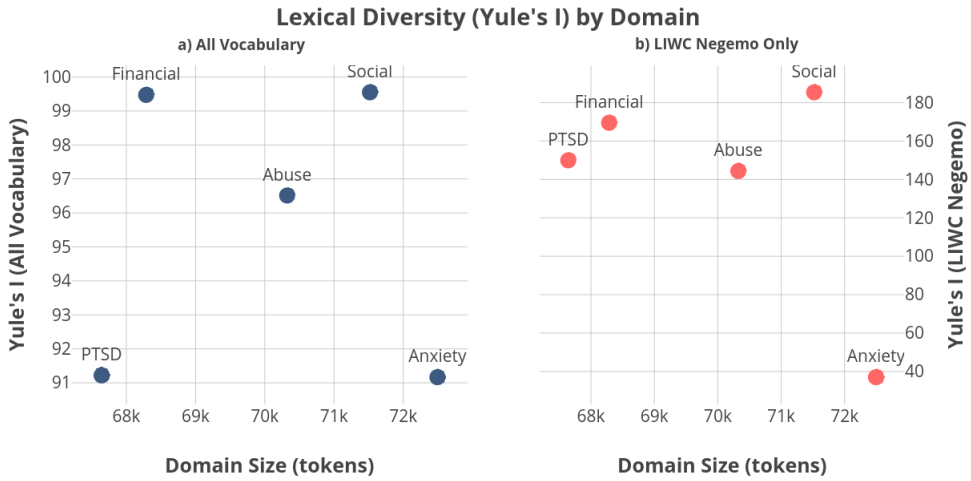

Figure 2: **Lexical Diversity by Domain.** Yule's I measure (on the y-axes) is plotted against domain size (on the x-axes) and each domain is plotted as a point on two graphics. a) measures the lexical diversity of all words in the vocabulary, while b) deletes all words that were not included in LIWC's negative emotion word list.

in each domain ("coverage" of the domain).

Results of the analysis are highlighted in Table 3. We first note that variety of expression depends on domain and topic; for example, the variety in the expression of negative emotions is particularly low in the financial domain (with 1.54% of words being negative emotion ("negemo") words and only 31% of "negemo" words used). We also see clear topic shifts among domains: the interpersonal domains contain roughly 1.5 times as many social words, proportionally, as the others; and domains are stratified by their coverage of the anxiety word list (with the most in the mental illness domains and the least in the financial domain).

We also examine the overall lexical diversity of each domain by calculating Yule's I measure (Yule, 1944). Figure 2 shows the lexical diversity of our data, both for all words in the vocabulary and for only words in LIWC's "negemo" word list. Yule's I measure reflects the repetitiveness of the data (as opposed to the broader coverage measured by our LIWC analysis). We notice exceptionally low lexical diversity for the mental illness domains, which we believe is due to the structured, clinical language surrounding mental illnesses. For example, posters in these domains discuss topics such as symptoms, medical care, and diagnoses (Figure 1, Table 2). When we restrict our analysis to negative emotion words, this pattern persists only for anxiety; the PTSD domain has comparatively little lexical variety, but what it does have contributes to its variety of expression for negative emotions.

By label. We perform similar analyses on data labeled stressful or non-stressful by a majority of annotators. We confirm some common results in the mental health literature, including that stressful data uses more first-person pronouns (perhaps reflecting increased self-focus) and that non-stressful data uses more social words (perhaps reflecting a better social support network).

Label	1st-Person %	"Posemo" %	"Negemo" %	"Anxiety" Cover.	"Social" %
Stress	9.81%	1.77%	3.54%	78%	8.35%
Non-Stress	6.53%	2.78%	1.75%	67%	11.15%

Table 4: **LIWC Analysis by Label.** Results from our analysis using LIWC word lists, with the same definitions as in Table 3. First-person pronouns ("1st-Person") use the LIWC "I" word list.

Measure	Stress	Non-Stress
% Conjunctions	0.88%	0.74%
Tokens/Segment	100.80	93.39
Clauses/Sentence	4.86	4.33
F-K Grade	5.31	5.60
ARI	4.39	5.01

Table 5: **Complexity by Label.** Measures of syntactic complexity for stressful and non-stressful data.

