Who Cares About Unlabeled Data? RoBERTa Does!

Dorian Smoljan, Fani Sentinella-Jerbić, Vladimir Rzaev

University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
{dorian.smoljan, fani.sentinella.jerbic, vladimir.rzaev98}@gmail.com

Abstract

Stress has become an irreplaceable part of modern life, and so the task of its detection and prevention is more important than ever. In recent years, social media has surfaced as an excellent source of stress-laden data, leading to the creation of specific social media based datasets focused on stress detection, such as Dreaddit. In this work, we take the natural step of utilizing the available data. We build upon prior research on the Dreaddit dataset and propose a number of improvements, the main ones being the use of additional unlabeled data and the use of Empath features.

1. Introduction

In today's busy world, stress feels omnipresent, and the number of people experiencing it on a daily basis only continues to grow. This is why various methods for detecting stress are more important than ever, as detection is the first part of identifying the many complex causes and treatments of stress. One large and relatively easy-to-access information pool of stress-laced data are social media platforms, such as Facebook, Reddit, or Twitter. Reddit is especially suited for this problem, as it allows users to anonymously post about their experiences, which often results in people talking more openly about their experiences, including stressful ones.

There have been a number of previous works focusing on detecting stress in social media, perhaps the most famous one being the work by Turcan and McKeown (2019), whose main contribution is the Dreaddit dataset for stress detection. Various other authors built upon Turcan et.al's work, most notably Knežević et al. (2021), who proposed a new model which managed to achieve state-of-the-art performance on this specific dataset. However, we find that the majority of previous work focused on using the labeled part of the Dreaddit dataset, failing to include the vastly larger unlabeled part.

In our work, we experiment with modifications of the solution used by Knežević et al. (2021), in order to determine their potential impact. The first is the use of the unlabeled Reddit dataset to additionally train the RoBERTa model, increasing its domain knowledge and allowing it to use word embeddings better suited to the problem at hand. The second is replacing the existing LIWC features used in the Dreaddit dataset with more accessible Empath features. Lastly, we experiment with different machine learning models for the final classification. We evaluate and compare all approaches used based on chosen performance metrics.

2. Related Work

As stress analysis in social media is a relatively new topic, new models are constantly being developed and the state-of-the-art changes constantly. A couple of related works include the work by Winata et al. (2018), in which LSTMs are used to detect stress based on a questionnaire, and unlabeled Twitter data is used to enhance learning, or the work of Lin et al. (2017), in which CNNs were used to detect stress from microblogging websites, using a set of both textual, visual and social features. However, the most significant existing work, upon which we base our own, is the work by Turcan and McKeown (2019), in which they introduce Dreaddit, a dataset for stress detection based on lexical characteristics of Reddit posts. They test the performance of the dataset using a couple of models and find that hybrid approach of using both word embbedings and lexical features with Logistic Regression works best. Matošević et al. (2021) build upon their work by comparing performance of different transformer-based models on the Dreaddit dataset. They conclude RoBERTa performs best for stress analysis with an F1 score of 0.8268. This is also confirmed by Selvadass et al. (2022). Another related paper is by Knežević et al. (2021) where a new ensemble-like model is introduced, which manages to achieve state-of-the-art results with an F1 score of 0.8355 on the aforementioned dataset. They train a RoBERTa classifier and feed its decision as a feature to the Logistic Regression model, alongside lexical LIWC features given in the original dataset. A visual representation of their model and its main components can be found in Figure 1.

