Studying the Influence of Toxicity and Emotion Features for Stress Detection on Social Media

Zeyad Alghamdi, Tharindu Kumarage, Mansooreh Karami, Faisal Alatawi, Ahmadreza Mosallanezhad and Huan Liu
School of Computing and Augmented Intelligence, Arizona State University, Tempe, USA
zalgham1@asu.edu
kskumara@asu.edu
mkarami@asu.edu
faalataw@asu.edu
amosalla@asu.edu
huanliu@asu.edu

Abstract: It is crucial to detect and manage stress as early as possible before it becomes a severe mental and physical health problem. Some authors even introduce stress as a "silent killer" to emphasize the significance of early stress management. Traumatic global events such as COVID-19 have amplified stress throughout online communities and it is quite common to see that social media users often vent about their problems or situations online. The ability to detect a person's stress from their posts on social media platforms like Reddit or Twitter in a timely manner can help early stress management and consequently counters mental health conditions. In order to detect stress from social media posts, we must obtain the characteristics that signal a user's stress. Which motivates us to study how salient features influence stress detection. On social media, text-based methods of communication predominantly overtake verbal forms, which makes these platforms a convenient rich medium with an extensive amount of text content to analyze a user's thoughts and emotions. We present a novel approach that helps improve stress detection on social media textual content with sentiment, emotion, and toxicity features. We design our framework based on multiple Transformer-based state-of-the-art sentiment, emotion, and toxicity analysis tools and models for feature extraction and discuss the stress detection tasks' interpretability via inspecting multiple dimensions. For the evaluation, we use publicly available and high-quality datasets where the social media posts are real, carefully selected and labeled. Our experiments show the influence of the proposed new feature dimensions on stress detection by comparing the state-of-the-art baselines and suggesting future directions in stress detection on social media. Furthermore, our extensive feature correlation analysis highlights different aspects, such as 1) *Positive* and *Negative* sentiment, 2) *Joy, Sadness,* and *Fear* emotions, and 3) *Obscene* and *Insult* toxicity as governing factors in improving stress detection performance.

Keywords: Stress Detection, Social Media, Natural Language Processing, Emotional Analysis, Toxicity Analysis

1. Introduction

The national library of medicine defines *stress* as a "feeling of emotional or physical tension. It can come from any event or thought that makes you feel frustrated, angry, or nervous" (MedlinePlus, 2016). Mental Health Foundation (2021) also adds to the definition that stress is our body's response to pressure that makes the person feel anxious, afraid, angry or aggressive, sad, irritable, frustrated, or depressed. Moreover, the American Psychological Association (APA) states that stress is a reaction to extant and future demands and pressures (APA, 2022). It can come from any event or thought that makes a person feel frustrated, angry, or nervous. According to the APA, global stressors are increasing due to reasons such as covid-19, inflation, and political climate. Furthermore, 27% of adults reported that stress impacts their day-to-day functioning; in fact, they report that most days, they are so stressed that they cannot function (APA, 2022). According to Mayo Clinic (2021), stress can harm your body, mood, and behavior, such as headaches, chest pain, anxiety, sadness, irritability, or anger outbursts. Similarly, Yaribeygi et al. (2017) have talked about the criticality of stress's impact on our lives and how it "can be either a triggering or aggravating factor for many diseases and pathological conditions. Therefore, it is crucial to understand stress and stressors more and counter stress before it develops into a complex mental or physical problem.

Many recent works have studied mental health and stress detection on social media as a proxy for identifying real-world stress and stressors (Naseem et al., 2022). This is because social media users often vent about their problems online and share their views, personal situations, and emotions with their followers, friends, and the public (Rastogi et al., 2022). Text-based social methods of communication are dominant when compared to verbal forms. This enables social media platforms to be a suitable medium with an ample amount of text content to study a user's thoughts and emotions. This led to many researchers praising and utilizing natural language processing (NLP) approaches to detect stress in social media (Sravanthi et al., 2020; Shao et al., 2021). State-of-

the-art works in stress detection on social media primarily focus on fine-tuning language models on a stress-labeled dataset (Rastogi et al., 2022; Naseem et al., 2022). Some works focused on studying much more fine-grained features, such as lexical and sentiment characteristics of the social media textual context towards stress detection (Thelwall, 2017; Turcan and McKeown, 2019). However, we could study additional feature dimensions to detect stress in textual content. For example, previously mentioned definitions of stress and its side effects reveal a pattern of stress-related emotions. Such as anger, aggression, and frustration reactions. Moreover, researchers claim that stressed users might exhibit toxic behavior and post toxic content (de Mesquita Neto and Becker, 2018). This necessitates the need to shed light on the emotional and the toxic features and their correlation on stress detection.

