Aalto University
CS-C4100 - Digital Health and Human Behavior

Project, topic 6:
Stress Detection from Social Media Posts

Contents

1 Introduction

Over the course of the 21st century, social media has become a space where people express their views and sometimes even disclose their own emotional states and personal experiences. While often sharing positive life experiences, users may express anxiety, overwhelm and other forms of negative emotional experiences that reflect stress. In the context of digital health, this creates a possibility to study stress-related language with machine learning and natural language processing techniques. Advanced methods of classifying stress-positive cases may help in monitoring population-level stress levels and well-being trends in populations or support interventions for high-stress individuals. The problem is challenging as stress is often not explicitly stated in social media contexts but it can be predicted to a degree from writing style and length.

Recent work (Rastogi et al., 2022) has introduced high-quality datasets to study this phenomenon using data from multiple sources and with different structures. This project studies how the content of these datasets can be used to build machine learning models that predict users' stress in binary classification form between stress-positive and stress-negative cases. Social media platforms are especially useful for machine learning training tasks since people often vent about their problems and thus provide textual representations and cues of the text content that is related to stress-positive individuals (Inamdar et al., 2023).

Previous work (Inamdar et al., 2023; Rastogi et al., 2022) suggest that transformer-based models outperform simple lexical-based models such as Logistic Regression on in-domain settings. This motivates the study of cross-domain performance to see how generalizable the performance can be across platforms. Harrigian et al. (2020) found that models trained on twitter tend to generalize better on Reddit than vice-versa. In this project I study how well models trained on Reddit perform on both within Reddit datasets as well as out-of-domain Twitter datasets. The results show that the models perform well on in-domain data but fail to generalize for cross-domain data.

2 Problem Formulation

For reasons mentioned in the introduction section, stress detection from textual social media posts is an important problem in order to generally understand the userbase across topics. Stress detection might also help find certain users who need mental health assistance. The goal of this project is to automatically identify from textual input, if a post on social media reflects stress-related emotional state. As mentioned, the problem is particularly interesting for large-scale monitoring of differences in mental health states of users.

In this project, stress detection with machine learning techniques is studied using four publicly available datasets collected from Reddit and Twitter (Rastogi et al., 2022). Machine learning models are then trained on this data and their ability to detect elevated stress on unseen textual inputs is assessed. The research questions formulated to study this problem are:

1. Can a pre-trained Transformer-based model (BERT) significantly outperform a traditional ML model (Logistic Regression) in social media stress detection trained on the same platform data?

2. How do stress detection models trained on data from one platform generalize when testing on data from other platforms (Reddit vs. Twitter)?

The research questions can be defined as binary supervised classification problem due to the nature of labeling the textual inputs as stress-positive ”1” or stress-negative ”0”. In other words, the goal is for the models to learn a function f that maps text to stress labels and can perform stress detection on new textual data with high predictive accuracy. Given the nature of supervised models learning from labels, the models should ideally capture the essential cues from textual data that relates to stress-positive posts but not overfit to the training data. Together, studying these research questions gives insight into how well stress can be detected from social media posts with different machine learning models and how well the training on data from a certain platform generalize to other platforms.

3 Dataset Description

This project uses four high-quality datasets provided by Rastogi et al. (2022) to study stress-detection from textual inputs. The datasets were collected from two major social media platforms ”Twitter” and ”Reddit”. They are designed and preprocessed to support studying stress-detection from textual inputs analytically across platforms. In the dataset, each textual post is associated either with stress-positive label ”1” or stress-negative label ”0”.

The datasets provided have two variants from each platform totaling four datasets, that have differences in length and depth of the datapoints:

- **Reddit Title:** This dataset contains only the title of the posts, providing less textual content per datapoint. N=5522
- **Reddit Combi:** Combines the title and body of each post into a single datapoint. This enables the models to study long-form stress-related inputs. N=3098
- **Twitter Full:** Tweets with stress-positive and stress-negative hashtags without filtering

for marketing or promotional content. N=8439

- **Twitter Non-Advert:** Cleaned version of the Twitter Full -dataset which has been cleaned from advertisement-like content using PLM-based denoising approach. N=1972

Together these datasets enable studying cross-platfrom and short-form vs long-form training data performance as well as the effect of having advertisement-related textual content in the training data for the prediction capability of the models.

