PeerJ Computer Science

Intent aware data augmentation by leveraging generative AI for stress detection in social media texts

Minhah Saleem and Jihie Kim

Department of Computer and AI, Dongguk University, Seoul, Republic of South Korea

ABSTRACT

Stress is a major issue in modern society. Researchers focus on identifying stress in individuals, linking language with mental health, and often utilizing social media posts. However, stress classification systems encounter data scarcity issues, necessitating data augmentation. Approaches like Back-Translation (BT), Easy Data Augmentation (EDA), and An Easier Data Augmentation (AEDA) are common. But, recent studies show the potential of generative AI, notably ChatGPT. This article centers on stress identification using the DREADDIT dataset and A Robustly Optimized BERT Pre-training Approach (RoBERTa) transformer, emphasizing the use of generative AI for augmentation. We propose two ChatGPT prompting techniques: same-intent and opposite-intent 1-shot intent-aware data augmentation. Same-intent prompts yield posts with similar topics and sentiments, while opposite-intent prompts produce posts with contrasting sentiments. Results show a 2% and 3% performance increase for opposing and same sentiments, respectively. This study pioneers intent-based data augmentation for stress detection and explores advanced mental health text classification methods with generative AI. It concludes that data augmentation has limited benefits and highlights the importance of diverse Reddit data and further research in this field.

Subjects Artificial Intelligence, Computational Linguistics, Natural Language and Speech, Network Science and Online Social Networks, Sentiment Analysis
Keywords Stress detection, Mental health, Data augmentation, Text classification, Generative AI, Sentiment analysis, Pre-trained language models, Natural language understanding

Submitted 27 February 2024
Accepted 5 June 2024
Published 8 July 2024

Corresponding authors
Minhah Saleem,
minhahsaleem1@gmail.com
Jihie Kim, jihie.kim@dgu.edu

Academic editor
Syed Sajid Ullah

Additional Information and Declarations can be found on page 17

DOI 10.7717/peerj-cs.2156

INTRODUCTION

In modern society, the need for effective mental health and stress detection mechanisms has become increasingly imperative. The pervasive impact of stress, touching individuals of all ages and backgrounds, underscores the urgency for accurate and accessible means of identification.

Early studies in psychology linked language with mental states (*Ramirez-Esparza et al., 2021*; *Rude, Gortner & Pennebaker, 2004*). Today, platforms like Reddit offer abundant linguistic data, facilitating open discussions on daily experiences and challenges, especially among the youth. This has fueled extensive natural language processing (NLP) research to identify stress patterns and trends.

Various approaches within the field encompass a spectrum of models, ranging from classical supervised techniques to deep models (*Turcan & McKeown, 2019*; *Ansari, Garg*

How to cite this article Saleem M, Kim J. 2024. Intent aware data augmentation by leveraging generative AI for stress detection in social media texts. *PeerJ Comput. Sci.* **10**:e2156 http://doi.org/10.7717/peerj-cs.2156

Table 1 **Example text posts that are wrongly classified by RoBERTa.**

Reddit post	Label	Prediction
Almost decided to live in my car, live with a crack head, travel the country (aka: begin my homeless life, cause I really had no money and if I did it would've gone to beer). I was losing my mind. After being heavily suicidal for a week I decided I can't live like this anymore - but I don't want to die right now. I planted a thought in my head. 'if you ever want to overcome this, you need to begin to change'.	Not stressed	Stressed
It's just us two and it's, really intense. She hugs me, tells me how much she's missed me. Reminisces about our relationship. Tells me how I broke her heart. She tells me about lads she's been with since and it felt like she was comparing them all to me and gets really emotional.	Stressed	Not stressed

& Saxena, 2021). Additionally, researchers have explored pre-trained large language models (*Liu et al., 2019*; *Devlin et al., 2019*) and their domain-specific variants (*Ji et al., 2022*; *Naseem et al., 2022*), as well as the application of zero and few-shot prompting for stress classification, leveraging models like GPT, Alpaca, and Flan (*Xu et al., 2023*). While certain domain-specific models demonstrate superior performance, it is noteworthy that A Robustly Optimized BERT Pretraining Approach (RoBERTa) emerges as the top-performing general model (*Kumar, Trueman & Cambria, 2022*), providing a robust baseline for further investigations.

Despite extensive research, only two publicly available datasets exist: the SAD dataset (*Mauriello et al., 2021*) with SMS-like sentences and the DREADDIT dataset (*Turcan & McKeown, 2019*), consisting of nearly 3,000 labeled Reddit posts spanning 10 subreddits. DREADDIT focuses on stress expressions with a negative attitude, not objective stress, making the classification more challenging. RoBERTa (*Liu et al., 2019*) struggles with classifying posts having subjective stress. Table 1 provides the misclassified examples and highlights the discrepancy between situations and users' emotional responses. For instance, in the first post, the user discusses financial issues and even expresses suicidal thoughts but is labeled as "Not stressed" because they also convey an intent to improve their condition. Conversely, the second post is labeled as "Stressed" despite the objectively non-stressful situation, highlighting the user's negative thinking leading to stress.

To enhance comparability, we conduct zero-shot text classification using ChatGPT. Despite outperforming existing literature for stress classification with zero-shot ChatGPT (*Xu et al., 2023*; *Yang et al., 2023*; *Lamichhane, 2023*), RoBERTa remains superior.

As the dataset heavily relies on user intent, making traditional machine learning techniques that leverage features or other methods incorporating linguistic features or external knowledge less effective, given their high dependence on stressors. Furthermore, due to the limited size of the dataset, data augmentation becomes a necessity, prompting extensive exploration of augmentation techniques in the field.

One such technique is token-level augmentation, wherein syntax and semantic meaning are preserved through the introduction of local modifications. Techniques like Easy Data Augmentation (EDA) (*Wei & Zou, 2019*) and An Easier Data Augmentation(AEDA)

(*Karimi, Rossi & Prati, 2021*), which apply random operations such as random token level insertion, are widely used in text classification and serve as benchmarks in our research. However, methods involving data noising prove unsuitable for social media data due to its inherent noise, and those introducing only local modifications are rendered ineffective, contributing to low corpus-level variability (*Anaby-Tavor et al., 2020*).

Various other techniques, including conditional (*Wu et al., 2019*; *Kumar, Choudhary & Cho, 2020*), unconditional (*Goodfellow et al., 2014*; *Kingma & Welling, 2013*), and graph-based methods (*Chen, Ji & Evans, 2020*), have been explored. Unconditional augmentation is context-agnostic, making it unsuitable for our problem. Back translation, a popular sentence-level technique, is employed for comparison, recognizing its limitations for the paragraph-based nature of Reddit posts.

These challenges highlight the necessity for post-level augmentation techniques that understand context, intent, and attitude. Generative AI, particularly ChatGPT, is chosen in existing literature for its robust capabilities, contextual understanding, and pre-training. Two implemented ideas involve simple rephrasing (*Dai et al., 2023*) and generating a mix of real and synthetic yet realistic samples (*Yoo et al., 2021*).