In our study, we hypothesize that emotions and toxicity features can correlate to stress labels alongside sentiment, which consequently influences the ss classification performance. We utilize state-of-the-art transformer-based pre-trained language models that were fine-tuned on social media content to extract sentiment, emotions, and toxicity features. Consequently, we use these extracted features to classify stress on social media posts. Furthermore, we conduct a feature correlation study to get a perspective on the effect of each individual sentiment, emotion, and toxic feature categories. To the best of our knowledge, we are the first to study to which extent stress can lead to toxic expressions on social media.

The major contributions of this paper are the following:

- We investigate novel feature dimensions for stress detection, which involves different Toxicity categories alongside the existing sentiment and emotion feature dimensions.
- We offer insights into how different emotion and toxic aspects relate to stress classification on social media.
- Empirical evaluation on benchmark datasets shows that our approach can improve stress detection across multiple social media platforms.

2. Related work

Stress Detection has been studied by researchers on different social media platforms. Turcan and McKeown (2019) have tackled it from a lexical perspective, where they have used Linguistic Inquiry and Word Count (LIWC) to analyze the lexical features of stress. TensiStrength stress and relaxation datasets were built on a lexical approach and a set of rules to detect direct and indirect expressions of stress or relaxation (Thelwall, 2017). Mauriello et al. (2021) conducted a linguistic analysis of the sentences to understand whether the lexical patterns differed enough for relatively accurate classification on their short conversation SMS-like stress dataset. Rastogi et al. (2022) have used popular lexicon-based methods such as VADER, Afinn, and So-cal to classify stress based on the sentiment of the text.

2.1 Sentiment analysis in stress detection

Sentiment analysis is an essential part of stress detection and is implemented in nearly all the studies in this area. Rastogi et al. (2022) have classified stress based on the output of the sentiment analysis models such as VADER, Afinn, and So-cal. Turcan and McKeown (2019) have used the pattern sentiment library for the lexical features. Thelwall (2017) illustrated in their work how stress detection is similar to the existing task of detecting sentiment in text. In recent years, pre-trained transformer-based models have been used to extract sentiment from textual content. For example, Cardiff-NLP's model (Barbieri et al., 2020) is one of the cutting-edge NLP tools for sentiment extraction in social media content as stated by Camacho-Collados et al. (2022) and Matos et al. (2022). Besides, this model has been implemented across multiple domains and related tasks, such as finance (Choya et al., 2022) and opinion polarity analysis (Medina et al., 2022).

2.2 Emotional cues in stress detection

Emotions play an integral part in stress detection. Rastogi et al. (2022) emphasize that analyzing mental health via text data can be done by recognizing emotions closely related to the state of stress/depression, such as anxiety, confusion, and annoyance. Users tend to post content with various negative expressions of emotions when stressed (Turcan and McKeown, 2019). Stress and negative emotions often co-occur, with each able to be a cause and effect of the other (Lazarus, 2006). Thelwall (2017) utilized a lexicon-based approach using terms related to anger and negative emotions. Mauriello et al. (2021) approach to stress detection is from the

emotional turmoil perspective, where stress results from the person's inner perceptions, emotional distress, or anxiety. On the other hand, Rastogi et al. (2022) have focused entirely on emotion classification for stress detection when building datasets and models. For this means, their proposed model classifies the sentences into the six basic emotions theorized by Paul Ekman (1992) (i.e., anger, disgust, fear, joy, sadness, and surprise). We also employ Ekman's model of core emotions because of its known useability in emotion categorization and psycho-linguistic studies (Rozado et al., 2022; Butt, 2022).

2.3 Toxicity relation to stress in social media

Toxicity and stress are often correlated. Fan et al. (2021) focused on the toxic language generated by people from all political spectrums who have taken social media sites like Twitter to vent their frustrations and give their opinions. Additionally, they stated that people become frustrated by the other group and start using toxic language. Similarly, de Mesquita Neto and Becker (2018) have claimed that in situations where stress is elevated, people tend to show toxic behavior and send toxic messages that contain complaining and insults. Emmerich et al. (2020) have suggested that a toxic behavior can be expressed when a person is under stress and finds an outlet to release stress. Moreover, Chong and Kwak (2022) introduced mental health and stress comments as toxicity triggers. Almerekhi et al. (2020) illustrate in their work that toxicity can be contagious and that toxic comments trigger other users to post toxic responses. Here we see a vicious cycle where stress influences toxic content and vice versa. Consequently, it motivates us to study the relationship between stress and toxicity in social media.