Social media data is noisy by default. Therefore, the dataset has been cleaned by removing links, repeated whitespaces and converting emojis to their corresponding textual representations. Extremely short samples were also filtered out in the preprocessing phase. Specifically for Twitter data, the full and non-advert datasets were differentiated using a transformer-based filter to remove advertisement-like text from the latter dataset. This reduced the size of the dataset for non-advert significantly (from 8900 datapoints to 2051) but imporved the quality of the data.

For the purposes of this project, I further cleaned the dataset from duplicates and reformatted the datasets to only have columns ”content” and ”label” since the column names were inconsistent in the datasets originally. The Reddit Combi -dataset contained both title and body of each post and I chose the content to be the body of each post to avoid overlap with the Reddit title dataset. The Twitter Full -datasets label column was initially named ”labels” which I changed to ”label” to be in line with the other datasets.

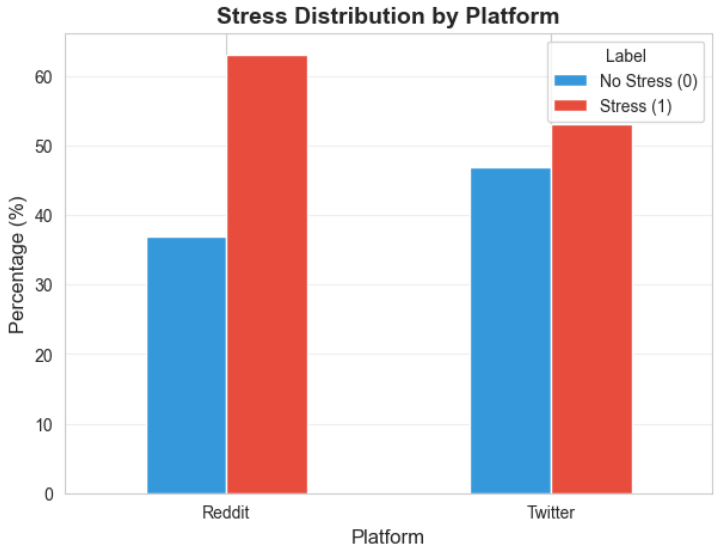

Figure 1: The distribution of stress-positive and stress-negative posts per platform

Figure 1 shows the distribution of stress-related posts for both platforms in the datasets after removing duplicates. Especially Twitter dataset seems to be fairly balanced whereas the Reddit

datasets have a slight overrepresentation of stress-positive posts. The distributions become relevant when choosing what metrics to study in terms of the success of the models.

Metric	Value
Average word count (No Stress, n=3186)	39.74
Average word count (Stress, n=5434)	89.50
Difference	49.76

Table 1: Text-Level Observation for Reddit Posts

Table 1 shows the average word counts for stress-positive and stress-negative posts. I chose to only compare word counts from reddit posts as the average word count differences between platforms and inconsistensies in the distributions might screw the data. For Reddit stress labels it turns out that stress-positive posts are significantly longer on average than stress-negative posts, which matches with my intuition. However, the difference in average word length (49.8 words!) is much more than I would have expected. This might relate to the need to ”vent” the negative feelings in the stress-positive subreddits resulting in longer posts.

Overall, the datasets offer excellent background to study stress detection in social media contexts. The notable difference in word length between stress-positive and stress-negative illustrate only one type of difference that there might exist between the emotional states of the users while writing posts. However, the length disparity may also negatively influence model behavior to attribute longer posts to be related to more stress-positivity which would be unideal.