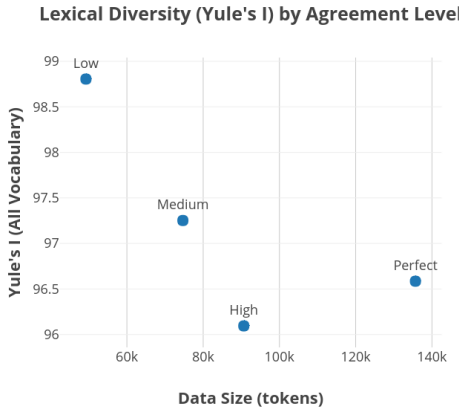

Figure 3: **Lexical Diversity by Agreement.** Yule's I measure (on the y-axis) is plotted against domain size (on the x-axis) for each level of annotator agreement. Perfect means all annotators agreed; High, 4/5 or more; Medium, 3/5 or more; and Low, everything else.

Additionally, we calculate measures of syntactic complexity, including the percentage of words that are conjunctions, average number of tokens per labeled segment, average number of clauses per sentence, Flesch-Kincaid Grade Level (Kincaid et al., 1975), and Automated Readability Index (Senter and Smith, 1967). These scores are comparable for all splits of our data; however, as shown in Table 5, we do see non-significant but persistent differences between stressful and non-stressful data, with stressful data being generally longer and more complex but also rated simpler by readability indices. These findings are intriguing and can be explored in future work.

By agreement. Finally, we examine the differences among annotator agreement levels. We find an inverse relationship between the lexical variety and the proportion of annotators who agree, as shown in Figure 3. While the amount of data and lexical variety seem to be related, Yule's I measure controls for length, so we believe that this trend reflects a difference in the type of data that encourages high or low agreement.

5 Methods

In order to train supervised models, we group the labeled segments by post and randomly select 10% of the posts ($\approx$ 10% of the labeled segments) to form a test set. This ensures that while there is a reasonable distribution of labels and domains in the train and test set, the two do not explicitly share any of the same content. This results in a total of 2,838 train data points (51.6% labeled stressful) and 715 test data points (52.4% labeled stressful). Because our data is relatively small, we train our traditional supervised models with 10-fold cross-validation; for our neural models, we break off a further random 10% of the training data for validation and average the predictions of 10 randomly-initialized trained models.

In addition to the words of the posts (both as bag-of-n-grams and distributed word embeddings), we include features in three categories:

Lexical features. Average, maximum, and minimum scores for pleasantness, activation, and imagery from the Dictionary of Affect in Language (DAL) (Whissel, 2009); the full suite of 93 LIWC features; and sentiment calculated using the `Pattern` sentiment library (Smedt and Daelemans, 2012).

Syntactic features. Part-of-speech unigrams and bigrams, the Flesch-Kincaid Grade Level, and the Automated Readability Index.

Social media features. The UTC timestamp of the post; the ratio of upvotes to downvotes on the post, where an upvote roughly corresponds to a reaction of "like" and a downvote to "dislike" (**upvote ratio**); the net score of the post (**karma**) (calculated by Reddit, $n_{\text{upvotes}} - n_{\text{downvotes}}$)[7]; and

[7] https://www.reddit.com/wiki/faq

the total number of comments in the entire thread under the post.

5.1 Supervised Models

We first experiment with a suite of non-neural models, including Support Vector Machines (SVMs), logistic regression, Naïve Bayes, Perceptron, and decision trees. We tune the parameters for these models using grid search and 10-fold cross-validation, and obtain results for different combinations of input and features.

For input representation, we experiment with bag-of-n-grams (for $n \in \{1..3\}$), Google News pre-trained Word2Vec embeddings (300-dimensional) (Mikolov et al., 2013), Word2Vec embeddings trained on our large unlabeled corpus (300-dimensional, to match), and BERT embeddings trained on our unlabeled corpus (768-dimensional, the top-level [CLS] embedding) (Devlin et al., 2019). We experiment with subsets of the above features, including separating the features by category (lexical, syntactic, social) and by magnitude of the Pearson correlation coefficient (r) with the training labels. Finally, we stratify the training data by annotator agreement, including separate experiments on only data for which all annotators agreed, data for which at least 4/5 annotators agreed, and so on.