We propose two data augmentation techniques using 1-shot ChatGPT prompting: same intent and opposite intent 1-shot intent-aware data augmentation. 1-shot learning involves using a single labeled example per new class, enabling the model to predict based on this sole example. In the same intent augmentation, ChatGPT generates a new post with the same topic and sentiment as a given post. Conversely, opposite intent augmentation prompts a post with the same topic but a different sentiment. Table 2 illustrates the same and opposite intent augmentation for "Stressed" and "Not Stressed" posts. For instance, consider the first stressed user post discussing challenging relationships with parents. In same intent augmentation, the emotional tone is mirrored, discussing unsupportive religious parents. While, for opposite intent augmentation, the user describes supportive attitude of parents.

Same intent augmentation aims to enhance sentiment learning through varied narrations, while opposite intent augmentation provides diverse perspectives on the same situation, aiding intent discernment. Our technique exhibits a significant 2% improvement with opposite intent and 3% with same intent augmentation compared to other methods. The success of 1-shot augmentation, maintaining consistent sentiment, proves promising in capturing contextual nuances and understanding user emotional responses. Additionally, we explored the impact of increasing the augmented data size and discovered that beyond a certain limit, the model's performance plateaus, indicating the need for more real-time diverse data.

To summarise, in this article, we introduce a stress prediction approach using Reddit posts and the RoBERTa transformer (*Liu et al., 2019*). Employing established data augmentation techniques, including Back-Translation (BT), Easy Data Augmentation (EDA) (*Wei & Zou, 2019*), and An Easier Data Augmentation(AEDA) (*Karimi, Rossi & Prati, 2021*), alongside a novel 1-shot intent-aware technique using GPT-3.5 (*Brown et al., 2020*), we achieve a 3% F1 score enhancement.

Our work is at the forefront of utilizing intent-based data augmentation to detect stress. The study offers a comprehensive examination of mental health text classification

Table 2 **Augmented text posts.**

Reddit post	Label	Same intent Aug.	Opposite intent Aug.
Did I mention my parents are religious? I don't know if this is normal for religious people to treat. Whenever I tell them I'm terrified of being homeless they tell me I'm a "acting like a baby" and "get over it" my parents parents did not treat them this way. They're basically mad because they(I guess 30 or 40 years ago were different when they were my age? Because they said they both lived on their own at 17 and that they find it creepy I'm 16+ and they find it creepy being around me).	Stressed	My folks are religious, did I mention that? Wondering if this treatment is typical. Every time I mention my fear of homelessness, they brush it off, calling me a "baby" and saying to "get over it." Their parents didn't treat them like this. They're hung up on their own past, saying they moved out at 17 and now find it "creepy" being around me at 16+. Stressful, right?	Feeling grateful for my supportive religious parents. They've always been understanding of my fears and never dismissive. Unlike some stories I've heard, they don't belittle my concerns or compare our experiences at my age. Any tips on how to show appreciation for such understanding parents?
This was a 'nice thought', and even though I had been told my mother to never leave a drink at the bar, it didn't cross my mind that there might be something put in this drink or that I was being pushed to be more intoxicated. After I started sipping on that drink, I felt all the alcohol hit me and felt drunk. He said I've ordered an uber for us, which I thought I was being dropped back at home; I don't actually remember the ride home in the Uber. We pulled up at his apartment complex, and he said come inside I know you're hungry I have pizza. I thought eating pizza would be a good idea after the alcohol, so I followed him inside.	Not stressed	Wow, what a wild night! It started off with a simple drink at the bar, and next thing I know, I'm enjoying pizza at this guy's place. Sure, I probably should've been more cautious, but everything turned out fine. Lesson learned though: always stay aware of your surroundings, even when things seem harmless.	This is so twisted! My gut told me to bail the moment I saw that drink, but I stupidly ignored it. Next thing I know, I'm practically blacked out in some stranger's Uber. And then, surprise, surprise, we're at his place, and suddenly I'm supposed to be grateful for pizza? Yeah, right. That's how predators work, luring you in with fake kindness. I'm lucky nothing worse happened, but seriously, I need to listen to my instincts next time.

techniques, evaluates the effects of preprocessing, and delves into the limitations of generative AI. Contributions include:

- A comprehensive analysis of stress classification,
- The proposal of 1-shot intent-aware data augmentation at post level for social media texts using generative AI
- Evaluation of two data augmentation techniques using 1-shot ChatGPT prompting: same intent and opposite intent 1-shot intent-aware data augmentation.

Despite observed limitations, our research advocates for continued innovation, emphasizing the importance of diverse, real-time data and recognizing the evolving nature of language on social media. Future exploration is needed in stress detection methodologies, model architectures, and modeling techniques.

RELATED WORKS

Background and motivation

The well-being of individuals and their ability to function effectively in life depends on their mental health, which encompasses emotional, psychological, and social aspects (*American Psychological Association, 2019*). Stress, a natural reaction to challenges, becomes problematic when it exceeds a certain level, affecting our physical, emotional, and behavioral responses (*American Psychological Association, 2019*). According to the Stress in America survey (2022), around 27% of adults experience debilitating stress (*American Psychological Association, 2022*), while 74% encounter overwhelming stress situations (*Mental Health Foundation, 2018*). Daily stress overload is a significant burden for young adults aged 18 to 34 (*American Psychological Association, 2022*). In their 2018 report, the Mental Health Foundation emphasized the link between stress, depression (51%), and heightened anxiety (61%), while also highlighting alarming rates of self-harm (16%) and suicidal thoughts (32%) among individuals dealing with stress (*Mental Health Foundation, 2018*). Generation Z's elevated stress levels signal a concerning trend (*American Psychological Association, 2020*).

The urgency to comprehend and deal with stress is emphasized by its profound effects on individual and societal well-being. Understanding and measuring stress involves a range of factors, like self-report surveys and observational studies (*Garg, 2023*; *Ramirez-Esparza et al., 2021*; *Rude, Gortner & Pennebaker, 2004*). In today's digital age, platforms such as Reddit are crucial, providing spaces for self-expression and the sharing of daily experiences (*Statista, 2022*). Social media enables adolescents to form communities, maintain friendships, and access social support (*Mental Health Foundation, 2018*; *U.S. Surgeon General Advisory, 2023*).

By employing NLP techniques, social media platforms can detect stress patterns in user-generated posts (*Statista, 2022*; *U.S. Surgeon General Advisory, 2023*).

Text classification for stress analysis

Numerous studies have investigated stress classification through NLP and machine learning, employing various models and approaches. Researchers have utilized classical supervised techniques (*e.g.*, support vector machines, logistic regression, naive Bayes, multi-layer perceptrons, decision trees) as well as deep models like convolutional neural networks (CNNs) and gated recurrent units (GRUs) (*Turcan & McKeown, 2019*; *Inamdar et al., 2023*; *Febriansyah et al., 2023*; *Selvadass, Bruntha & Priyadharsini, 2022*).

Another set of approaches involves fine-tuning pre-trained, large language models, such as Bidirectional Encoder Representations from Transformers (BERT) (*Devlin et al., 2019*), RoBERTA (*Liu et al., 2019*), and their variants like Mental BERT (*Ji et al., 2022*), Mental RoBERTA (*Ji et al., 2022*) and PHS-BERT (*Naseem et al., 2022*). Additionally, integration of linguistic features has shown promise for stress detection, with notable results (*Ilias, Mouzakitis & Askounis, 2023*; *Wang et al., 2023*). For the DREADDIT dataset, *Kumar, Trueman & Cambria (2022)* achieved an F1 score of 83.90 using RoBERTa along with linguistic inquiry and word count (LIWC) features. *Yang, Zhang & Ananiadou (2022)* introduced KC-NET, incorporating context-aware post-encoding, knowledge-aware dot product attention, and supervised contrastive learning, achieving an impressive F1 score of 83.50 on the same dataset.