4 Methods

The purpose of research question 1 is to benchmark classical and transformer-based models on stress detection within Reddit-retrieved posts and then compare the model behavior across platforms in research question 2. The data was examined by visualizing stress-positive and stress-negative distributions and The datasets contain high-quality data and was preprocessed carefully by the authors. Only some cleaning of duplicates was left to do. To train the models, the reddit Title and Reddit Combi datasets were combined into one dataset that contained all the textual content of the data points and their labels. This dataset was then split to train/test split with 80/20 distribution.

The experiments were done in Python using common scientific libraries pandas, numpy, and scikit-learn. Transformer model was implemented using the Hugging Face Transformers library and PyTorch. Random seeds were set to enhance reproducibility. Machine learning

models used in this project were logistic regression with TF-IDF for the classical baseline model and BERT for transformer-based model. For logistic regression, 5000 features were retained and stopwords were removed in preprocessing phase. The TF-IDF was fitted only on training data after which it transformed both training data and test sets. Training loss was monitored to confirm model convergence. The performance of logistic regression was assessed using Accuracy, Precision, Recall, and F1-Score.

The transformer-based model was BERT and it was fine-tuned using the BertForSequenceClassification framework for binary classification. The training was completed on my laptop CPU and it took about 1.5 hours. The model training was performed on Hugging Face Trainer API and the configuration included one epoch and a batch size of 32, with weight decay value of 0.01. The default learning rate was used. Logging occured every 50 steps to monitor training progress. Evaluation metrics for this transformer-based approach mirrored the Logistic Regression model: Accuracy, Precision, Recall and F1-Score.

For research question 2, I used the same models trained with the Reddit training set and evaluated them on test data retrieved from Twitter. Baseline Logistic Regression with TF-IDF and transformer-based BERT was run through with this new data. For Logistic Regression, TF-IDF vectorizer trained on Reddit data was used to transform the new data from Twitter and for BERT, the training and model weights were applied straight to the data from twitter without fine-tuning to evaluate cross-platform generalization.

The two models were compared on the performance of the same tests using Accuracy, Precision, Recall and F1-Score. This approach offered a straightforward way to study the models' differences in performance on the same data. To illustrate performane differences, a bar chart was created to compare the performance of both models in each metric side-by-side.

5 Results

After applying the cleaning steps describer earlier and combining Reddit Title and Reddit Combi (Rastogi et al., 2022) to one dataset, table 2 shows the label distribution of this new dataset. As one can see, this imbalance is not extreme but might severely affect model performance. This means that metrics that focus on stress class might appear strong.

Category	Count
Stress-positive	5,430
Stress-negative	3,183

Table 2: Stress vs. No Stress distribution

5.1 Research question 1

The first research question asks whether a transformer-based model (BERT) performs better than a traditional model (Logistic Regression with TF-IDF) when both are trained and tested on the same data obtained from Reddit.

Metric	Value
Accuracy	0.90
Precision (no stress)	0.90
Recall (no stress)	0.81
F1-score (no stress)	0.86
Precision (stress)	0.90
Recall (stress)	0.95
F1-score (stress)	0.92

Table 3: Logistic Regression Performance Metrics on Reddit test set

Table 3 shows logistic regression performance on Reddit test set. One can see that the model is quite reliable and the metrics vary around 90% throughout different metrics. Recall for stress-negative class stands out as lowest, likely due to imbalance of the label classes in the training data. This suggests that the classifier tends to lean towards predicting stress-positive content. Conversely, recall for stress-positive posts is rather high (95%). In other words, the model performs well in finding stress-positive posts and is unlikely to miss stressed textual content but may identify non-stressed posts as stress-positive.

Metric	Value
Accuracy	0.94
Precision (no stress)	0.93
Recall (no stress)	0.92
F1-score (no stress)	0.93
Precision (stress)	0.96
Recall (stress)	0.96
F1-score (stress)	0.96

Table 4: BERT Performance Metrics on Reddit test set

BERT was fine-tuned for one epoch on the same Reddit training set and evaluated on the same test set as Logistic regression. Table 4 shows the performance metrics for BERT across both classes. The transformer-based model seems to perform rather well on all performance metrics. For stress-negative, BERT achieved a F1-score of 0.93 while for stress-positive the model achieved 0.96 F1-score. Unlike with Logistic Regression, BERT does not show that strong decline in the asymmetry of the recall metric between stress-positive and stress-negative classes.

