Predicting stress levels based on text input in Reddit posts: A natural language processing perspective on mental health

Brett Balazic (1), Majd Boulos (1), Joshua Onichino (1), Brenda Truong (1), Ann Huang (2)

Dawson College (1), McGill University (2)

Abstract - While COVID-19 was a blatant health problem that hospitalized millions, mental health issues burgeoned as a more latent epidemic, propagating implicitly within our population, especially among adolescents. Social media provide a forum for anonymous conversations that allow those who live with mental health challenges to express themselves and find community. It is against this backdrop that we investigated the ability of social media posts to predict users' potential mental health challenges. We postulated that the anonymity that characterizes most social media posts reduces users' inhibitions, which in turn enhances the ability to detect and understand their lived experiences connected with mental health. From this, we further hypothesized that a careful analysis of user's entries and diction would illuminate any particular health challenges that they might encounter. To test these points, we analyzed a dataset of Reddit posts downloaded from ten different subreddits related to, for example, anxiety, PTSD, and surviving abuse, to study how users discuss mental health experiences. We quantified the different profiles of language expression in different subreddits by measuring the frequency of mental health related keywords. Further, we used supervised machine learning algorithms to predict whether Reddit users have high stress levels based on their posts. These algorithms have been trained on our dataset where each Reddit user has been manually labelled as high or low stress level. We then used unsupervised machine learning algorithms to predict the subreddit into which a random post should figure, based on the diction and vocabulary used. As such, we took the view that these algorithms ultimately should be capable of predicting the mental health disorder to which a Reddit user is most susceptible, based on their posts. Through this study, we achieved four goals; (1) we identified the distribution of stressed individuals across subreddits, (2) we determined the language used by those with varying stress levels across subreddits, (3) we trained a machine learning algorithm to predict the stress level of Reddit users based on language used in their posts (99.1% within training samples, and 71% for generalization), and (4) we trained a machine learning algorithm to predict the subreddit to which any given post belongs (68.9% within training samples, 67.3% for generalization). The applications of such algorithms are numerous and promising. We anticipate that our algorithms will contribute to ongoing research in mental health symptomatology, patient screening, and accessibility of mental wellness resources.

Key words – *mental health, machine learning, Reddit, anxiety, PTSD, healthcare, natural language processing, logistic regression, stress levels*

Introduction

The COVID-19 pandemic resulted in global lockdown and social gathering restrictions that lasted for several months. Many were confined to their homes during this period. This led to an insurgence of research conducted in the field of mental health that has sought to investigate the impact of this spell of isolation, unprecedented in recent generations, on mental wellness. Moreover, the pandemic may have compounded pre-existing, prevalent mental health issues within youth populations. Statistics Canada reports that in 2019, nearly one in five youth aged 15 to 17 report that their mental health was "fair" or "poor" [1]. Restrictions on social interactions imposed to minimize COVID-19 transmissions may have exacerbated adverse outcomes for youth mental wellness.

Moreover, the symptomatology of mental health issues presents numerous challenges when compared to physical ailments. This is because, unlike physical health problems, which result in manifest, concrete symptoms in patients, mental health issues can fester within a population unnoticed and undiagnosed. The Princeton Public Health Review estimates that 450 million people worldwide live with a mental illness, yet most of these people, about 400 million, do not receive the treatment that they need [2].

Further, as is consistent with similar situations of natural disaster and other crises, social media usage has been shown to increase during the COVID-19 pandemic [3]. These platforms allowed many to communicate openly, sharing their unfiltered thoughts as either an identified or anonymous user. These posts offer a window into a user's mental state. This is particularly true of social media outlets such as Reddit, which offers users subspaces, known as subreddits, to discuss issues that are especially relevant to them.

Subreddits such as r/anxiety, r/ptsd, r/relationships, among others, were of particular interest to this study as these communities are populated by individuals who write anonymous posts that relate to their mental state.

It is postulated that the anonymity that characterizes most social media posts reduces users' inhibitions, which in turn enhances the ability to detect and understand their lived experiences connected with mental health. It is further hypothesized that a careful analysis of user's entries and diction would illuminate any health challenges that they might encounter.

While recent studies have sought to investigate mental health symptomatology with machine learning (ML), the relationship between social media and mental health issues has not been explored sufficiently using ML algorithms [4]. This paper investigates whether it is possible to determine the stress level of a social media user based on the diction they use in a post. We explored this problem using ML algorithms given that the relative recency of AI and ML's rise to prominence are such that these have been integrated only minimally to-

date within contemporary mental healthcare practices.