3. Methodology

In this section, we go through the proposed approach. Figure 1 indicates the model's architecture. As previously mentioned, our model and experiments aim to understand how emotion and toxicity affect stress detection. To this end, we use components to extract the emotion and toxicity of text. Mainly our approach consists of the following three feature extraction components: (1) a sentiment analysis component to extract the sentiment of text, (2) a toxicity component to detect the toxicity-related scores of a given text, and (3) an emotion analysis component that categorizes the emotion of the text.

These components are based on pre-trained transformers, which are deep neural networks that learn the words' context and, consequently, the meaning via a mechanism known as self-attention, a form of differentially weighting the significance of each part of the input sentence when constructing word embeddings. These transformer architectures have reached prediction accuracies that match human annotations for text classification tasks (Devlin et al., 2018). We go through the details of each component in the following subsections.

3.1 Sentiment analysis component

To detect the general sentiment of text, we use an existing pre-trained transformer-based language model, RoBERTa, that was fine-tuned on millions of social media textual content (i.e., Tweets and posts) (F. Barbieri et al., 2020). This model outputs three scores, indicating the probability of the input text being *positive, negative,* or *neutral.*

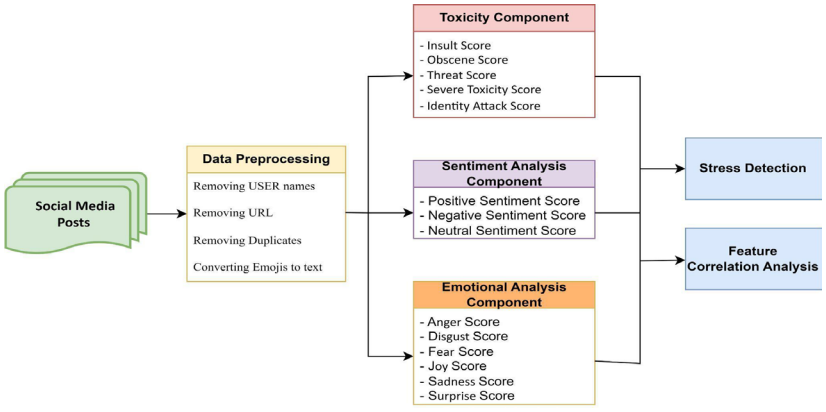

Figure 1: Our approach pipeline. It consists of Data input and preprocessing, Feature Extraction stage (i.e., Toxicity, Sentiment, and Emotions), Stress Classification, and Correlation Analysis.

3.2 Emotions analysis component

Emotions were calculated using a publicly available Transformer Distil-RoBERTa-based model. This model was trained and fine-tuned on a combination of multiple emotion datasets for English, predicting Ekman's (1992) six basic emotions: *anger, disgust, fear, joy, sadness,* and *surprise*. The dataset used to train and fine-tune this model is diverse and is a combination of several publicly available datasets. The datasets include social media data from Twitter and Reddit, student self-post reports, and TV dialogs. Using such a diverse dataset increases the model's generalizability (Hartmann, 2022).

3.3 Toxicity detection component

To analyze the toxicity of the textual content, we used the state-of-the-art model, Detoxify (Hanu and Unitary team, 2020). Detoxify is based on transformer model BERT-base-uncased and has two variants, depending on the training set: (1) "Original" Detoxify, which is trained on Wikipedia comments, and (2) "Unbiased" Detoxify, which is trained based on the Jigsaw dataset. Both models output the probability of a given text in the following categories: *toxicity, severe toxicity, obscene, threat, insult*, and *identity attack*. We have tested our framework with both variants. Due to the performance similarity between the two variants, we use the Original Detoxify throughout the paper.

4. Experimental setting

In this section, we present details about the datasets and the implementation details of feature extraction and stress classification.

4.1 Datasets

We selected multiple popular datasets curated around the stress detection task. Our focus was on versatile datasets that can showcase the performance of stress detection models and their interpretability. The datasets were of different textual lengths, from short Twitter tweets to long Reddit posts. All the datasets' contents were pre-processed by removing the usernames and links and duplicates in addition to converting the emojis to text. It is worth mentioning that not all datasets have balanced label distribution. A summary of the datasets used is in Table 1. We go through the details of each dataset in the following subsections.

Table 1: Experimented Datasets and characteristics. The datasets are called balanced if the distribution of labels is around 50% (±5%).