3. Datasets

In our experiments, we used four different datasets. The first and most important one was **Dreaddit**, a dataset first introduced by Turcan and McKeown (2019), and used in the Knežević et al. (2021)'s work we build upon. It consists of Reddit posts from 10 different subreddits, all of which are the ones with a higher probability of having a stress-related post (i.e. r/PTSD, r/domesticviolence etc.). In total 187,444 Reddit posts were collected and 3,553 of those have been annotated with a binary stress label, namely as either stress or not stress, using the majority vote as the final label. They also collect available social characteristics of each post as well as calculate lexical features using the Linguistic Inquiry and Word Count Tool (Tausczik and Pennebaker, 2010), further referred to as LIWC. The labeled part of the dataset is then split into a train and test set in the ratio of 80:20. Unfortunately, the unlabeled part

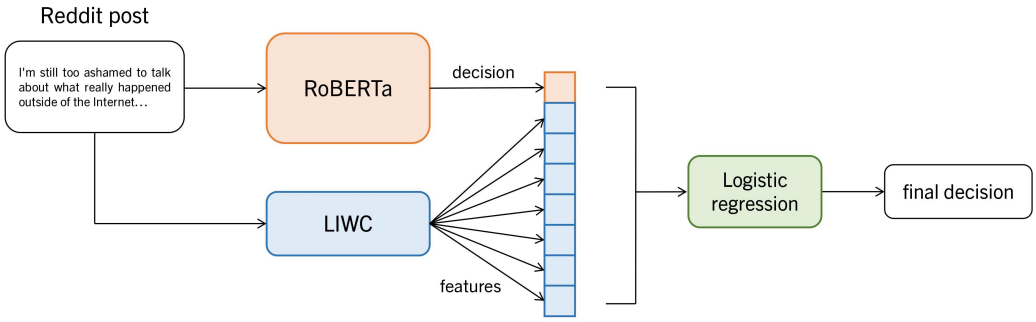

Figure 1: Representation of the Knežević et al. (2021)'s model.

of the dataset isn't included with the rest of the data, so we weren't able to include it in our experiments even though it served as the initial inspiration for our paper.

As we still needed a large unlabeled dataset with similar characteristics as the Dreaddit dataset in order to perform our desired experiments, we decided to build **our own unlabeled dataset**. We used the PushshiftAPI[1] through PSAW[2] API wrapper to scrape as many Reddit posts as we could from the same subreddits used in the Dreaddit dataset, to make the unlabeled data as similar as possible to the existing labeled data. Due to certain usage limits, we were able to scrape 60,844 posts in total. We publish the scraped dataset alongside this paper, in the hopes someone will find it useful.

In addition to the scraped data, we decided to incorporate an additional unlabeled dataset found on the Huggingface dataset repository, called **Reddit Mental Health Posts**[3]. It is comprised of Reddit posts taken from five mental health related subreddits: r/adhd, r/depression, r/aspergers, r/ocd, and r/ptsd. Although not all of the subreddits included are the same as in the Dreaddit dataset, we figured the domain is similar enough that it could be of use. The original dataset consisted of 151,381 posts, but after removing the deleted and removed posts, we were left with 64,853 posts in total.

Finally, we used a **large unlabeled dataset from TakeLab**, kindly provided to us by research associate Josip Jukić, consisting of 2,430,112 posts and comments from various Reddit subreddits (not counting those shorter than three words). We incorporated this dataset as its domain was not related to mental health or stress, whereas other datasets domains were. This enabled us to determine whether the domain similarity of the unlabeled dataset to the domain of the task effects performance or not.

4. Experimental Setup

We recognize previously mentioned Knežević et al. (2021)'s approach as promising but also identify room for improvement across all three parts of their model; the RoBERTa classifier, the lexical features, and the final decision classifier.

4.1. RoBERTa Classifier

One of the works which had a major impact in the field of NLP in the last couple of years was certainly BERT (Devlin et al., 2018), a transformer model for language understanding. It leveraged the power of pretraining on large amounts of data, such as WikiText, with the possibility of fine-tuning the topmost layer of the model to achieve state-of-the-art performance in a wide range of NLP tasks. Since then, a number of improvements to the base BERT model have been published, one of them being the RoBERTa model by (Liu et al., 2019), which improves on the BERT model, mainly by tuning parts of its training process.