As Figure 2 shows, comparing the models confirms that transformer-based BERT performs better in the same-platform setting than baseline Logistic Regression in detecting stress-positive cases. Accuracy of BERT was 4.8 percentage point higher and F1-score 3.6 percentage points higher compared to the baseline model. Therefore answer to research question 1 within same platform training and testing is that transformer-based BERT seem to outperform traditional baseline Logistic Regression with TF-IDF.

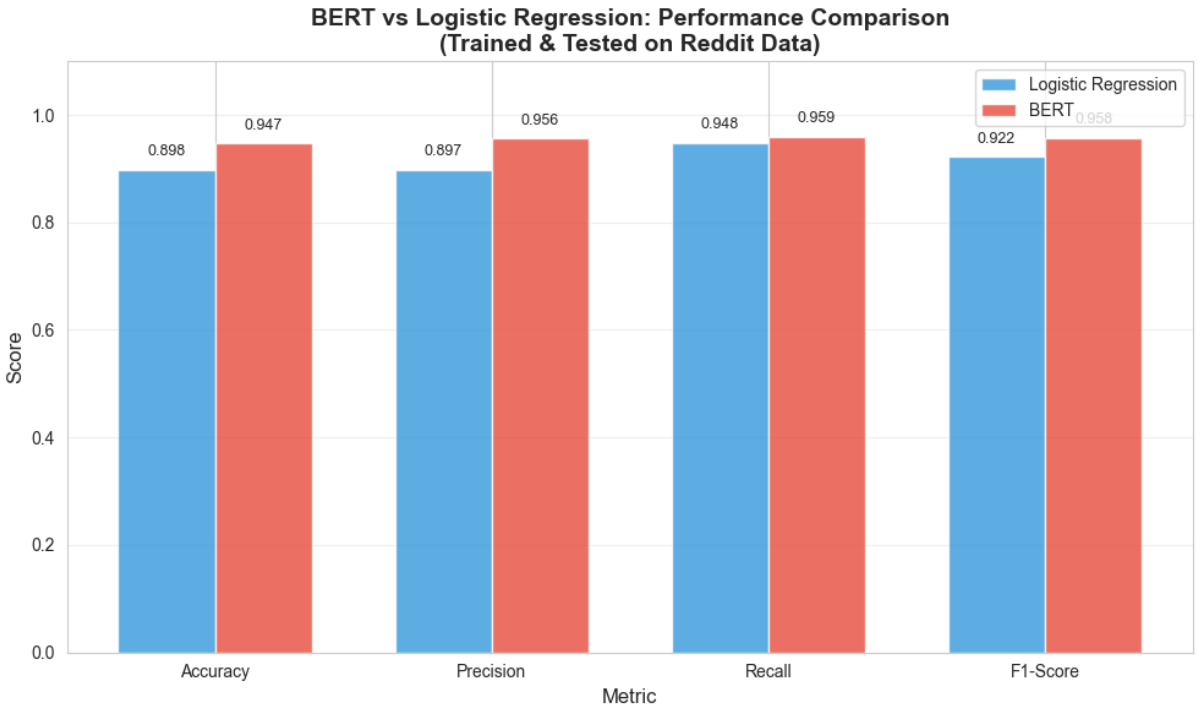

Figure 2: Comparison of model performance

5.2 Research question 2

The second research question tests whether Reddit-trained models can transfer their success to text retrieved from completely other platform, Twitter. The style of the texts can highly differ between the platforms, making the stress-detection task more challenging for the models. The models might have learned platform-specific patterns and some elements such as hashtags are more prominent in Twitter datasets.