We finally experiment with neural models, although our dataset is relatively small. We train both a two-layer bidirectional Gated Recurrent Neural Network (GRNN) (Cho et al., 2014) and Convolutional Neural Network (CNN) (as designed in Kim (2014)) with parallel filters of size 2 and 3, as these have been shown to be effective in the literature on emotion detection in text (e.g., Xu et al. (2018); Abdul-Mageed and Ungar (2017)). Because neural models require large amounts of data, we do not cull the data by annotator agreement for these experiments and use all the labeled data we have. We experiment with training embeddings with random initialization as well as initializing with our domain-specific Word2Vec embeddings, and we also concatenate the best feature set from our non-neural experiments onto the representations after the recurrent and convolutional/pooling layers respectively.

Finally, we apply BERT directly to our task, fine-tuning the pretrained BERT-base[8] on our classification task for three epochs (as performed in Devlin et al. (2019) when applying BERT to any task). Our parameter settings for our various models are available in the appendix.

[8]Using the implementation available at https://github.com/huggingface/pytorch-transformers

6 Results and Discussion

We present our results in Table 6. Our best model is a logistic regression classifier with Word2Vec embeddings trained on our unlabeled corpus, high-correlation features ($\geq$ 0.4 absolute Pearson's r), and high-agreement data (at least 4/5 annotators agreed); this model achieves an F-score of 79.8 on our test set, a significant improvement over the majority baseline, the n-gram baseline, and the pre-trained embedding model, (all by the approximate randomization test, $p < 0.01$). The high-correlation features used by this model are LIWC's clout, tone, and "I" pronoun features, and we investigate the use of these features in the other model types. Particularly, we apply different architectures (GRNN and CNN) and different input representations (pretrained Word2Vec, domain-specific BERT).

We find that our logistic regression classifier described above achieves comparable performance to BERT-base (approximate randomization test, $p > 0.5$) with the added benefits of increased interpretability and less intensive training. Additionally, domain-specific word embeddings trained on our unlabeled corpus (Word2Vec, BERT) significantly outperform n-grams or pretrained embeddings, as expected, signaling the importance of domain knowledge in this problem.

We note that our basic deep learning models do not perform as well as our traditional supervised models or BERT, although they consistently, significantly outperform the majority baseline. We believe this is due to a serious lack of data; our labeled dataset is orders of magnitude smaller than neural models typically require to perform well. We expect that neural models can make good use of our large unlabeled dataset, which we plan to explore in future work. We believe that the superior performance of the pretrained BERT-base model (which uses no additional features) on our dataset supports this hypothesis as well.

In Table 7, we examine the impact of different feature sets and levels of annotator agreement on our logistic regressor with domain-specific Word2Vec embeddings and find consistent patterns supporting this model. First, we

Model	P	R	F
Majority baseline	0.5161	1.0000	0.6808
CNN + features*	0.6023	0.8455	0.7035
CNN*	0.5840	0.9322	0.7182
GRNN w/ attention + features*	0.6792	0.7859	0.7286
GRNN w/ attention*	0.7020	0.7724	0.7355
n-gram baseline*	0.7249	0.7642	0.7441
n-grams + features*	0.7474	0.7940	0.7700
LogReg w/ pretrained Word2Vec + features	0.7346	0.8103	0.7706
LogReg w/ fine-tuned BERT LM + features*	0.7704	0.8184	0.7937
LogReg w/ domain Word2Vec + features*	0.7433	0.8320	0.7980
BERT-base*	0.7518	0.8699	0.8065

Table 6: **Supervised Results.** Precision (P), recall (R), and F1-score (F) for our supervised models. Our best model achieves 79.80 F1-score on our test set, comparable to the state-of-the-art pretrained BERT-base model. In this table, "features" always refers to our best-performing feature set (≥ 0.4 absolute Pearson's r). Models marked with a * show a significant improvement over the majority baseline (approximate randomization test, $p < 0.01$).