Datasets	Label Balance	Topics (subreddits or hashtags)	Dataset Collection Period
Dreaddit Reddit (Turcan and McKeown, 2019)	balanced	Interpersonal conflict, Mental illness, financial need: r/relationships, r/homeless, r/food pantry, r/assistance, r/stress, r/ptsd, r/almosthomeless,r/anxiety, r/survivorsofabuse, r/domesticviolence	Between January 2017 and November 2018.
Social media articles Twitter (Rastogi, et al., 2022)	Twitter-Full balanced	Stress and non-stress: #Stress, #Stressed, #Tired, #FeelingUseless, #MentalHealthMatters, #MentalHealth, #FeelingStressed, #IamStressed, #Fatigued, #PandemicBlues, #Happiness, #Happy, #Delighted,#Joy,#Blessed.	Between September 2019 and September 2021.
	Twitter Non-Advert unbalanced		
Social media articles Reddit (Rastogi et al., 2022)	Reddit Title balanced	Stress and non-stress: r\Stressed, r\Stress, r\Depressing, r\Depression r\MentalHealth, r\Happy, r\MadeMeSmile, r\MakeMeSmile, r\Wholesome.	
	Reddit Combined unbalanced		

4.1.1 Dreaddit

Turcan and McKeown (2019) collected Reddit posts and then annotated them using Amazon Mechanical Turk (AMT) (Amazon Mechanical Turk, 2018) to determine whether the posts exhibited the stress of the person who posted. It is an excellent resource for analyzing Reddit corpus. The dataset excels due to its complex and diverse data and characteristics across multiple topics, as shown in Table 1. The dataset includes 3,553 posts and focuses on major stressful topics such as homelessness and domestic violence. Posts in Dreaddit are collected from 10 subreddits. Naseem et al. (2022) used Dreaddit to investigate public health surveillance on social media and evaluate stress detection using social media posts.

4.1.2 Social media articles: Twitter and Reddit

Rastogi et al. (2022) collected a stress detection dataset from Reddit and Twitter. According to the datasets' creators, additionally, it effectively depicts the widespread stress levels caused by COVID-19. The two datasets were labeled automatically using a rule-based and pre-trained language model (PLM). They both include tweets and posts related to stress and non-stress topics, as shown in Table 1. In addition, the authors provided two Twitter datasets: (1) Twitter-Full: a dataset that includes all tweets, and (2) Twitter Non-Advert: which exclude tweets related to advertisements. Also, Reddit has two datasets: (1) Reddit Combined: includes posts body and title, and (2) Reddit Title: includes titles only.

4.2 Implementation details

Our experiment consisted of acquiring the scores for different toxicity categories mentioned in section 3.3. In addition to the six emotional features scores discussed in section 3.2 and the sentiment scores mentioned in 3.1. That will give us a total of 14 features. The evaluation was based on our model's accuracy in detecting the label of stress and the F1 score, which is an essential metric due to some datasets' imbalanced labels. This micro-averaging score will help us depict a comparable performance.

In order to evaluate the effectiveness of the model, we conducted a comparative analysis using multiple classification algorithms, including Random Forest, Multi-layer Perceptron (MLP), Adaptive Boosting, Gradient Boosting, and Extreme Gradient Boosting (XGBoost). Our experiments indicated that, on average, gradient boosting achieved higher accuracy compared to other classification algorithms. Specifically, for the Dreadit and Reddit title dataset Twitter non-advert, a majority of the feature combinations yielded high accuracy when using Gradient Boosting. Thus, we use Gradient Boost to present the findings of our work, as demonstrated in Table 2. Furthermore, for the Twitter-full dataset, Gradient Boosting was among the top performers in terms of accuracy. However, it was slightly behind the top performer, with an average difference of less than 1% compared to other algorithms. Please refer to **Appendix Table A.1** for more details.

We employed the StratifiedKFold cross-validation method to balance the class labels per fold and prevent overfitting to the dataset. Additionally, to examine the correlation and interpretability of the features, we utilized Pearson's correlation coefficient to measure the statistical relationship between variables. Furthermore, we conducted hypothesis testing using the P-value to confirm the statistical significance of our findings.

5. Experiment results and discussion

In this section, we show an ablation study where we reveal each feature set's influence on the stress detection task. To present and discuss our results, we categorize our experiment into the following two research questions.

RQ1. Can we use emotions and toxicity as features to better detect stress?

Our study shows that, yes, using emotions and toxicity can provide an improvement to the task of stress detection. Our experiment results in Table 2 show that combining all the features, including emotions and toxicity, gives us better F1 and accuracy scores. Moreover, we compare our results with the state-of-the-art baselines provided by the authors of the datasets Turcan and McKeown (2019) and Rastogi et al. (2022). Our Gradient boosting classifier with all three feature categories outperforms the lexical methods presented in the above papers. Also, for some datasets, our approach outperforms the transformer-based model. For example, in the Dreaddit dataset, the authors achieved an F1 score of 80.65% via the pre-training Bert-base model. In comparison, our approach accomplished an increase of 2.23% to the F1 score without the need to fine-tune a large language model-based classifier. Meanwhile, for the other datasets, the pre-trained Roberta model

achieved an average increase of 3.94% in the F1 score compared to our model. However, this difference in performance is minimal if we consider that our approach needed no language model training and fine-tuning on stress-labeled datasets. In particular, while our approach uses a maximum of 14 features and a simple classifier training to detect stress, the Roberta model uses millions of parameters with an extensive fine-tuning process. Therefore, in addition to the simplicity, the performance of our approach is competitive and can be relied on as a baseline for detecting stress.