Methods

The Dreaddit Dataset

Dreaddit, a dataset of lengthy social media posts in five categories, each including different ways of expressing stress level through stressful and non-stressful text, with a subset of the data annotated by human annotators, was used in this study [5]. These five categories corresponded to “abuse,” “anxiety,” “financial,” “PTSD,” and “social.” From these larger domains, data were extracted from four different subreddits: r/assistance, r/survivorsofabuse, r/ptsd, and r/relationships (see Table 1).

Text Processing

The text of the posts was first cleaned to prepare them for processing by machine learning. This procedure involved the following steps:

- Lowercasing: All text is converted to lowercase to standardize it and reduce the complexity of the vocabulary.
- Stop word removal: Stop words are common words that do not contribute much to the meaning of the text, such as "the," "and," and "of". Removing these words can reduce noise in the text data and improve the accuracy of the analysis.
- Removing special characters and punctuations: This involved removing all special characters and punctuation marks from the text because these characters do not contribute to the meaning of the text and can interfere with the analysis.
- Encoding: This involved converting the text into a numerical representation that can be processed by algorithms. Word Count Vectorizer is used for the encoding step, which will be introduced below.

Word Count Vectorizer

Word Count Vectorizer is a popular text preprocessing technique used in Natural Language Processing (NLP) to transform text into a numerical format that can be easily analyzed and processed by ML algorithms. In this technique, each Reddit post is represented as a vector, where each dimension represents a specific word in the corpus vocabulary, and the value in each dimension represents the frequency of the corresponding word in the document. This way, ML algorithms can process the data as numerical inputs, making it possible to apply mathematical operations to the text data (see Figure 1).

Logistic Regression

Logistic regression is a type of statistical analysis used to model the probability of a certain event occurring, given a set of independent variables. It is a commonly used algorithm in ML for classification tasks, where the goal is to predict a categorical outcome variable based on one or more input variables. In this study, one aim was to predict the stress level of Reddit users (high-stress versus low-stress) based on the vectorized form of their Reddit posts.

The output of logistic regression is a probability score between 0 and 1, which represents the likelihood of a particular

observation belonging to a specific class. The model makes predictions by comparing the probability score to a threshold value, typically 0.5, and assigns the observation to the class with the highest probability. Logistic regression works by fitting a logistic function to the input variables, which maps the input variables to the output probability score. The model estimates the coefficients of the logistic function, which represent the strength and direction of the relationship between the input variables and the output probability.

In this study, logistic regression is used to predict the probability that a Reddit user has high-stress level (P(high-stress)) based on the user's post. Users are classified as high-stress if they have P(high-stress) greater than 0.5. The coefficient for each unique word represents the strength that the appearance of each word associates with a high stress level post, which facilitates the interpretation of a fitted model (see Figure 2).

Programming Language

All the programming efforts are done in Python within Google Colab. We employed the Word Count Vectorizer and the logistic regression model from the sklearn library for our data analysis.

Results

i) Determining the distribution of highly stressed individuals across different subreddits.

The posts that are labelled are categorized by the subreddit to which they belong and the stress level they represent. The figure below provides the total amount of high-stress and low-stress posts for each subreddit (see Figure 3). The graph shows that r/ptsd and r/relationships have the most labelled posts and that r/stress and r/food pantry provide the least.

ii) Language used by stressed individuals

Determining the frequency of words was done by creating word clouds that represent the commonly used words for certain subreddits and stress levels. Most of the words that appear frequently are those that are used in daily language like "time," "many," and "make." Apart from these more common words, language that is associated with mental health challenges tended to appear at high frequencies in high-stress word clouds in comparison to low-stress clouds. These words include "help," "therapist," "anxiety," "attack," "trauma," "abuse," "crying," "nightmare," "medication," and "doctor". The figures below illustrate the progression of the techniques that were employed to determine the language patterns used by those with varying stress levels (see Figures 4 and 5).

In the first word cloud below, the most frequently used words in the PTSD subreddit are visualized. However, it was found that most of the words visualized below are common words used in daily language and are thus not informative about the stress level of the Reddit user. Some of these words include "time," "need," and "know" (see Figure 4).