		Agreement Threshold for Data			
		Any Majority	60% (3/5)	80% (4/5)	100% (5/5)
Features	None	75.40	76.31	78.48	77.69
	All	76.90	77.12	77.10	78.28
	LIWC	77.91	78.91	78.16	77.66
	DAL	75.58	77.06	78.05	77.06
	Lexical	76.42	77.92	77.54	77.88
	Syntactic	74.63	75.49	76.66	76.19
	Social	76.67	76.45	78.38	78.06
	$\|r\| \geq 0.4$	77.44	78.76	**79.80**	78.52
	$\|r\| \geq 0.3$	77.01	78.28	79.38	78.31
	$\|r\| \geq 0.2$	77.53	78.61	79.02	78.28
	$\|r\| \geq 0.1$	76.61	77.07	76.32	77.48

Table 7: **Feature Sets and Data Sets.** The results of our best classifier trained on different subsets of features and data. Features are grouped by type and by magnitude of their Pearson correlation with the train labels (no features had an absolute correlation greater than 0.5); data is separated by the proportion of annotators who agreed. Our best score (corresponding to our best non-neural model) is shown in bold.

see a tradeoff between data size and data quality, where lower-agreement data (which can be seen as lower-quality) results in worse performance, but the larger 80% agreement data consistently outperforms the smaller perfect agreement data. Additionally, LIWC features consistently perform well while syntactic features consistently do not, and we see a trend towards the quality of features over their quantity; those with the highest Pearson correlation with the train set (which all happen to be LIWC features) outperform sets with lower correlations, which in turn outperform the set of all features. This suggests that stress detection is a highly lexical problem, and in particular, resources developed with psychological applications in mind, like LIWC, are very helpful.

Finally, we perform an error analysis of the two best-performing models. Although the dataset is nearly balanced, both BERT-base and our best logistic regression model greatly overclassify stress, as shown in Table 8, and they broadly overlap but do differ in their predictions (disagreeing with one another on approximately 100 instances).

We note that the examples misclassified by both models are often, though not always, ones with low annotator agreement (with the average percent agreement for misclassified examples being 0.55 for BERT and 0.61 for logistic regression). Both models seem to have trouble with less explicit expressions of stress, framing negative ex-

		Gold	
		0	**1**
LogReg	**0**	241	105
	1	49	320

		Gold	
		0	**1**
BERT	**0**	240	106
	1	48	321

		BERT	
		0	**1**
LogReg	**0**	237	51
	1	53	374

Table 8: **Confusion Matrices**. Confusion matrices of our best models and the gold labels. 0 represents data labeled not stressed while 1 represents data labeled stressed.

Text	Gold Label	Agreement	Subreddit Name	Models Failed
Hello everyone, A very close friend of mine was in an accident a few years ago and deals with PTSD. He has horrific nightmares that wake him up and keep him in a state of fright. We live in separate provinces, so when he does have his dreams it is difficult to comfort him. Each time he calls, and I struggle with what to say on the phone.	Not Stress	60%	ptsd	Both
I asked the other day if they've set a date. He laughed in my face and said 'no' as if it were the most ridiculous thing he's ever heard. He comes home late, and showers immediately. Then, he showers every morning before he leaves. He doesn't talk to my mum and I, at all, and he's cagey and secretive about everything, to the point of hostility towards my sister.	Stress	60%	domesticviolence	BERT
If he's the textbook abuser, she is the textbook victim. She keeps giving him chances and accepting his apologies and living in this cycle of abuse. She thinks she's the one doing something wrong. I keep telling her that the only thing she is doing wrong is staying with this guy and thinking he will change. I tell her she does not deserve this treatment.	Not Stress	100%	domesticviolence	LogReg

Table 9: **Error Analysis Examples**. Examples of test samples our models failed to classify correctly."BERT" refers to the state-of-the-art BERT-base model, while "LogReg" is our best logistic regressor described in section 6.

periences in a positive or retrospective way, and stories where another person aside from the poster is the focus; these types of errors are difficult to capture with the features we used (primarily lexical), and further work should be aware of them. We include some examples of these errors in Table 9, and further illustrative examples are available in the appendix.