To further illustrate how significant each feature category is for the task of stress classification, we compare the average improvement when using sentiment alone versus when using sentiment, emotional, and toxicity. Here we observe an average improvement of 2.14% in accuracy and an average improvement of 1.85% in F1 score for all datasets. On the other hand, if we compare the average improvement when using the combinations of toxicity features and sentiment versus sentiment alone, we get an increase of 0.57% in accuracy and 0.7% in the F1 score. Here we emphasize that our aim in this study is not to build the best-performing stress classifier but to study the influence of toxicity and emotions on the task at hand. As a result, we conducted further analysis of our next research question to examine the proposed features and their implication for stress detection.

Table 2: Overall stress detection performance (Gradient Boosting model) using Accuracy and F1 scores. We mainly focus on F1 due to imbalanced labeling in some datasets.

Best Features of each model or class/dataset	Dreaddit		Twitter Full		Twitter non-advert		Reddit combined		Reddit title	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Toxicity features	68.44	72.78	63.51	66.56	68.1	75.95	87.84	93.46	71.13	69.12
Emotional features	75.35	77.62	80.57	81.68	81.27	84.12	91.04	94.96	89.14	89.03
Sentiment features	81.56	82.62	80.15	80.49	80.51	83.58	91.52	95.23	90.59	90.28
Toxicity, Emotional features	77.13	79.36	80.75	81.89	82.28	84.98	91.68	95.34	90.11	89.28
Sentiment, Emotional features	80.32	81.83	82.58	83.2	83.8	86.15	92.16	95.56	91.49	91.31
Sentiment, toxicity features	81.21	82.85	80.98	81.52	81.27	84.12	92.03	95.49	91.68	91.73
All features	**82.37**	**82.88**	**82.58**	**83.24**	**85.06**	**87.31**	**92.96**	**96.2**	**92.04**	**91.84**

RQ2. Which sentiment, emotion, and toxic aspects benefit stress detection the most?

This research question explores the influence of all the sentiment, emotion, and toxic features on the stress label via correlation analysis. We started by studying the affect of different sentiments on stress detection. As shown in Table 3, the *positive* sentiment has the highest inverse correlation for the stress label for all datasets, with an absolute average of 0.67. Additionally, the *negative* sentiment is directly correlated to the stress label for all datasets, with an absolute average of 0.59. Sentiment analysis is an essential part of stress detection and has been implemented in previous research Rastogi et al. (2022) and Thelwall (2017). We observe that neutral sentiment has conflicting correlations across the different datasets. We attribute this abnormality to the dataset labeling process. For instance, Turcan and McKeown (2019) Dreaddit dataset annotation was manual, and the annotators were instructed to label a post as stressful only if that post displayed stress and a negative attitude. In contrast, Rastogi et al. (2022) labeling process labeled a post as stressful if that post displayed stress and either a neutral or negative attitude. As a result, we see a positive correlation between the stress label to the neutral sentiment in the (Rastogi et al., 2022) datasets.

For Emotions, we observe that the emotion of *Joy* is inversely correlated to stress, and it shows the most emotional absolute average correlation, with an absolute average of around 0.54. *Sadness* and *Fear* are directly correlated to stress and only show an average weak correlation of around 0.25 and 0.27, respectively, across all datasets. We found that these findings are consistent with the definition of stress and its relationship with different types of emotions (Mental Health Foundation, 2021; Mayo Clinic, 2021). Even though *Anger* emotion was mentioned in many definitions of stress and stress behavior (MedlinePlus, 2016; Mental Health Foundation, 2021) and (APA, 2022; Mayo Clinic, 2021), it has a weaker correlation with an average of 0.16. This tells us that not all stress-related emotions will appear in social media textual content. Moreover, based on our observations, the average correlations of *Disgust* and *Surprise* to stress are relatively much lower than other emotions.

Table 3: Pearson correlation scores for all features extracted across all datasets. We bold the feature with the most correlation for every feature set for all datasets. The asterisk * features have p>0.05.