To tackle this problem and visualize only the words that are highly indicative of the users' stress level, an additional sentiment analysis

was performed before drawing the word cloud. Sentiment analysis involves using machine learning algorithms to analyze each word in the Reddit posts and identify the sentiment expressed. Most of the common words like "time," "need," and "know" have a sentiment score of zero, meaning that they are neutral expressions. It is only the subset of words with a negative sentiment score that are included with this technique, meaning that the word clouds produced contain negative sentiments, for subsequent analysis. Their frequencies were counted over posts in the PTSD subreddit and visualized in the second word cloud below (see Figure 5).

iii) Training a machine learning algorithm to predict the stress level of Reddit users based on language used in their posts.

Logistic regression was used to create a curve of best fit around the training data to predict the stress level of other posts. The model was given the vectorized text inputs and was allowed to make predictions, where posts that had a P(high-stress) value greater than 0.5 were defined to be high stress level posts. Then, the parameters were optimized to improve the quality of the predictions.

The confusion matrix of the predicted stress level of a Reddit post is visualized below (see Figure 6).

The figure shows that, of the posts evaluated, 293 (67.4%) of the high stress level labels were correctly predicted and 334 (76.3%) of the low-stress level labels were correctly predicted. Overall, the algorithm was able to predict the stress level of a post with a 71% accuracy. In comparison, it can predict the stress level of the training data with an accuracy of 99.1%.

iv) Training a machine learning algorithm to predict the subreddit to which any given post belongs.

The results obtained in this part of the study can potentially demonstrate if it is possible to predict the mental health problem to which any individual is most susceptible based on their Reddit posts. A multi-class logistic regression is used to predict the subreddit of a post. The training accuracy and the testing accuracy were obtained, which are 68.9% and 67.3% respectively.

Discussion and Conclusion

Our results suggest that the logistic regression model that we employed attained the two objectives central to this study, *iii)* determining the stress level of users based on their posts, and *iv)* identifying the subreddit to which a given post belongs, within a reasonable margin of accuracy. Further, our results illustrate the recurrent language used by individuals with high stress levels. Among the common words frequently used were "help," "time," and "need" while "attack," "bad," and "hard" were among the more frequently used words with a negative sentiment score. We suggest that future studies use these language patterns as the precipice for investigating other avenues of study.

While we were able to achieve strong results for each of the questions we posed at the outset of our study, our findings are limited by three key factors. First, our logistic regression model was able to predict the

stress level of the Reddit users based on their posts for only 71% of the posts under generalization. While this result is strong, it only differs from the result that would be due to random chance by 20 percentage points. Second, our model was able to predict the subreddit to which a given post belongs for only 67.3% of the posts under generalization. Third, the sole data source from which results were obtained was Reddit. Ideally, a researcher would have drawn from multiple social media platforms to avoid potential bias or patterns that might be present among users of a single platform.

We posit that these limitations do not detract from the strength of this study, but instead provide a window into future avenues of investigation. For example, the implementation of more complex ML algorithms or language models can be employed using a method similar to that which was employed in this study to yield results with higher accuracy. Further, language models that are more refined will be able to draw upon the language patterns identified in this study to determine, with greater precision, stress levels from text-based input. Further, we suggest that future research be conducted using a similar methodology but with posts from different languages. Each language provides those who use it with a unique and distinct set of tools to express themselves and, as such, it would be fascinating to see how ML can be used to predict mental health challenges across different tongues.

A principal aim of this project was to contribute to the democratization of healthcare. As these technologies develop over time, we can imagine the broad, bilateral uses and benefits of ML algorithms in the context of adolescent mental health. In particular, a user who contemplates that they might be experiencing mental illness could use these algorithms to relay a post they shared on social media. The algorithm would in turn yield preliminary diagnoses and support the process of the user's search for targeted professional care. Alternatively, we envision the development of a chatbot with which patients could have conversations, similar in some respects to ChatGPT, which could in turn make preliminary diagnoses, determine which patients require the most urgent care, and suggest some immediate steps that could be taken to lessen the burden of the mental challenges they are experiencing. However, we appreciate the complexity of this issue given the ethical considerations and the fact that the accuracy of predictions made by ML algorithms must be improved considerably before being integrated into clinical practice.

At the same time, such algorithms hold the promise of facilitating mental health professionals' work with and treatment of such patients. The COVID-19 pandemic exposed the frailties of our healthcare system, and we feel that a lack of resources and time is central to this issue. We anticipate that the integration of a fully developed ML algorithm will not only ameliorate patient health outcomes but also allow physicians to see and treat more patients. In this way, our research can play a role in the development of innovative, practical, and accessible tools

that will advance individual and population mental health outcomes.