There are two distinct parts when it comes to tuning RoBERTa for a specific NLP task. The first and most important one is fine-tuning the head of the model on the labeled data we want to classify, label, etc. The head is a (small) neural network that takes the outputs of the core transformer model and maps them to an appropriate output space. So when training the head, we're effectively just training the neural network at the output of the transformer model. However, if in addition to labeled data we also possess a large enough quantity of unlabeled data, it is possible to tune RoBERTa in an additional way. We do this by continuing the training of its (base) transformer model on our unlabeled data, in a process known as "domain adaptation". This is done by training the model on a masked language modeling task, in which certain parts of the input sequence are labeled with a special "MASK" token, and the model's task is to predict the masked tokens based on their context. By doing this, we allow the model to learn word embeddings better suited to the problem domain, which can in turn increase the performance on our desired task. The main benefit of this approach is that it requires only unlabeled data, which is usually far more abundant and cheaper to acquire than labeled data. The entire process of tuning the RoBERTa model is shown in Figure 2.

Knežević et al. (2021) used predictions from a fine-tuned RoBERTa transformer as one of the features of the final classifier. We identify the opportunity of additionally training RoBERTa for the stress detection task using the explained domain adaptation process, enabling it to learn

[1] https://github.com/pushshift/api
[2] https://github.com/dmarx/psaw
[3] https://huggingface.co/datasets/solomonk/reddit_mental_health_posts

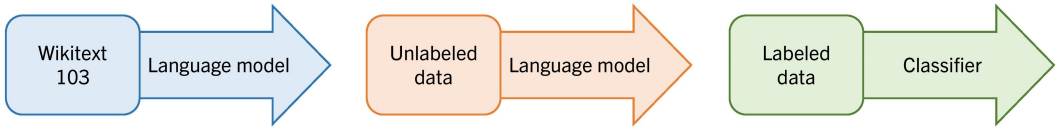

Figure 2: Illustration of the RoBERTa tuning process, modified from [5].

Table 1: Hyperparameters for fine-tuning RoBERTa on stress detection task

Hyperparameter	Value
Tokenizer max length	512
Train batch size	10
Learning rate	5e-6
No. epochs	4
L2 regularization factor	1e-2

Table 2: Results of fine-tuned RoBERTa classifier for accuracy, precision, recall and F1 score as evaluation metrics.

Unlabeled dataset	Accuracy	Precision	Recall	F1
Base	0.8097	0.8494	**0.7879**	0.8175
Scraped	0.8181	0.8804	0.7877	0.8315
Mental health	**0.8194**	**0.8933**	0.7791	**0.8323**
General Reddit posts	0.7722	0.8341	0.7360	0.7820

word embeddings better suited for stress detection, and thus improve its performance on the classification task. After domain adaptation using unlabeled data, we proceed in the usual manner by fine-tuning the head of the RoBERTa model on the labeled Dreaddit dataset.

As for hyperparameters, for training the RoBERTa language model, we used the default parameters given in the Huggingface training script[4]. For training the head, we used the Adam optimizer (Kingma and Ba, 2014) and achieved the best results using the parameters in Table 1. We used the same hyperparameters for training the head as (Knežević et al., 2021) in order to compare the achieved results.

4.2. "Lexical" Features

For lexical features, Knežević et al. (2021) used previously prepared features from Dreaddit obtained through LIWC, the industry-leading software tool for text analysis. It works by counting percentage of words in text that correspond to several psychologically meaningful categories in a lexicon. We try using a more widely accessible tool because LIWC is a paid software. We craft new features through the free Empath tool (Fast et al., 2016), which uses neural embeddings to discover topic-related terms instead of a lexicon. One additional strength of it is the ability to create new lexical categories by using a small set of seed words, for example, the words "dog" and "cat" for the category of "pets". Although we don't make use of this ability, it could be utilized in future work.