Comparative evaluation of Alpaca, Alpaca-LoRA, FLAN-T5, GPT-3.5, and GPT-4 in zero-shot prompting, few-shot prompting, and instruction fine-tuning revealed that GPT provides the highest accuracy in zero-shot prompting, while FLAN-T5 excels in few-shot prompting (*Xu et al., 2023*). *Yang et al. (2023)* comprehensively compared language model

Table 3 ChatGPT prompts used in literature for stress classification.

Technique	Zero-shot prompt
Xu et al. (2023)	TextData + "As a psychologist, read the post on social media and answer the question." + "Is the poster [stressed]?" + "Only return yes or no"
Yang et al. (2023)	Post: [Post] Question: Does the poster suffer from stress?
Lamichhane (2023)	"Which of the classes can the following post be assigned to? Only reply with answers from one of the classes: stressed, non-stressed."

(LLM) performances, highlighting the effectiveness of ChatGPT with expert-written few-shot examples. The best zero-shot classification performance using ChatGPT is an F1 score of 73% (*Lamichhane, 2023*). Some of the prompts used in literature for zero-shot classification using ChatGPT can be seen in Table 3. The varying performance highlights the dependence of zero-shot classification on query prompts.

In summary, the literature reveals a diverse array of approaches ranging from classical supervised techniques to advanced pre-trained language models, each demonstrating varying degrees of success in stress classification tasks. Despite the superior performance of domain-specific models, RoBERTa turns out to be the best-performing general text classification model (*Kumar, Trueman & Cambria, 2022*).

Data augmentation

Widely used token-level augmentation techniques involve preserving syntax and semantic meaning through well-designed local modifications. The most popular technique is EDA (*Wei & Zou, 2019*) which includes random operations such as deletion, insertion, replacement, and swapping. An AEDA only performs random insertion of punctuation marks and promises better performance (*Karimi, Rossi & Prati, 2021*). Language models (LMs), including Conditional BERT (*Wu et al., 2019*) and paraphrasing techniques like round-trip or back translation (*Edunov et al., 2018*; *Xie et al., 2020*; *Corbeil & Ghadivel, 2020*), play a crucial role in data augmentation for NLP tasks. Some research explores end-to-end models for paraphrasing with additional syntactic information and latent variables (*Kumar et al., 2019*; *Prakash et al., 2016*; *Iyyer et al., 2018*; *Gupta et al., 2018*).

Unconditional data augmentation models like generative adversarial networks (GANs) (*Goodfellow et al., 2014*) and variational autoencoders (VAEs) (*Kingma & Welling, 2013*), along with graph-based methods (*Chen, Ji & Evans, 2020*), are also explored in the literature. Conditional generation methods, utilizing pre-trained language models like GPT-2 (*Anaby-Tavor et al., 2020*) and T5 (*Lee et al., 2021*), generate extra text conditioned on labels for data augmentation.

Recent efforts leverage Generative AI, such as ChatGPT, proposing novel techniques (*Yoo et al., 2021*; *Dai et al., 2023*). *Yoo et al. (2021)* suggests a technique using large-scale language models to generate realistic text samples, incorporating soft-labels for knowledge distillation. AugGPT (*Dai et al., 2023*) rephrases each sentence in training samples into multiple conceptually similar but semantically different samples for downstream model training.

Challenges and opportunities

The Dreaddit dataset leverages Reddit posts for stress classification, but its limited size challenges language model performance, therefore, data augmentation is crucial. Unconditional data augmentation techniques (*Goodfellow et al., 2014*; *Kingma & Welling, 2013*) generate random texts but lack contextual understanding essential for stress detection. Token-level augmentations benefit simpler tasks more, hindering performance in complex ones (*Chen et al., 2023*). EDA (*Wei & Zou, 2019*) improves with machine learning techniques (*Ansari, Garg & Saxena, 2021*), but limited benefits are observed for BERT and RoBERTa (*Feng et al., 2021*). In mental health classification on Reddit, RoBERTa and EDA yield minimal enhancement, attributed to EDA's efficacy in smaller datasets (*Murarka, Radhakrishnan & Ravichandran, 2021*). Data noising (*Wei & Zou, 2019*; *Karimi, Rossi & Prati, 2021*) is unsuitable for social media noise. Sentence-based augmentation is also ineffective for Reddit due to large paragraphs (*Ansari, Garg & Saxena, 2021*). Local changes yield structurally similar sentences, making token or sentence-level techniques ineffective for RoBERTa in stress classification (*Anaby-Tavor et al., 2020*). Despite challenges, generative AI shows promise for superior data augmentation (*Yoo et al., 2021*; *Dai et al., 2023*), particularly with RoBERTa. Further research in this direction can overcome mental health classification limitations.

METHODOLOGY

In this section, we detail the methodology employed in our research, focusing on the use of two generative AI techniques for data augmentation using 1-shot learning to enhance stress detection in Reddit posts.

Model

Our baseline stress detection model, RoBERTa (*Liu et al., 2019*), optimizes BERT pretraining with enhanced performance, denoising autoencoding, and robust hyperparameter tuning. Chosen for its proven efficacy on mental health datasets, RoBERTa serves as a general yet effective baseline, substantiated by literature and memory considerations.

Dataset

Major mental health studies (*Garg, 2023*; *Ansari, Garg & Saxena, 2021*) predominantly utilize Reddit data. Benchmark datasets for stress detection include the Stress Annotated Dataset (SAD) (*Mauriello et al., 2021*) and DREADDIT (*Turcan & McKeown, 2019*).

SAD (*Mauriello et al., 2021*) comprises 6,850 SMS-like sentences categorized into nine stressor categories, offering daily stressor classifications and a collection of sentences derived from stress management literature, live chatbot conversations, crowdsourcing, and web scraping.

DREADDIT (*Turcan & McKeown, 2019*) presents a text corpus of 190,000 posts from five Reddit community categories, with 3.5K labeled segments for stress identification extracted from 3,000 posts. Covering ten subreddits related to abuse, social issues, anxiety, PTSD, and financial matters from January 2017 to November 2018, the dataset emphasizes expressions of stress with a negative attitude. This aligns with our stress classification research, emphasizing understanding user intent. Annotators focused on identifying both stress and a negative attitude in posts. The training set comprises 2,838 data points (51.6% labeled as stressful), and the test set has 715 data points (52.4% labeled as stressful).

Given the relatively small size of the dataset, data augmentation techniques are deemed necessary to enhance its robustness and effectiveness in stress classification.

Baseline augmentation techniques

Back translation

Back Translation, a widely adopted data augmentation technique, is implemented for comparison. Beyond traditional paraphrasing, our approach introduces multi-language translation: English to Korean to Arabic to French and back to English, generating diverse expressions.

EDA

EDA (*Wei & Zou, 2019*) enriches the dataset by random deletion, synonym replacement, insertion, and word shuffling. Widely used in stress detection literature, EDA serves as a benchmark for comparison.