The cross-comparison results for Logistic Regression show much worse performance for Logistic regression in classifying stress compared to Reddit test-set. F1-score dropped from 0.92 in Reddit test set to 0.76 on Twitter test set. This suggests that keyword-driven cues learned from reddit generalize somewhat poorly to Twitter. The model likely benefits from the fact that stress discussion shares similar vocabulary across platforms. Although this performance is significantly worse on the in-domain Reddit test-set, the results imply that

Metric	Value
Accuracy	0.74
Precision (no stress)	0.75
Recall (no stress)	0.67
F1-score (no stress)	0.71
Precision (stress)	0.73
Recall (stress)	0.80
F1-score (stress)	0.76

Table 5: Logistic Regression Performance Metrics (Trained on Reddit, Tested on Twitter)

baseline model trained on a certain platform still has some capability to generalize to other platforms.

Metric	Value
Accuracy	0.68
Precision (no stress)	0.61
Recall (no stress)	0.87
F1-score (no stress)	0.71
Precision (stress)	0.81
Recall (stress)	0.51
F1-score (stress)	0.62

Table 6: BERT Performance Metrics (Trained on Reddit, Tested on Twitter)

Table 6 shows results for performance of Reddit-trained BERT on Twitter test set. In this context the transformer-based methods showed much sharper degradation on this new data. For BERT, F1-score dropped from 0.96 in Reddit test set to 0.62 in Twitter. The pattern is notably different from the baseline model. The precision is quite high but recall drops significantly for stress class. These metrics indicate that BERT correctly identifies many stress-positive metrics when it predicts stress but misses almost half of the positive cases. It seems that the fine-tuned BERT model learned stress detection more highly on Reddit specific cues rather than general language-wise. Therefore, its performance on Twitter data drops more than with baseline Logistic regression.

Figure 3 compares the evaluation settings for both transformer-based model and Logistic Regression. The result shows that the in-domain ranking is reversed in the cross-domain performance: baseline Logistic Regression outperforms transformer-based BERT on all metrics except for precision. This might indicate that BERT features associated with Reddit specifically to be related to stress detection although this might not have been the case. Perhaps the imbalance in datapoints in the reddit training set also contributed to this unexpected lower performance. Moreover, the text length metric studied in dataset description section might give insight into this as the BERT model might have learned to associate shorter posts with stress-negativity although this maybe was not the case in the Twitter test set.