7 Conclusion and Future Work

In this paper, we present a new dataset, Dreaddit, for stress classification in social media, and find the current baseline at 80% F-score on the binary stress classification problem. We believe this dataset has the potential to spur development of sophisticated, interpretable models of psychological stress. Analysis of our data and our models shows that stress detection is a highly lexical problem benefitting from domain knowledge, but we note there is still room for improvement, especially in incorporating the framing and intentions of the writer. We intend for our future work to use this dataset to contextualize stress and offer explanations using the content features of the text. Additional interesting problems applicable to this dataset include the development of effective distant labeling schemes, which is a significant first step to developing a quantitative model of stress.

Acknowledgements

We would like to thank Fei-Tzin Lee, Christopher Hidey, Diana Abagyan, and our anonymous reviewers for their insightful comments during the writing of this paper. This research was funded in part by a Presidential Fellowship from the Fu Foundation School of Engineering and Applied Science at Columbia University.

References

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

Fares Al-Shargie, Masashi Kiguchi, Nasreen Badruddin, Sarat C. Dass, and Ahmad Fadzil Mohammad Hani. 2016. Mental stress assessment using simultaneous measurement of eeg and fnirs. *Biomedical Optics Express*, 7(10):38823898.

Andrew P. Allen, Paul J. Kennedy, John F. Cryan, Timothy G. Dinan, and Gerard Clarke. 2014. Biological and psychological markers of stress in humans: Focus on the trier social stress test. *Neuroscience & Biobehavioral Reviews*, 38:94124.

Marilia A. Calcia, David R. Bonsall, Peter S. Bloomfield, Sudhakar Selvaraj, Tatiana Barichello, and Oliver D. Howes. 2016. Stress and neuroinflammation: a systematic review of the effects of stress on microglia and the implications for mental illness. *Psychopharmacology*, 233(9):1637–1650.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1724–1734.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*.

Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. 2018. SMHD: a large-scale resource for exploring online language usage for multiple mental health conditions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1485–1497, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Sharath Chandra Guntuku, Anneke Buffone, Kokil Jaidka, Johannes C. Eichstaedt, and Lyle H. Ungar. 2018. Understanding and measuring psychological stress using social media. *CoRR*, abs/1811.07430.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *CoRR*, abs/1408.5882.

J. Peter Kincaid, Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159174.

Huijie Lin, Jia Jia, Jiezhong Qiu, Yongfeng Zhang, Guangyao Shen, Lexing Xie, Jie Tang, Ling Feng, and Tat-Seng Chua. 2017. Detecting stress based on social interactions in social networks. *IEEE Transactions on Knowledge and Data Engineering*, 29(09):1820–1833.

Sonia J. Lupien, Bruce S. McEwen, Megan R. Gunnar, and Christine Heim. 2009. Effects of stress throughout the lifespan on the brain, behaviour and cognition. *Nature Reviews Neuroscience*, 10(6):434–445.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA. Curran Associates Inc.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015.

R.J. Senter and E.A. Smith. 1967. Automated readability index.

Tom De Smedt and Walter Daelemans. 2012. Pattern for python. *Journal of Machine Learning Research*, 13:2063–2067.

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

Cynthia Whissel. 2009. Using the revised dictionary of affect in language to quantify the emotional undertones of samples of natural language. *Psychological Reports*, 105(2):509–521.

Genta Indra Winata, Onno Pepijn Kampman, and Pascale Fung. 2018. Attention-based LSTM for psychological stress detection from spoken language using distant supervision. *CoRR*, abs/1805.12307.

Peng Xu, Andrea Madotto, Chien-Sheng Wu, Ji Ho Park, and Pascale Fung. 2018. Emo2vec: Learning generalized emotion representation by multi-task training. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA@EMNLP 2018, Brussels, Belgium, October 31, 2018*, pages 292–298.

George Udny Yule. 1944. *The statistical study of literary vocabulary*. Cambridge Univ. Pr.

Xin Zuo, Tian Li, and Pascale Fung. 2012. A multilingual natural stress emotion database. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 1174–1178, Istanbul, Turkey. European Language Resources Association (ELRA).