We compute predefined Empath categories for each post from the dataset and use them as new features. We remove newly created features with high mutual correlation as they would disturb the performance of machine learning models. For this, we use the ϕ_k correlation measure[6] (Baak et al., 2020) with a threshold of 0.85. The choice of this correlation measure stems from claims that it captures non-linear dependencies and works well on data with mixed categorical and numerical features. We would argue the RoBERTa prediction is certainly more categorical than other features used so we opt for this measure.

Lastly, we compute if the performance of the model changes significantly as a result of the replacement using a significance level of 0.05 with the statistical McNemar's Test (McNemar, 1947).

4.3. Final Decision Classifier

Original Dreaddit creators concluded Logistic Regression performed best on the dataset, however, Knežević et al. (2021) didn't provide it as the motivation behind using it for their model as well. Therefore we also inspect the performance of other machine learning algorithms on this model architecture. We use Logistic Regression, Support Vector Machine, and Random Forest. Additionally, we perform a grid search of each model's hyperparameters through 5-fold cross-validation on the train set.

5. Results and Discussion

Performance of different model variations is presented in Table 2 and Table 3 using accuracy, precision, recall, and F1 score as evaluation metrics. We discuss this further through the commentary of each model component.

Fine-tuned RoBERTa. The results of fine-tuning the RoBERTa head for the stress classification task are shown in Table 2. We can see that both RoBERTa models additionally trained on domain-specific unlabeled data (scraped

[4] https://github.com/huggingface/transformers/blob/main/examples/pytorch/language-modeling/run_mlm.py

[5] https://huggingface.co/course/chapter7/3?fw=tf

[6] https://github.com/kaveio/phik

Table 3: Performance of different model architectures used.

Model	Accuracy	Precision	Recall	F1
Majority baseline	0.5161	0.5161	0.5161	0.5161
RoBERTa (scraped) + Empath + RF	0.8126	0.7752	0.8970	0.8317
RoBERTa (scraped) + LIWC + RF	0.8168	0.7793	0.8997	0.8352
RoBERTa (scraped) + LIWC + SVM	0.8168	0.7742	0.9106	0.8369
RoBERTa (scraped) + Empath + SVM	0.8168	0.7742	0.9106	0.8369
RoBERTa (scraped) + Empath + LR	0.8224	0.7881	0.8970	0.8390
RoBERTa (scraped) + LIWC + LR	0.8224	0.7827	0.9079	0.8407
RoBERTa (mental health) + LIWC + RF	0.8126	0.7739	0.8997	0.8321
RoBERTa (mental health) + Empath + RF	0.8140	0.7719	0.9079	0.8344
RoBERTa (mental health) + LIWC + LR	0.8196	0.7791	0.9079	0.8385
RoBERTa (mental health) + LIWC + SVM	0.8182	0.7710	**0.9214**	0.8395
RoBERTa (mental health) + Empath + SVM	0.8181	0.7710	**0.9214**	0.8395
RoBERTa (mental health) + Empath + LR	0.8238	0.7845	0.9079	**0.8417**
RoBERTa (base) + LIWC + LR - Knežević et al. (2021)	0.81259	0.78043	0.88618	0.82995
RoBERTa (base) + LIWC subset + LR - Knežević et al. (2021)	**0.82657**	**0.81654**	0.85637	0.83598

stress-related posts and mental health related posts) managed to outperform the base RoBERTa model in both accuracy and F1 score. Using the McNemar's test with a significance level of 0.05 we determine that there is a significant difference in performance when comparing both of these additionally trained models to the base RoBERTa model, so we can conclude that the model achieves statistically significant improvement when additionally trained on domain-specific unlabeled data. However, the domain-unspecific general Reddit dataset did not manage to outperform the base RoBERTa model, achieving lower both F1 and accuracy scores, even though it was much larger than the other two unlabeled datasets. This indicates that it is important that the domain of the unlabeled data matches, or at least is similar to the domain of the given task, in this example stress detection.