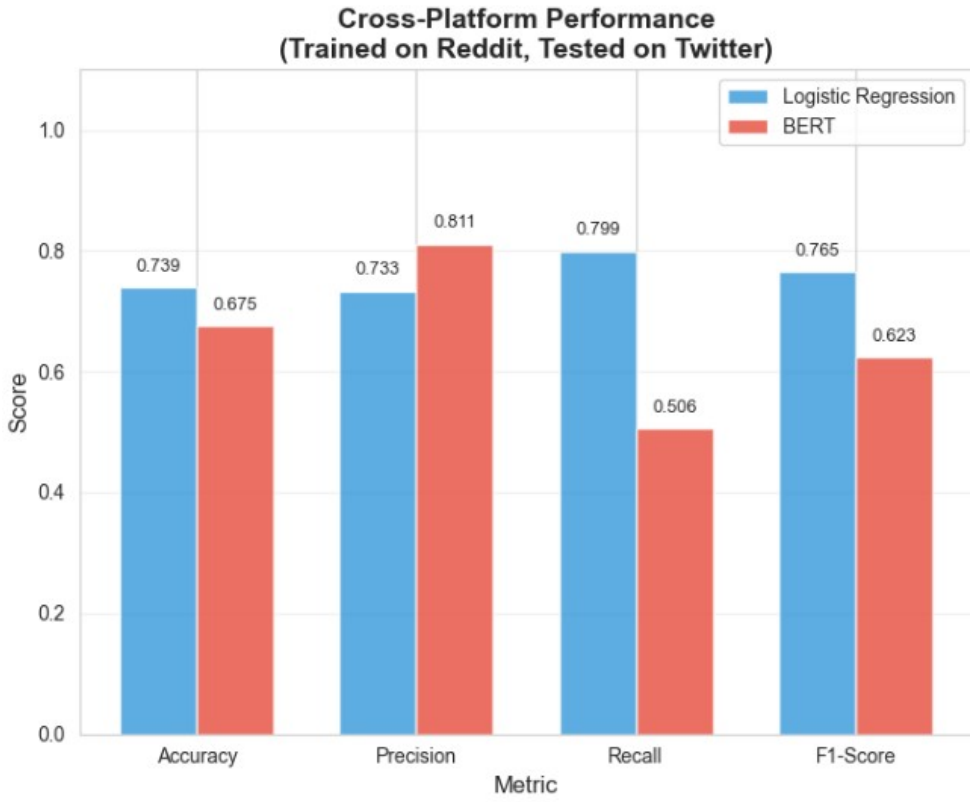

Figure 3: Comparison of model performance

These results indicate that under the configurations used in this project, reddit-trained stress detection models do not generalize to Twitter without significant loss in performance. Notably, transformer-based BERT suffered from more severe performance loss than baseline model in this scenario.

6 Conclusion & Discussion

The results indicate that in-domain performance of models stays relatively high. Both models trained on Reddit dataset performed very well also on Reddit-domain test data while BERT achieved nearly state-of-the-art level results for this binary classification setup. The higher performance of BERT suggests that context embeddings are better than obvious keywords in capturing domain specific stress expressions.

For cross-platform stress detection both models performed poorly. Transformer-based BERT performed worse out of the two models compared. This might indicate that in Twitter, stress signals are more expressed in shorter form such as emojis and hashtags rather than context compared to Reddit's narrative style. To improve performance in future machine-learning projects, the models should learn stress-detection data from multiple platforms contained in

their training data. I believe that this would yield much better results for cross-platform tests even for platforms not contained in the training data initially.

An important conclusion from the results is that model complexity does not guarantee better cross-domain results. The transformer-based BERT took almost 2 hours to train on my (slowish) laptop. Still baseline Logistic Regression with TF-IDF outperformed BERT in this cross-domain experiment. Possible explanation is that Logistic Regression learns more lexical cues on stress detection that generalize better on different platforms whereas fine-tuned transformers may learn platform-specific semantics that do not apply in cross-domain settings. The findings align well with previous work. For in-domain Reddit test set the performance metrics were outstanding but they transformed poorly to Twitter as in previous studies (Harrigian et al., 2020)

There are some limitations to consider for the results of the project. Firstly, the labels from the training data are treated as ground truth for stress-positive and stress-negative textual labeling although the training data is only approximated to represent stress-positive and stress-negative textual posts from corresponding forums. The nature of labeling stress in reality is not a binary process but a spectrum, which the nature of the training data does not address. Secondly, the task is only limited to English text and cannot be generalized to detect stress in other languages. Moreover, there may exist significant differences in how people express stress in different cultures also within the domain of English language. Lastly regarding the model performance, the training data was only limited to a small dataset and broader training data might yield very different results for the model comparison.

Machine learning models for stress detection offer substantial possibilities for monitoring mental health and understanding sentiment related to topics on social media contexts. thus having reliable machine learning models to predict stress levels based on minimal data may offer significant help in mitigating negative mental health effects in the population. Studying this phenomenon in the future and developing useful models stays important as the userbase of social media platforms contninues to grow. Possible directions for future studies include also how the performance of models outside their training domain can be maximized.

References

Harrigian, K., Aguirre, C., & Dredze, M. (2020). Do models of mental health based on social media data generalize? *Findings of the association for computational linguistics: EMNLP 2020*, 3774–3788.

Inamdar, S., Chapekar, R., Gite, S., & Pradhan, B. (2023). Machine learning driven mental stress detection on reddit posts using natural language processing. *Human-Centric Intelligent Systems*, *3*(2), 80–91.

Rastogi, A., Liu, Q., & Cambria, E. (2022). Stress detection from social media articles: New dataset benchmark and analytical study. *2022 International Joint Conference on Neural Networks (IJCNN)*, 1–8. https://doi.org/10.1109/IJCNN55064.2022.9892889