AEDA

AEDA (*Karimi, Rossi & Prati, 2021*), a faster technique performing random insertion of punctuation marks, claims better and stable performance. Included for efficiency and potential improvement in stress detection tasks, it complements our baseline.

These augmentation techniques, chosen based on prevalence, relevance to stress detection, and potential impact on model robustness, are systematically compared against each other and the baseline RoBERTa model to assess their performance in Reddit posts.

Proposed augmentation techniques

In this section, we introduce a novel approach to data augmentation, termed 1-shot intent-aware data augmentation, leveraging generative AI models, specifically ChatGPT (GPT 3.5). Unlike traditional token-level augmentation, our methodology delves deeper into understanding user posts at a semantic level. An overview of our proposed method can be seen in Fig. 1.

1-shot same intent data augmentation

This strategy involves instructing the generative AI (ChatGPT) to generate a post similar to a provided example post, maintaining both the topic and emotion. For example, if the original post is labeled as "stressed," the generated post should also evoke a "stressed" sentiment. We conducted a series of experiments with different prompts and manually analyzed the generated posts. Our meticulous experimentation led us to adopt the following prompt:

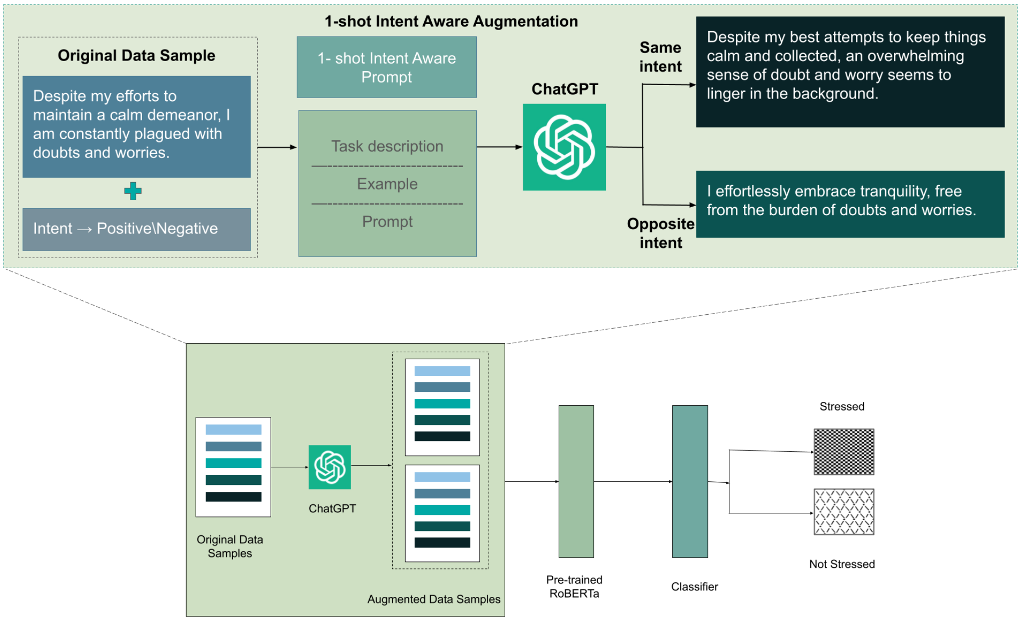
1-shot Intent Aware Augmentation
Original Data Sample
Despite my efforts to maintain a calm demeanor, I am constantly plagued with doubts and worries.
Intent → Positive\Negative
1- shot Intent Aware Prompt
Task description
Example
Prompt
ChatGPT
Same intent
Despite my best attempts to keep things calm and collected, an overwhelming sense of doubt and worry seems to linger in the background.
Opposite intent
I effortlessly embrace tranquility, free from the burden of doubts and worries.
Original Data Samples
ChatGPT
Augmented Data Samples
Pre-trained RoBERTa
Classifier
Stressed
Not Stressed

Figure 1 Overview of proposed methodology.

Full-size DOI: 10.7717/peerjcs.2156/fig-1

Box 1.

"Generate a unique precise reddit user post on r/SR having same topic and S sentiment like the following example post: \n 'input_text'."

Here, SR represents the subreddit of the original post, S represents the sentiment (either 'positive' or 'negative'), and input_text is the input Reddit post.

For a Reddit post on r/anxiety labeled as "not stressed", the prompt is as follows:

Box 2.

"Generate a unique precise reddit user post on r/anxiety having same topic and positive sentiment like the following example post: \n 'Almost decided to live in my car, live with a crack head, travel the country (aka: begin my homeless life, cause I really had no money and if I did it would've gone to beer). I was losing my mind. After being heavily suicidal for a week I decided I can't live like this anymore - but I don't want to die right now. I planted a thought in my head. 'if you ever want to overcome this, you need to begin to change'."

1-shot opposite intent data augmentation

Similar to the previous strategy, this augmentation method involves providing an example post and instructing the generative AI (ChatGPT) to generate a post with the opposite emotion. The prompt selection process for opposite intent augmentation was more exhaustive, and after meticulous experimentation, we settled on the following prompt:

Box 3.

"following post is written by a L person, write a similar post from the perspective of a op_L person: \n 'input_text'."

Here, L represents the true label of the original post, op_L is the opposite label (*e.g.*, if L is "not stressed", op_L is "stressed"), and input_text is the input Reddit post.

For example, the prompt for a Reddit post on r/anxiety labeled as "not stressed" is as follows:

Box 4.

"following post is written by a not stressed person, write a similar post from the perspective of a stressed person: \n 'Almost decided to live in my car, live with a crack head, travel the country (aka: begin my homeless life, cause I really had no money and if I did it would've gone to beer). I was losing my mind. After being heavily suicidal for a week I decided I can't live like this anymore - but I don't want to die right now. I planted a thought in my head. 'if you ever want to overcome this, you need to begin to change'."

We hypothesize that an increase in dataset size through augmentation will lead to improved model performance. Considering the pivotal role of user intent in determining stress, augmenting data with generative AI can provide the classification model with a more nuanced understanding of user intent, surpassing the significance of situational context alone. The expectation is that a more diverse dataset will enhance the stress detection model's accuracy and robustness.

EXPERIMENTAL RESULTS AND EVALUATION

Model training

Our baseline model is RoBERTa (*Liu et al., 2019*), chosen for enhanced training efficiency. Training utilized early stopping with key parameters, including a batch size of 8, learning rate of 1×10^{-6}, Adam optimizer, and binary cross-entropy loss. The dataset was already divided into training and test sets. We then split the training data into training and validation sets with a ratio of 80:20, while keeping the test split unchanged. Data splits and experiments were conducted on an NVIDIA GeForce RTX 2080 Ti using the PyTorch framework.

Model evaluation metrics

Our stress detection model is evaluated using precision, recall, F1 score, and accuracy, tailored to the binary classification (Stressed or not Stressed).

Accuracy

Widely used, accuracy measures correct predictions, providing a broad view of model correctness.

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}}.$$

Precision

Precision measures true positives to total positives, indicating the model's ability to minimize false positives.

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}.$$

Recall

Recall assesses the model's ability to capture all instances of the positive class.