References

[1] Government of Canada, S. C. (2022, May 11). *Canadian Health Survey on Children and Youth, 2019*. The Daily, 2019. Retrieved April 23, 2023, from https://www150.statcan.gc.ca/n1/daily-quotidien/200723/dq200723a-eng.htm

[2] PPHR. (2017, April 30). *Untreated Mental Illnesses: The Causes and Effects – Princeton Public Health Review*. Princeton University. Retrieved April 23, 2023, from https://pphr.princeton.edu/2017/04/30/untreated-mental-illnesses-the-causes-and-effects/

[3] Venegas-Vera, A. V., Colbert, G. B., & Lerma, E. V. (2020). *Positive and negative impact of social media in the COVID-19 era*. Reviews in cardiovascular medicine, 21(4), 561–564. https://doi.org/10.31083/j.rcm.2020.04.195

[4] Qasrawi, R., Vicuna Polo, S. P., Abu Al-Halawa, D., Hallaq, S., & Abdeen, Z. (2022). *Assessment and Prediction of Depression and Anxiety Risk Factors in Schoolchildren: Machine Learning Techniques Performance Analysis*. JMIR formative research, 6(8), e32736. https://doi.org/10.2196/32736

[5] Turcan, E., & McKeown, K. (2019, October 31). *Dreaddit: A Reddit Dataset for Stress Analysis in Social Media*. arXiv.org. Retrieved April 23, 2023, from https://arxiv.org/abs/1911.00133

[6] Researcher (n.d.). *Linear regression vs. logistic regression: Understanding 13 key differences*. Spiceworks. Retrieved April 23, 2023, from https://www.spiceworks.com/tech/artificial-intelligence/articles/linear-regression-vs-logistic-regression/

[7] Researcher (n.d.). *Countvectorizer in python*. Educative. Retrieved April 23, 2023, from https://www.educative.io/answers/countvectorizer-in-python

Appendix

Domain	Subreddit Name	Labeled posts
Abuse	r/domesticviolence	388
	r/survivorsofabuse	315
Anxiety	r/anxiety	650
	r/stress	78
Financial	r/assistance	355
	r/almosthomeless	99
	r/homeless	220
	r/foodpantry	43
PTSD	r/ptsd	711
Social	r/relationships	694

Table 1. Labelled posts for each subreddit from the Dreaddit domains that are analyzed in this study. There exist five separate domains, each of which consisting of at least one subreddit for a total of 3553 labelled posts.

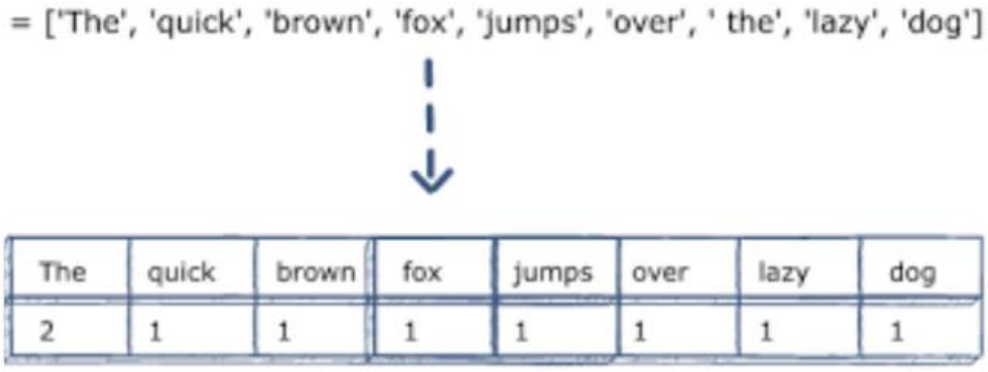

Figure 1. Example of the way in which Word Count Vectorizer interprets data. This image illustrates an example of how Word Count Vectorizer converts text data into numerical vectors, where each element represents the count of a word.