Feature/ Dataset		Dreaddit	Twitter Full	Twitter Non-Advert	Reddit Combined	Reddit Title	Absolute Average Correlation |r|
Sentiment	**Negative**	**0.610**	0.513	0.47	0.624	0.766	*0.597*
	Positive	-0.481	**-0.686**	**-0.696**	**-0.71**	**-0.807**	***0.676***
	Neutral	-0.398	0.463	0.444	-0.007*	0.225	0.307
Emotions	**Joy**	-0.260	**-0.582**	**-0.554**	**-0.622**	**-0.664**	***0.536***
	Sadness	0.162	0.258	0.211	0.246	0.402	*0.256*
	Fear	0.285	0.339	0.360	0.15	0.22	*0.271*
	Anger	0.107	0.165	0.148	0.121	0.261	0.160
	Disgust	0.061	0.162	0.16	0.100	0.219	0.140
	Surprise	-0.082	-0.049	-0.102	-0.131	-0.117	0.096
Toxicity	**Insult**	0.089	**0.102**	**0.090**	0.085	0.125	0.098
	Obscene	**0.151**	0.079	0.047	**0.123**	**0.131**	**0.106**
	Threat	0.058	0.044	0.046	0.057	0.121	0.065
	Severe Toxicity	0.085	0.052	0.0278*	0.067	0.126	0.072
	Identity Attack	0.054	0.056	0.085	0.054	0.038	0.057

Finally, we observe that the toxicity features' presence is mainly in *the Insult* and *Obscene* categories, with an average Pearson correlation of about 0.1, and 0.106, respectively. Nonetheless, other toxicity categories, such as *Threat, Severe Toxicity, and Identity Attack*, show an average below 0.1 to stress textual content. Despite toxic behavior being linked to stress, as other researchers have suggested (Fan et al., 2021; de Mesquita Neto and Becker, 2018; Emmerich et al., 2020), we observe a weak correlation between stress and toxicity. Consequently, indicating that a stressed person does not necessarily exhibit toxic behavior.

To summarize, the correlation coefficients were statistically significant, revealing that Positive and Negative sentiments and the emotions of (joy, sadness, and fear) are essential features for stress detection. Moreover, contradictory to what has been stated in stress and stress behavior definitions and related works, the anger emotions and toxicity features show a much weaker correlation to stress. This indicates that a stressed person does not always exhibit toxic behavior.

6. Conclusion and future work

Stress is crucial to detect with all means possible. Stress in social media can be detected more accurately if we include meaningful features. Our approach of using toxicity and emotional features in addition to the sentiment features shows potential compared to the baselines. Although including emotion and toxicity features improve stress detection, the performance difference indicates there is still room for improvement. We have shown that in a subset of features, such as positive and negative sentiment, the emotion of joy correlates mostly with stress. On top of that, some emotions, such as sadness and fear, have a weak to moderate correlation to stress. Moreover, based on the previously mentioned stress definitions and behaviors, the anger emotion should significantly correlate to stress. However, our observation is that anger emotion only shows a low correlation to stress. Similarly, related works claim stress and toxicity behavior are linked, yet we observe toxicity features having a weaker correlation than previously mentioned features.

A limitation we faced was finding a fairground to compete with pre-trained transformer baselines. As we have mentioned, our approach achieved competitive performance results. We stipulate that by fusing our acquired features with textual embedding, we can achieve greater accuracy and F1 scores. Nonetheless, that can be a future project to create a robust, generalizable stress classifier. Moreover, the bias in data selection might affect the toxicity and stress relationship analysis. For instance, there is a potential that the posts expressing stress in a toxic manner were not collected nor permitted to remain on the social media platform due to possibly violating community guidelines or being too toxic (Twitter Help Center, 2019; Reddit, 2019).

An interesting future direction is to explore classification bias and generalizability. Previous studies have shown that bias and generalizability are connected, and debiasing machine learning and deep learning models can lead to more generalization (Chuang, 2021). Additionally, we plan to explore some aspects of stylometry (writing style of the user), such as punctuation, as seen in the work of (Karami, 2022) but in the context of stress detection.

Acknowledgment:

This work was supported by the Office of Naval Research under Award No. N00014-21-1-4002. Opinions, interpretations, conclusions, and recommendations are those of the authors.

References

Almerekhi, H., Kwak, H., Salminen, J. and Jansen, B.J., 2020, April. Are these comments triggering? predicting triggers of toxicity in online discussions. In Proceedings of The Web Conference 2020 (pp. 3033-3040).

Amazon Mechanical Turk (2018). Amazon mechanical turk. [online] Mturk.com. Available at: www.mturk.com/.

Apa.org. (2022). Stress in America 2022. [online] Available at: https://www.apa.org/news/press/releases/stress/2022/concerned-future-inflation.

Butt, S., Sharma, S., Sharma, R., Sidorov, G. and Gelbukh, A., 2022. What goes on inside rumour and non-rumour tweets and their reactions: A Psycholinguistic Analyses. Computers in Human Behavior, p.107345.