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

F1 score

F1 Score harmonizes precision and recall, providing a balanced measure of the model's effectiveness in binary classification.

$$F1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}.$$

In summary, accuracy, precision, recall, and F1 score collectively offer a nuanced understanding of the stress detection model's effectiveness in binary classification, contributing unique insights to our comprehension.

Implementation of RoBERTa and data preprocessing

During the implementation of the baseline, a notable disparity in the performance of RoBERTa using raw and preprocessed data was observed. The employed data preprocessing techniques included:

1. Converting data to lowercase
2. Removing punctuation marks
3. Removing stop words and frequent words
4. Removing extra spaces and emojis
5. Removing words and digits containing digits
6. Removing URL links and HTML tags
7. Stemming and lemmatization
8. Expanding contractions

Examples of posts, both before and after preprocessing, are presented in Table 4.

We compared the performance of RoBERTa with and without preprocessing, evaluating accuracy, F1 score, recall, and precision (see Table 5). Corresponding confusion matrices

Table 4 **Effect of preprocessing techniques.**

Original post	Preprocessed post
‘He said he had not felt that way before, suggeted I go rest and so ..TRIGGER AHEAD IF YOUIŔE A HYPOCONDRIAC LIKE ME: i decide to look up “feelings of doom” in hopes of maybe getting sucked into some rabbit hole of ludicrous conspiracy, a stupid “are you psychic” test or new age b.s., something I could even laugh at down the road. No, I ended up reading that this sense of doom can be indicative of various health ailments; one of which I am prone to.. So on top of my “doom” to my gloom..I am now fń worried about my heart. I do happen to have a physical in 48 h.’	‘said felt way sugget go rest trigger ahead youir hypocondriac like decid look feel doom hope mayb get suck rabbit hole ludicr conspiraci stupid psychic test new age b someth could even laugh road end read sens doom indic variou health ailment one prone top doom gloomi fn worri heart happen physic hour’
‘Hey there r/assistance, Not sure if this is the right place to post this.. but here goes =) I’m currently a student intern at Sandia National Labs and working on a survey to help improve our marketing outreach efforts at the many schools we recruit at around the country. We’re looking for current undergrad/grad STEM students so if you’re a STEM student or know STEM students, I would greatly appreciate if you can help take or pass along this short survey. As a thank you, everyone who helps take the survey will be entered in to a drawing for chance to win one of three $50 Amazon gcs.’	‘hey rassist sure right place post go im current student intern sandia nation lab work survey help improv market outreach effort mani school recruit around countri look current undergradgrad stem student stem student know stem student would greatli appreci help take pa along short survey thank everyon help take survey enter draw chanc win one three amazon gc’

Table 5 **Performance comparison of RoBERTa with and without preprocessing.**

Method	Precision	Recall	F1-score	Accuracy
With preprocessing	74	74	74	75
Without preprocessing	**81**	**80**	**80**	**80**

are shown in Fig. 2. Results indicate that preprocessing tends to increase the number of posts labeled as stressed, possibly due to an enhanced focus on stressors rather than intent.

The investigation into data preprocessing revealed that RoBERTa performed commendably even without preprocessing. However, preprocessing led to a loss of contextual information.

Zero-shot stress classification with ChatGPT

For a better comparison we also performed zero-shot stress classification using ChatGPT. In contrast to the zero-shot classification results reported in the literature, we achieved an accuracy of 74.96 and an F1 score of 73.91 with precision and recall being 78.25 and 74.39 respectively using the following prompt:

Box 5.

“Classify the following reddit post as ‘stressed’ or ’not stressed’ using labels 1 and 0 respectively. Your response should not be anything except the label:” + post

Even though we achieved results better than that reported in literature, ChatGPT still falls short from RoBERTa in stress classification.

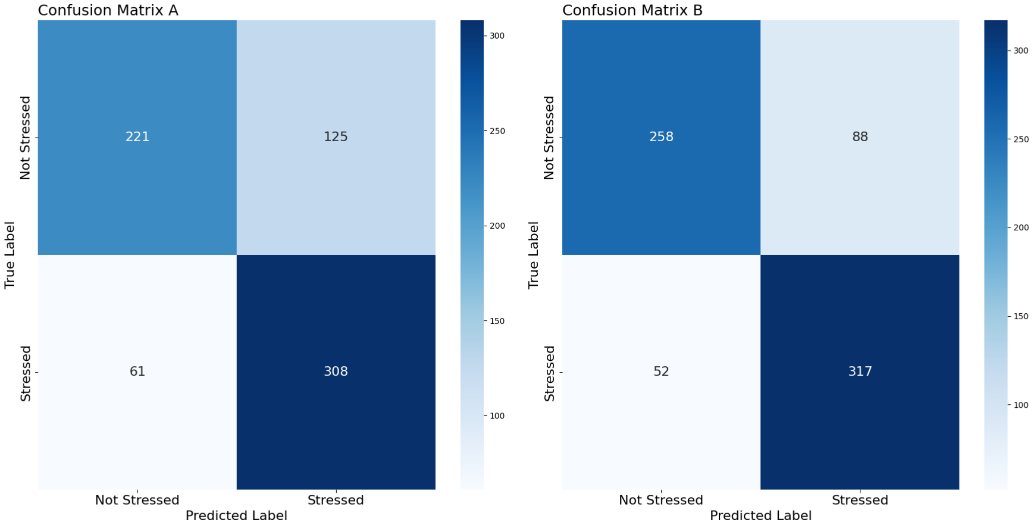

Figure 2 **Confusion matrices.** (A) With preprocessing. (B) Without preprocessing.
Full-size DOI: 10.7717/peerjcs.2156/fig-2

Table 6 **Performance comparison of the data augmentation techniques.** The bolded values represent the statistical metrics of the technique that delivers the best performance in terms of accuracy and F1 score.

Method	Precision	Recall	F1-score	Accuracy
Original	81	80	80	80
BT	81	80	80	*81*
EDA	79	79	79	79
AEDA	80	80	80	80
BT + EDA (concatenated)	80	80	80	80
BT + AEDA (concatenated)	*82*	*81*	*81*	*81*
EDA + AEDA (concatenated)	81	*81*	*81*	*81*
BT + EDA + AEDA (concatenated)	81	80	80	80
BT + EDA (both)	81	*81*	*81*	*81*
BT + AEDA (both)	*82*	*81*	*81*	*81*
BT + EDA + AEDA (both)	81	*81*	*81*	*81*
Opposite intent	**82**	**82**	**82**	**82**
Same intent	**83**	**83**	**83**	**83**

1-shot intent-aware data augmentation

The comparison of experimental results utilizing 1-shot intent-aware data augmentation with state-of-the-art techniques such as Back Translation (BT), EDA, and AEDA individually and in various combinations is presented in Table 6. '*Concatenated*' in the table indicates that the augmentation techniques were applied to data and then the data was concatenated. While '*both*' indicates that back translation was done on data and then the other augmentation technique was applied on both original and back translated texts.