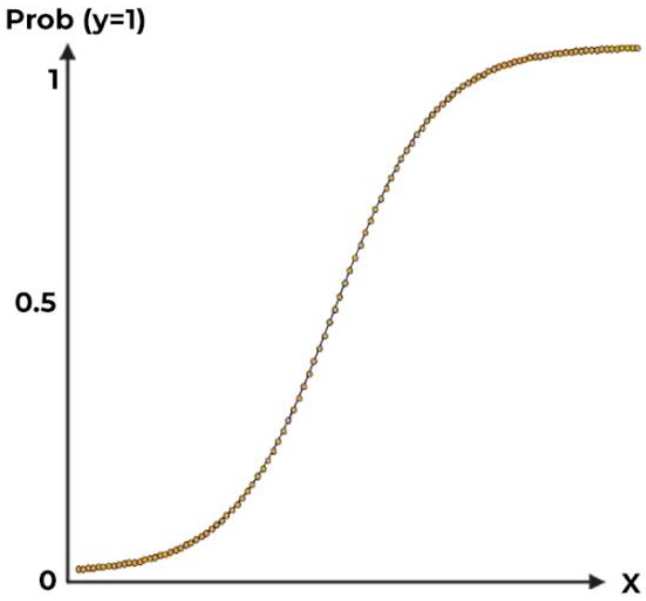

Figure 2. Example of the way in which logistic regression fits data. Logistic regression uses input variables to calculate probabilities of binary outcomes and classify data accordingly.

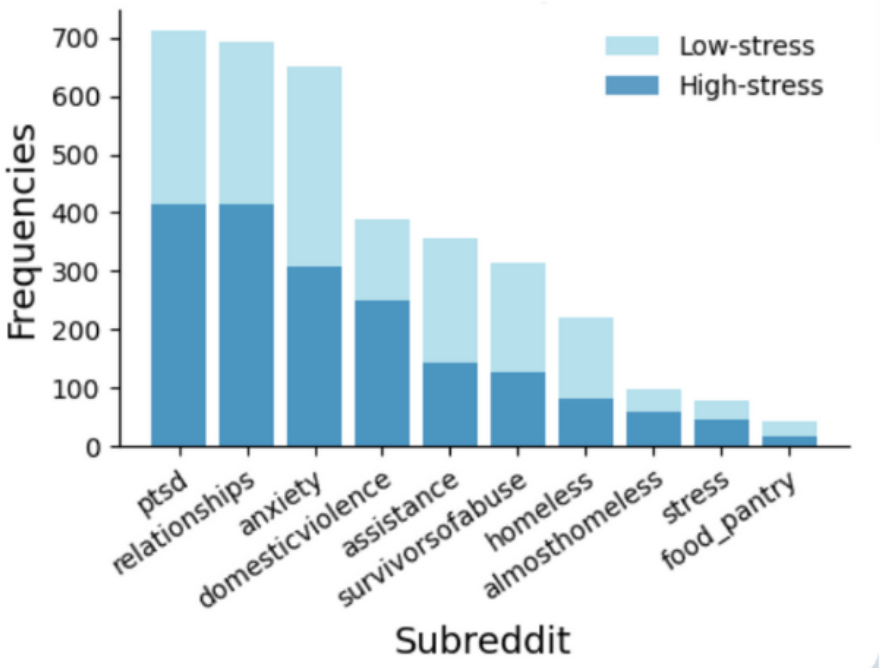

Figure 3. The distribution of stressed individuals across subreddits. This bar graph illustrates the frequency of labelled posts as a function of the subreddit to which they belong. Each bar is composed of two smaller dark and light blue bars indicating a high- or low-stress post, respectively. This sorting method allows for a visual representation of the composition of each subreddit.