Camacho-Collados, J., Rezaee, K., Riahi, T., Ushio, A., Loureiro, D., Antypas, D., Boisson, J., Espinosa-Anke, L., Liu, F., Martínez-Cámara, E. and Medina, G., 2022. Tweetnlp: Cutting-edge natural language processing for social media. arXiv preprint arXiv:2206.14774.

Chong, Y.Y. and Kwak, H., 2022, May. Understanding Toxicity Triggers on Reddit in the Context of Singapore. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 16, pp. 1383-1387).

Choya, J., Wanga, B., AlShelahia, A. and Saigala, R.,2022. External Forces on Financial Markets: Evidence from the GameStop Short Squeeze and Flash Crash.

Chuang, Y.S., Gao, M., Luo, H., Glass, J., Lee, H.Y., Chen, Y.N. and Li, S.W., 2021, August. Mitigating Biases in Toxic Language Detection through Invariant Rationalization. In Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021) (pp. 114-120).

de Mesquita Neto, J.A. and Becker, K., 2018. Relating conversational topics and toxic behavior effects in a MOBA game. Entertainment computing, 26, pp.10-29.

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2019, January. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In NAACL-HLT (1).

Ekman P., "An argument for basic emotions," Cognition and Emotion, vol. 6, no. 3–4, pp. 169–200, May 1992,

Emmerich, K., Krekhov, A. and Krüger, J., 2020, November. " Pls Uninstall" On the Interplay of the COVID-19 Pandemic and Toxic Player Behavior in Competitive Gaming. In Extended Abstracts of the 2020 Annual Symposium on Computer-Human Interaction in Play (pp. 224-228).

Fan, H., Du, W., Dahou, A., Ewees, A.A., Yousri, D., Elaziz, M.A., Elsheikh, A.H., Abualigah, L. and Al-qaness, M.A., 2021. Social media toxicity classification using deep learning: Real-world application uk brexit. Electronics, 10(11), p.1332.

F. Barbieri, J. Camacho-Collados, L. Neves, and L. Espinosa-Anke. Twitter-roberta-base for sentiment analysis. https://huggingface.co/cardiffnlp/twitter-roberta-basesentiment, 2020..

Hanu, L., 2020. Unitary team. Detoxify. Github. https://github.com/unitaryai/detoxify.

Hartmann, J. (2022) Emotion English DistilRoBERTa-base. Available at: https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/.

Hoang, P.G., Nguyen, L.T. and Nguyen, K., 2021, August. UIT-E10dot3 at SemEval-2021 Task 5: Toxic Spans Detection with Named Entity Recognition and Question-Answering Approaches. In Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021) (pp. 919-926).

Ji, S., Zhang, T., Ansari, L., Fu, J., Tiwari, P. and Cambria, E., 2021. Mentalbert: Publicly available pretrained language models for mental healthcare. arXiv preprint arXiv:2110.15621.

Karami, M., Mosallanezhad, A., Mancenido, M. V., & Liu, H. 2022. "Let's Eat Grandma": Does Punctuation Matter in Sentence Representation?.

Lazarus, R. S. (2006). Stress and emotion: A new synthesis. Springer Publishing Company.

Matos, B., Lima, R.C., Almeida, J.M., Gonçalves, M.A. and Santos, R.L., 2022. On the Presence of Abusive Language in Mis/Disinformation. In the International Conference on Social Informatics (pp. 292-304). Springer, Cham.

Mayo Clinic (2021). Stress Management. [online] Mayo Clinic. Available at: www.mayoclinic.org/healthy-lifestyle/stress-management/in-depth/stress-symptoms/art-20050987.

Mauriello, M.L., Lincoln, T., Hon, G., Simon, D., Jurafsky, D. and Paredes, P., 2021, May. 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 (pp. 1-7).

Medina, G.M., Collados, J.C. and Cámara, E.M., 2022. NLP4SM: Natural Language Processing for social media.

MedlinePlus (2016). Stress and your health. [online] Medlineplus. Available at: https://medlineplus.gov/ency/article/003211.htm.

Mental Health Foundation (2021). Stress. [online] www.mentalhealth.org.uk. Available at: https://www.mentalhealth.org.uk/explore-mental-health/a-z-topics/stress.

Naseem, U., Lee, B.C., Khushi, M., Kim, J. and Dunn, A., 2022, May. Benchmarking for Public Health Surveillance tasks on Social Media with a Domain-Specific Pretrained Language Model. In Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP (pp. 22-31).

Rastogi, A., Liu, Q. and Cambria, E., 2022, July. Stress Detection from Social Media Articles: New Dataset Benchmark and Analytical Study. In 2022 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.

Reddit (2019). Content Policy - Reddit. [online] Redditinc.com. Available at: https://www.redditinc.com/policies/content-policy.