It is noteworthy that intent-aware augmentation markedly enhances the model performance by 3%, with opposite intent augmentation achieving an F1 score of 82%

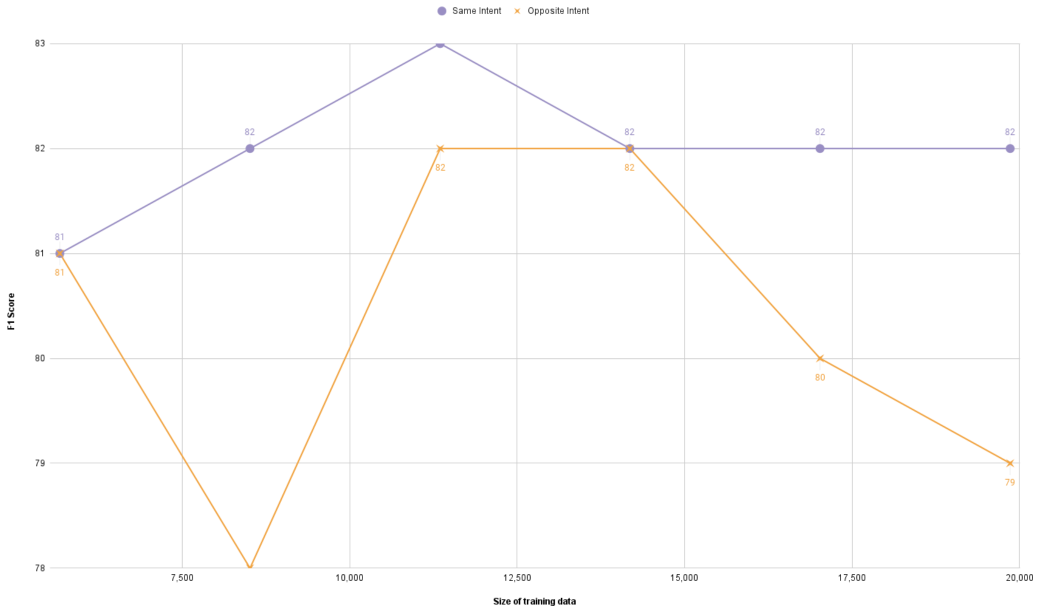

Figure 3 Effect of augmented training data size on F1 score.
Full-size DOI: 10.7717/peerjcs.2156/fig-3

and same intent augmentation yielding an F1 score of 83%. BT and AEDA do not improve the RoBERTa's performance, while EDA causes further degradation.

Size of augmented data

Exploration of different augmentation sizes for our proposed 1-shot intent-aware augmentation technique (same and opposite intent) was conducted where we gradually increased the size of augmented data and observed its effect on the performance of the model as shown in Fig. 3. The observed improvement in model results was noted, but a saturation point in data augmentation effectiveness was identified, emphasizing the need for real-time data beyond a certain threshold.

Statistical analysis of augmented data

We conducted a comprehensive statistical analysis, employing T-test, P-value, and BLEU score to assess the impact of data augmentation on model performance. The statistical values for our proposed technique, compared with state-of-the-art methods, are presented in Table 7.

T-test

The t-test compares means of two groups to determine a significant difference. In data augmentation, it assesses if there is a significant mean difference between the original and augmented datasets. A low p-value compared to t-test indicates a significant difference.

Table 7 **Statistical analysis of data augmentation techniques.** The bolded values represent the statistical metrics of the technique that delivers the best performance in terms of accuracy and F1 score.

Method	*T*-test	*P*-value	Avg. BLEU score
Original	0	1	1
BT	1.01	0.31	0.09
EDA	6.75	0	0.03
AEDA	0	0.99	0.13
Same intent	**2.07**	**0.04**	**0.05**
Opposite intent	0.8	0.42	0.06

P-value

The *p*-value assesses evidence against a null hypothesis in statistical tests. In data augmentation, it determines the significance of the observed mean difference between the original and augmented datasets. A low *p*-value suggests a significant difference.

Bilingual Evaluation Understudy Score

The Bilingual Evaluation Understudy Score (BLEU) score, common in NLP and machine translation, quantifies the similarity between original and augmented datasets. A lower BLEU score indicates less similarity, implying better data diversity.

DISCUSSION

Our examination of the influence of preprocessing on stress detection in social media datasets using RoBERTa has uncovered significant insights. Preprocessing affects RoBERTa's ability to capture sentiment, but it shines with raw data, demonstrating its skill with unfiltered text inputs. In stress detection, model efficacy with raw data, as evident in platforms like Reddit, relies on users expressing emotions through nuanced textual cues like punctuation marks and case variations. This affects the way sentiment is narrated, resulting in RoBERTa's improved performance when using unprocessed data.

The DREADDIT dataset presents a unique challenge as conventional preprocessing techniques, including BT, EDA, and AEDA, show poor performance as seen in Fig. 4. Although EDA demonstrates the highest data diversity in statistical analyses (refer to Table 7), the introduction of noise disrupts the delicate balance in user-conveyed sentiment. This highlights the inadequacy of techniques relying solely on synonym replacement or rephrasing for stress detection.

The significance of intent-aware augmentation for creating posts with preserved sentiment and matching context is emphasized by our findings. 1-shot data augmentation, maintaining consistent sentiment, shows promise, significantly improving model performance. Varied accounts of a stress-inducing event aid the model in comprehending the contextual subtleties and possible triggers for a user's emotional response.

Generative AI for 1-shot opposite intent augmentation faces challenges, despite its superior performance relative to current state-of-the-art techniques. The limitations of RoBERTa become apparent in its difficulty to comprehend conflicting emotions, particularly in complex emotional contexts, exposing challenges in stress classification. The

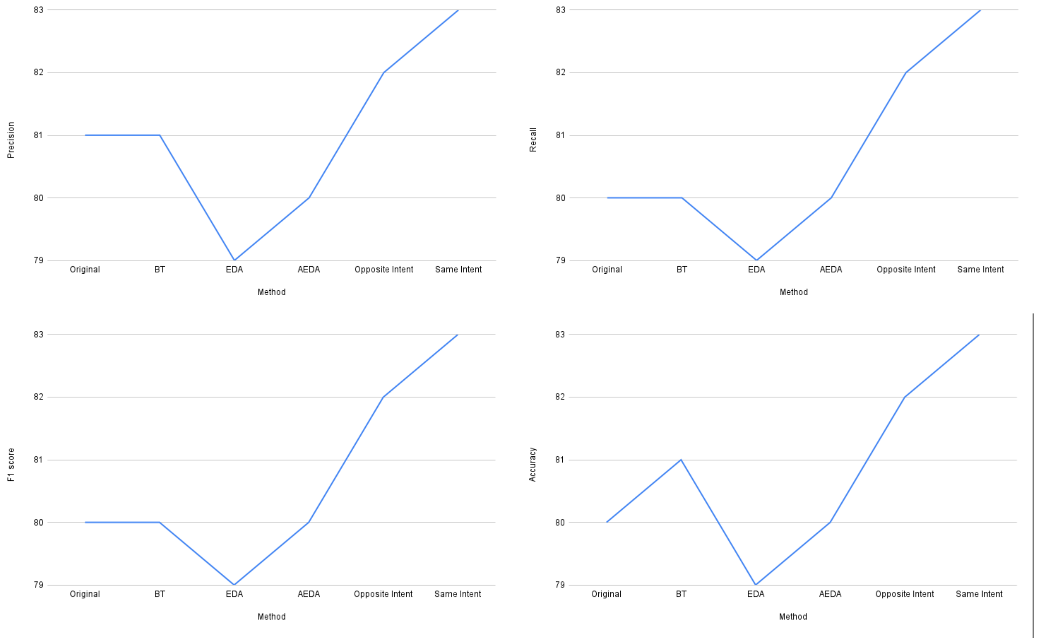

Figure 4 **Graphical performance comparison of the data augmentation techniques.**
Full-size DOI: 10.7717/peerjcs.2156/fig-4

limitation extends to advanced NLP systems, emphasizing their incomplete understanding of human emotions, especially in mental health contexts. It highlights the need to enhance NLP for accurate interpretation and response to complex emotional states.