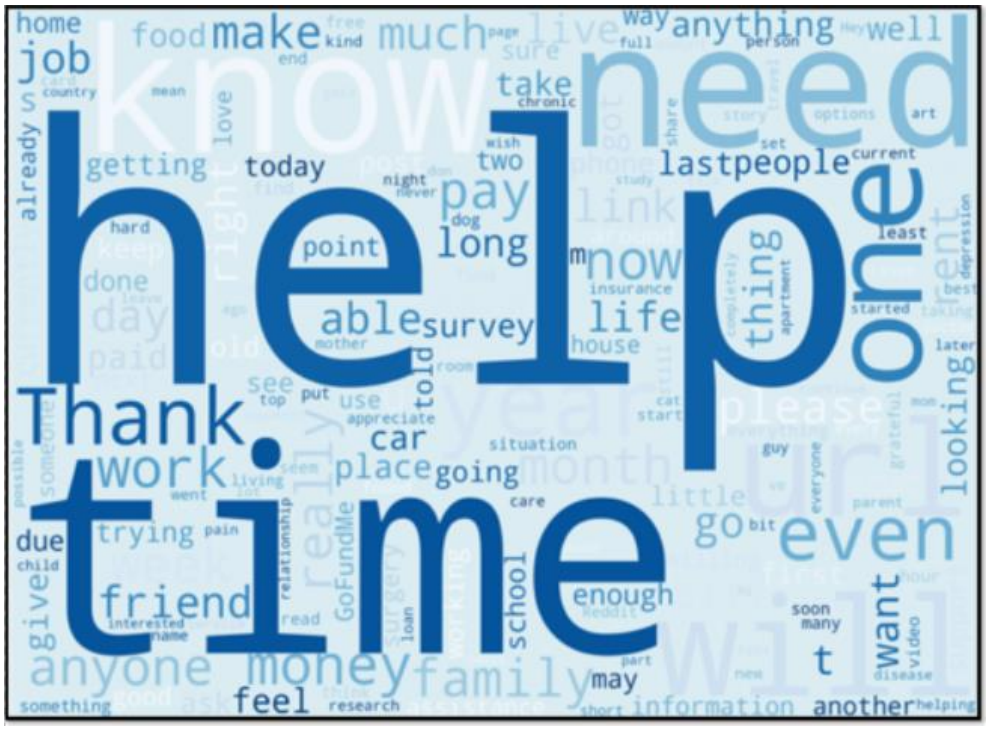

Figure 4. Word cloud of the most frequently used words in posts from the PTSD subreddit.

Figure 5. Word cloud of words with a negative sentiment score in posts from the PTSD subreddit.

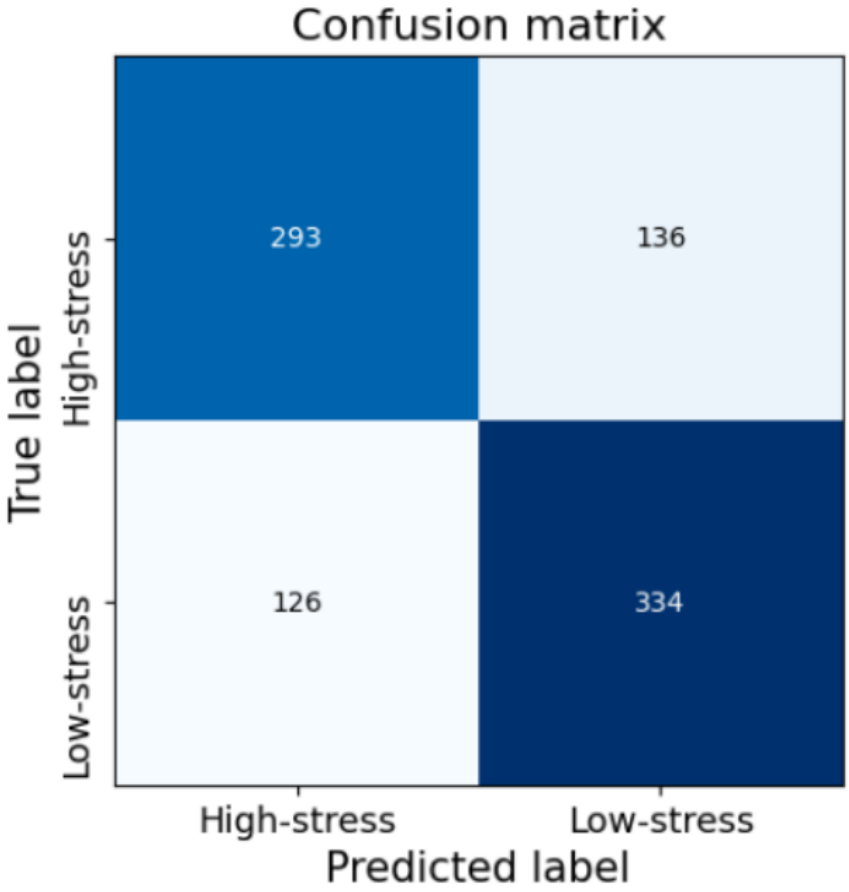

Figure 6. The logistic regression confusion matrix. This matrix shows the number of data points that correspond to each combination of true label and predicted label. At the centre of each matrix is the number of posts sorted by the algorithm.