Rozado, D., Hughes, R. and Halberstadt, J., 2022. Longitudinal analysis of sentiment and emotion in news media headlines using automated labelling with Transformer language models. Plos one, 17(10), p.e0276367.

Shao, R., Shi, Z. and Zhang, D., 2021. Social media and emotional burnout regulation during the COVID-19 pandemic: multilevel approach. Journal of Medical Internet Research, 23(3), p.e27015.

Sravanthi, T., Hema, V., Reddy, S.T., Mahender, K. and Venkateshwarlu, S., 2020, December. Detection of Mentally Distressed Social Media Profiles Using Machine Learning Techniques. In IOP Conference Series: Materials Science and Engineering (Vol. 981, No. 2, p. 022056). IOP Publishing.

Thelwall, M., 2017. TensiStrength: Stress and relaxation magnitude detection for social media texts. Information Processing & Management, 53(1), pp.106-121.

Turcan, E. and Mckeown, K., 2019, November. Dreaddit: A Reddit Dataset for Stress Analysis in Social Media. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)* (pp. 97-107)..

Twitter Help Center (2019). The Twitter Rules. [online] Twitter.com. Available at: https://help.twitter.com/en/rules-and-policies/twitter-rules.

Yaribeygi, H., Panahi, Y., Sahraei, H., Johnston, T.P. and Sahebkar, A., 2017. The impact of stress on body *function: A review. EXCLI journal, 16, p.1057.*

Appendix:

Table A.1: Additional detailed results comparing multiple classification algorithms (Gradient boosting= GB, random-forest=RF, Multi-layer Perceptron= MLP, ADA boosting= ADAB, XGboost= XGB)

Feature set / Accuracy	*GB*	*RF*	*MLP*	ADAB	*XGB*	*GB*	*RF*	*MLP*	ADAB	*XGB*
	Dreadit					*Reddit title*				
Toxicity features	**68.44**	66.31	61.17	68.26	69.5	**71.13**	67.33	56.56	65.16	67.78
Emotional features	75.35	74.11	**78.01**	74.11	74.82	89.14	**89.32**	88.24	88.33	88.6
Sentiment features	**81.56**	76.42	80.5	79.96	81.21	**90.59**	89.14	90.23	**90.59**	90.5
Toxicity, Emotional features	**77.13**	75.35	74.82	76.42	76.42	**90.11**	89.32	88.42	88.96	89.5
Sentiment, Emotional features	80.32	80.14	**81.21**	79.43	80.67	91.49	91.49	**91.86**	91.04	91.76

Feature set / Accuracy	*GB*	*RF*	*MLP*	ADAB	*XGB*	*GB*	*RF*	*MLP*	ADAB	*XGB*
	Dreadit					*Reddit title*				
Toxicity features	**68.44**	66.31	61.17	68.26	69.5	**71.13**	67.33	56.56	65.16	67.78
Emotional features	75.35	74.11	**78.01**	74.11	74.82	89.14	**89.32**	88.24	88.33	88.6
Sentiment features	**81.56**	76.42	80.5	79.96	81.21	**90.59**	89.14	90.23	**90.59**	90.5
Toxicity, Emotional features	**77.13**	75.35	74.82	76.42	76.42	**90.11**	89.32	88.42	88.96	89.5
Sentiment, toxicity features	**81.21**	80.14	80.32	79.79	80.67	**91.68**	91.04	91.04	90.68	91.4
All features	**82.37**	80.67	80.32	78.19	81.21	92.04	91.58	91.67	92.04	**92.58**

Feature set / Accuracy	*GB*	*RF*	*MLP*	ADAB	*XGB*	*GB*	*RF*	*MLP*	ADAB	*XGB*
	Twitter non-advert					*Reddit combi*				
Toxicity features	**68.1**	65.57	57.72	66.33	66.33	87.84	87.2	87.68	87.2	**88.0**
Emotional features	**81.27**	80.25	79.75	80.51	81.01	91.04	91.36	91.84	**92.48**	91.85
Sentiment features	80.51	78.67	**81.4**	81.28	81.16	**91.52**	91.2	92.16	**91.52**	91.04
Toxicity, Emotional features	**82.28**	79.75	80.25	81.52	81.77	91.68	92.0	**92.32**	92.16	91.84
Sentiment, Emotional features	83.8	**84.56**	**84.56**	81.77	83.54	92.16	92.8	93.6	92.16	**92.8**
Sentiment, toxicity features	81.27	81.77	**82.53**	80.76	82.03	92.03	91.68	**92.48**	91.04	91.2
All features	**85.06**	84.3	**85.06**	83.29	83.8	**92.96**	92.48	93.28	92.32	92.0