Our findings stress the importance of real-time data diversity, as 1-shot intent-aware data augmentation yields diminishing returns. While the DREADDIT dataset is valuable for stress classification, its collection period (January 1, 2017, to November 19, 2018) emphasizes the need for more recent datasets that capture the evolving language used on social media platforms.

Moreover, the lackluster zero-shot classification of ChatGPT suggests a reliance on prompts by pre-trained models. Despite being advanced, accurately categorizing human emotions based solely on post context is still difficult. To enhance performance in mental health tasks, it is essential to use concise prompts that describe human emotions.

CONCLUSION

In conclusion, this study focuses on the complexities of stress detection and highlights its importance in mental health applications. Our study examines the effectiveness of intent-based data augmentation using the DREADDIT dataset and RoBERTa, demonstrating a 3% performance improvement with 1-shot techniques. Despite this, the value of real-time diverse data cannot be overstated because of the risk of diminishing returns. Thoughtful preprocessing plays a crucial role in social media-based stress detection, as evidenced by the impact it has on RoBERTa's performance. Our prompt design for ChatGPT, serving as a classifier, outshines other prompts documented in the literature, thereby emphasizing the effectiveness of our prompting technique.

This pioneering work contributes insights into intent-based augmentation for stress detection, scrutinizing preprocessing and generative AI limitations. By acknowledging challenges, we advocate for continuous research and innovation, emphasizing the necessity for diverse, real-time data and improved model architectures for smaller datasets. To enhance intent recognition in stress detection, future work should concentrate on enhancing model architectures for smaller datasets and utilizing unlabeled Reddit data.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2024-2020-0-01789), and the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2024-RS-2023-00254592) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
The MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center): IITP-2024-2020-0-01789.
The Artificial Intelligence Convergence Innovation Human Resources Development supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation): IITP-2024-RS-2023-00254592.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Minhah Saleem conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Jihie Kim conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The data and code are available in the Supplemental Files, at GitHub and Zenodo:

- https://github.com/Minhah-Saleem/IADA

- Minhah Saleem. (2024). Minhah-Saleem/IADA: v1.0 (v1.0). Zenodo. Available at https://doi.org/10.5281/zenodo.10693893.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.2156#supplemental-information.

REFERENCES

Anaby-Tavor A, Carmeli B, Goldbraich E, Kantor A, Kour G, Shlomov S, Tepper N, Zwerdling N. 2020. Do not have enough data? Deep learning to the rescue!. *Proceedings of the AAAI Conference on Artificial Intelligence* **34(05)**:7383–7390.

Ansari G, Garg M, Saxena C. 2021. Data augmentation for mental health classification on social media. In: Bandyopadhyay S, Devi SL, Bhattacharyya P, eds. *Proceedings of the 18th international conference on natural language processing (ICON)*. National Institute of Technology Silchar, Silchar, India: NLP Association of India (NLPAI), 152–161.

American Psychological Association. 2019. Stress. *Available at https://www.apa.org/topics/stress* (accessed on 30 November 2023).

American Psychological Association. 2020. Stress in America™ 2020: a national mental health crisis. *Available at https://www.apa.org/news/press/releases/stress/2020/sia-mental-health-crisis.pdf* (accessed on 30 November 2023).

American Psychological Association. 2022. Stress in America 2022: concerned for the future, beset by inflation. *Available at https://www.apa.org/news/press/releases/stress/2022/concerned-future-inflation* (accessed on 30 November 2023).

Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler D, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D. 2020. Language models are few-shot learners. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H, eds. *Advances in neural information processing systems, vol. 33*. Curran Associates, Inc, 1877–1901.

Chen H, Ji Y, Evans D. 2020. Finding friends and flipping frenemies: automatic paraphrase dataset augmentation using graph theory. In: Cohn T, He Y, Liu Y, eds. *Findings of the Association for Computational Linguistics: EMNLP 2020*. New York: Association for Computational Linguistics, 4741–4751 DOI 10.18653/v1/2020.findings-emnlp.426.

Chen J, Tam D, Raffel C, Bansal M, Yang D. 2023. An empirical survey of data augmentation for limited data learning in NLP. *Transactions of the Association for Computational Linguistics* **11**:191–211.

Corbeil J-P, Ghadivel HA. 2020. Bet: a backtranslation approach for easy data augmentation in transformer-based paraphrase identification context. ArXiv arXiv:2009.12452.

Dai H, Liu Z, Liao W, Huang X, Wu Z, Zhao L, Liu W, Liu N, Li S, Zhu D, Cai H, Li Q, Shen D, Liu T, Li X. 2023. Chataug: leveraging ChatGPT for text data augmentation. ArXiv arXiv:2302.13007.

Devlin J, Chang M-W, Lee K, Toutanova K. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein J, Doran C, Solorio T, eds. *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)*. New York: Association for Computational Linguistics, 4171–4186 DOI 10.18653/v1/N19-1423.

Edunov S, Ott M, Auli M, Grangier D. 2018. Understanding back-translation at scale. In: Riloff E, Chiang D, Hockenmaier J, Tsujii J, eds. *Proceedings of the 2018 conference on empirical methods in natural language processing*. New York: Association for Computational Linguistics, 489–500 DOI 10.18653/v1/D18-1045.

Febriansyah MR, Nicholas , Yunanda R, Suhartono D. 2023. Stress detection system for social media users. *Procedia Computer Science* **216**:672–681 DOI 10.1016/j.procs.2022.12.183.

Feng SY, Gangal V, Wei J, Chandar S, Vosoughi S, Mitamura T, Hovy E. 2021. A survey of data augmentation approaches for NLP. In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 968–988.

Garg M. 2023. Mental health analysis in social media posts: a survey. *Archives of Computational Methods in Engineering* **30(3)**:1819–1842 DOI 10.1007/s11831-022-09863-z.

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. 2014. Generative adversarial nets. *Proceedings of the 27th International Conference on Neural Information Processing Systems* **27**:2672–2680.

Gupta A, Agarwal A, Singh P, Rai P. 2018. A deep generative framework for paraphrase generation. In: *Proceedings of the AAAI conference on artificial intelligence, vol. 32(1)*.

Ilias L, Mouzakitis S, Askounis D. 2023. Calibration of transformer-based models for identifying stress and depression in social media. *IEEE Transactions on Computational Social Systems* 1–12 DOI 10.1109/TCSS.2023.3283009.

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 DOI 10.1007/s44230-023-00020-8.