LIWC vs Empath features. The replacement of the former with the latter showed no degradation of the model at the 0.05 level of significance. What is more, when used in combination with RoBERTa fine-tuned on the mental health dataset, Empath manages to outperform LIWC. We believe this is meaningful because Empath is a free alternative to LIWC and offers previously suggested additional functionalities. On the other hand, it uses neural representations instead of lexicons, making it a more complex and less interpretative model.

Machine learning algorithm for the final decision. SVM and Random Forest algorithms didn't show improvement in comparison to Logistic Regression, however, grid search did show improvement over Knežević et al. (2021)'s model, meaning they probably skipped this essential part of building their model.

Additionally, through simple analysis of several models based on coefficients assigned to features by Logistic Regression, we find that features that are related to social interactions such as "messaging", "family", "friends", "meeting", "speaking", and "listen" tend to influence the decision of the model toward the post not expressing stress. This could possibly be a hint to socializing being a good remedy for stress, which is similar to a conclusion psychologists have pointed to in the past (Lürzel et al., 2011); (VanKim and Nelson, 2013).

6. Conclusion and Future Work

In this paper, we expanded upon Knežević et al. (2021)'s model for detecting stress in Reddit posts on several fronts. First, we managed to improve the model by using unlabeled data to adapt RoBERTa classifier to the stress detection domain. Using smaller amount of domain-specific unlabeled data showed better performance than using a larger amount of generic data. We were also able to replace golden standard lexicon-based LIWC features with more accessible and flexible embeddings-based Empath features without degrading the model performance. Lastly, by trying several different machine learning algorithms with different hyperparameter settings for the final decision classifier, we concluded Logistic Regression is the best for this particular task and model architecture. We hope our contribution inspires future work, especially for the case of using the domain adaptation approach since today the amount of unlabeled data for almost any task is present in enormous amounts on the Internet.

7. Acknowledgements

We would like to thank research assistant Josip Jukić for providing us with the general Reddit dataset as well as associate professor Marko Subašić for giving us access to the hardware used during training.

Reproducibility

Our scraped data and model implementation is available at `https://github.com/fsentin/dreaddit`.

References

M. Baak, R. Koopman, H. Snoek, and S. Klous. 2020. A new correlation coefficient between categorical, ordinal and interval variables with Pearson characteristics. *Computational Statistics & Data Analysis*, 152:107043.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

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*, pages 4647–4657.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*.

Domagoj Knežević, Josipa Tomić, and Nikola Vugdelija. 2021. Linguistic Features Impact in Stress Analysis. In Josip Jukić and Jan Šnajder, editors, *Text Analysis and Retrieval Course Project Reports*, pages 41–43, Zagreb. Faculty of Electrical Engineering and Computing.

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. volume 29, pages 1820–1833.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Stephanie Lürzel, Sylvia Kaiser, and Norbert Sachser. 2011. Social interaction decreases stress responsiveness during adolescence. *Psychoneuroendocrinology*, 36(9):1370–1377.

Lovro Matošević, Filip Sosa, and Marko Gašparac. 2021. Stressformers: Transfering Knowledge for Stress Analysis in Social Media. In Josip Jukić and Jan Šnajder, editors, *Text Analysis and Retrieval Course Project Reports*, pages 56–59, Zagreb. Faculty of Electrical Engineering and Computing.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Salomi Selvadass, P. Malin Bruntha, and K. Priyadarshini. 2022. Stress Analysis in Social Media using ML Algorithms. In *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pages 1502–1506.

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):24–54.

Elsbeth Turcan and Kathy McKeown. 2019. Dreaddit: A Reddit Dataset for Stress Analysis in Social Media. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 97–107, Hong Kong, November. Association for Computational Linguistics.

Nicole A VanKim and Toben F Nelson. 2013. Vigorous Physical Activity, Mental Health, Perceived Stress, and Socializing among College Students. *American Journal of Health Promotion*, 28(1):7–15.

Genta Indra Winata, Onno Pepijn Kampman, and Pascale Fung. 2018. Attention-based LSTM for Psychological Stress Detection from Spoken Language Using Distant Supervision. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, apr.