Iyyer M, Wieting J, Gimpel K, Zettlemoyer L. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In: Walker M, Ji H, Stent A, eds. *Proceedings of the 2018 conference of the North American chapter of the Association for Computational Linguistics: human language technologies, volume 1 (long papers)*. New Orleans, Louisiana: Association for Computational Linguistics, 1875–1885 DOI 10.18653/v1/N18-1170.

Ji S, Zhang T, Ansari L, Fu J, Tiwari P, Cambria E. 2022. MentalBERT: publicly available pretrained language models for mental healthcare. In: Calzolari N, Béchet F, Blache P, Choukri K, Cieri C, Declerck T, Goggi S, Isahara H, Maegaard B, Mariani J, Mazo H, Odijk J, Piperidis S, eds. *Proceedings of the thirteenth language resources and evaluation conference*. Marseille, France: European Language Resources Association, 7184–7190.

Karimi A, Rossi L, Prati A. 2021. AEDA: an easier data augmentation technique for text classification. In: Moens M-F, Huang X, Specia L, Yih SW-t, eds. *Findings of the*

Association for Computational Linguistics: EMNLP 2021. New York: Association for Computational Linguistics, 2748–2754 DOI 10.18653/v1/2021.findings-emnlp.234.

Kingma DP, Welling M. 2013. Auto-encoding variational Bayes. ArXiv arXiv:1312.6114.

Kumar A, Bhattamishra S, Bhandari M, Talukdar P. 2019. Submodular optimization-based diverse paraphrasing and its effectiveness in data augmentation. 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)*. 3609–3619.

Kumar A, Trueman TE, Cambria E. 2022. Stress identification in online social networks. In: *2022 IEEE international conference on data mining workshops (ICDMW)*. 427–434 DOI 10.1109/ICDMW58026.2022.00063.

Kumar V, Choudhary A, Cho E. 2020. Data augmentation using pre-trained transformer models. ArXiv arXiv:2003.02245.

Lamichhane B. 2023. Evaluation of ChatGPT for NLP-based mental health applications. ArXiv arXiv:2303.15727.

Lee K, Guu K, He L, Dozat T, Chung HW. 2021. Neural data augmentation via example extrapolation. ArXiv arXiv:2102.01335.

Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. 2019. RoBERTa: a robustly optimized BERT pretraining approach. ArXiv arXiv:1907.11692.

Mauriello ML, Lincoln T, Hon G, Simon D, Jurafsky D, Paredes P. 2021. SAD: a stress annotated dataset for recognizing everyday stressors in SMS-like conversational systems. In: *Extended abstracts of the 2021 CHI conference on human factors in computing systems, CHI EA '21*. New York: Association for Computing Machinery, DOI 10.1145/3411763.3451799.

Mental Health Foundation. 2018. Stress: are we coping? *Available at https://www.mentalhealth.org.uk/sites/default/files/2022-08/stress-are-we-coping.pdf* (accessed on 30 November 2023).

Murarka A, Radhakrishnan B, Ravichandran S. 2021. Classification of mental illnesses on social media using RoBERTa. In: *Proceedings of the 12th international workshop on health text mining and information analysis*. 59–68.

Naseem U, Lee BC, Khushi M, Kim J, Dunn A. 2022. Benchmarking for public health surveillance tasks on social media with a domain-specific pretrained language model. In: Shavrina T, Mikhailov V, Malykh V, Artemova E, Serikov O, Protasov V, eds. *Proceedings of NLP power! The first workshop on efficient benchmarking in NLP*. New York: Association for Computational Linguistics, 22–31 DOI 10.18653/v1/2022.nlppower-1.3.

Prakash A, Hasan SA, Lee K, Datla V, Qadir A, Liu J, Farri O. 2016. Neural paraphrase generation with stacked residual LSTM networks. In: Matsumoto Y, Prasad R, eds. *Proceedings of COLING 2016, the 26th international conference on computational linguistics: technical papers*. Osaka, Japan: The COLING 2016 Organizing Committee, 2923–2934.

Ramirez-Esparza N, Chung C, Kacewic E, Pennebaker J. 2021. The psychology of word use in depression forums in English and in Spanish: testing two text analytic approaches. *Proceedings of the International AAAI Conference on Web and Social Media* **2(1)**:102–108 DOI 10.1609/icwsm.v2i1.18623.

Rude S, Gortner E-M, Pennebaker J. 2004. Language use of depressed and depression-vulnerable college students. *Cognition and Emotion* **18(8)**:1121–1133 DOI 10.1080/02699930441000030.

Selvadass S, Bruntha PM, Priyadharsini K. 2022. Stress analysis in social media using ML algorithms. In: *2022 4th international conference on smart systems and inventive technology (ICSSIT)*. 1502–1506 DOI 10.1109/ICSSIT53264.2022.9716396.

Statista. 2022. Topic: Reddit. *Available at https://www.statista.com/topics/5672/reddit/#topicOverview* (accessed on 30 November 2023).

Turcan E, McKeown K. 2019. Dreaddit: a Reddit dataset for stress analysis in social media. In: Holderness E, Jimeno Yepes A, Lavelli A, Minard A-L, Pustejovsky J, Rinaldi F, eds. *Proceedings of the tenth international workshop on health text mining and information analysis (LOUHI 2019)*. New York: Association for Computational Linguistics, 97–107 DOI 10.18653/v1/D19-6213.

U.S. Surgeon General Advisory. 2023. Social media and youth mental health. *Available at https://www.hhs.gov/sites/default/files/sg-youth-mental-health-social-media-advisory.pdf* (accessed on 30 November 2023).

Wang X, Zhang H, Cao L, Zeng K, Li Q, Li N, Feng L. 2023. Contrastive learning of stress-specific word embedding for social media based stress detection. In: *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining, KDD '23*. New York: Association for Computing Machinery, 5137–5149 DOI 10.1145/3580305.3599795.

Wei J, Zou K. 2019. EDA: easy data augmentation techniques for boosting performance on text classification tasks. In: Inui K, Jiang J, Ng V, Wan X, eds. *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*. New York: Association for Computational Linguistics, 6382–6388 DOI 10.18653/v1/D19-1670.

Wu X, Lv S, Zang L, Han J, Hu S. 2019. Conditional bert contextual augmentation. In: *Computational science—ICCS 2019: 19th international conference, Faro, Portugal, June 12–14, 2019, Proceedings, Part IV 19*. 84–95.

Xie Q, Dai Z, Hovy E, Luong T, Le Q. 2020. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems* **33**:6256–6268.

Xu X, Yao B, Dong Y, Gabriel S, Yu H, Hendler J, Ghassemi M, Dey AK, Wang D. 2023. Mental-LLM: leveraging large language models for mental health prediction via online text data. ArXiv arXiv:2307.14385.

Yang K, Ji S, Zhang T, Xie Q, Kuang Z, Ananiadou S. 2023. Towards interpretable mental health analysis with large language models. In: *Proceedings of the 2023 conference on empirical methods in natural language processing*. 6056–6077.

Yang K, Zhang T, Ananiadou S. 2022. A mental state knowledge—aware and contrastive network for early stress and depression detection on social media. *Information Processing & Management* **59(4)**:102961 DOI 10.1016/j.ipm.2022.102961.

Yoo KM, Park D, Kang J, Lee S-W, Park W. 2021. GPT3Mix: leveraging large-scale language models for text augmentation. ArXiv arXiv:2104.08826.