Stress Detection from Social Media Articles: New Dataset Benchmark and Analytical Study

Aryan Rastogi[1], Qian Liu[2], Erik Cambria[2]
1. Department of Electrical Engineering, Indian Institute of Technology Indore, India
2. School of Computer Science and Engineering, Nanyang Technological University, Singapore

Abstract—**Stress detection is a basic and essential task for examining the mental health of a given population. With the rapid digitalization leading to text-based forms of communication gaining dominance over spoken ones, there is now the chance to develop analytical studies for stress detection directly from textual inputs in social media. However, only a limited number of benchmarks are publicly available. To this end, we create four high quality datasets based on Twitter and Reddit, which are designed particularly for the task of stress detection from social media texts. The main contributions are three-folds: 1) for each dataset, we provide a detailed description on our dataset construction process, including data collection, data preprocessing and annotation; 2) we perform a comparative study on the performance of different rule-based and machine learning-based approaches on the proposed datasets as the new benchmark of this field; 3) we study the feasibility and reliability of existing systems. Extensive experiments show that Transformer-based models outperform lexical-based and embedding-based methods. Also, we observe that existing methods based on sentiment polarity detection cannot be directly adapted to the task of stress detection and there remains space for further improvements. We hope this study could pave the way for future studies in the field of computational stress detection.**

Index Terms—**Stress Detection, Sentiment Analysis**

I. Introduction

Emotion classification is one of the widespread and well known applications in the domain of text analytics [1]–[3]. It provides valuable information about the underlying sentiment and tone of the text, and helps in inferring the overall reception of the public towards a certain topic. Stress detection [4] is a special case of emotion classification wherein smart systems are developed to detect underlying stress/depression in individuals. Stress is defined as a state of imbalance between one's internal demands and his/her ability to meet those demands [5], and is widely regarded as a medical problem. A number of surveys carried out during the Covid-19 pandemic [6] has substantiated the occurrence of stress amongst people belonging to different walks of life. Thus, stress detection is important to adjudge the mental health of a populace, and it provides the government agencies and the private corporations valuable feedback about the potential stress-inducing factors, and allows them to frame their policies and agendas taking into due consideration the mental stability of their people [7], [8]. While a number of studies have been carried out to detect stress in individuals based on physiological and biological parameters [9], [10], the field dealing with stress detection from text based input has emerged as one of the popular areas of research in the recent times.

The reason for this is the high digital presence of people on social media, due to which they often share their views, experiences and personal emotions with their followers and friends. Hence, data acquired from Social Media platforms offer a ready resource for building a stress detection system by leveraging the power of machine learning based models.

However, there are a number of challenges towards creating a high quality dataset for the task of stress detection. The main reason is that the data fetched from social media is highly interlaced with noise, which may affect the model predictions negatively. To alleviate this problem, this work primarily seeks to address the aforementioned issues in creating smart systems for stress detection from social media articles. Our basic idea for dataset preparation is to utilize transfer learning to develop automated systems for labelling of raw data, since it is quite cumbersome to enlist manual annotators for the same and it only serves to reduce the effectiveness of the overall detection system. More specifically, we construct our datasets[1] from two of the most popular social media platforms - Twitter and Reddit.

This ensures that a balanced study is conducted based on the age demographics; Twitter includes representation from individuals from all age groups [11], Reddit is mostly used by teenagers and people in their early twenties [12]. Furthermore, we use transfer learning to develop smart systems for data annotation, and then provide a comparative study on the performance of different well-known lexicon, embedding and pre-trained language models (PLMs) on our datasets. The applicability of the sentiment polarity classification task for stress classification is also explored. We also show the effectiveness of PLMs in achieving state-of-the-art results as compared to other approaches for this task.

Our main contributions can be summarized as follows:

1) We construct four stress detection datasets based on two most popular social media platforms - Reddit and Twitter, which are released to the research community to better understand the nature and extent of stress in the population and address rising mental health concerns. This would address two challenges: lack of large-scale datasets for model learning, and laborious data annotation, which compromise learning a smart and reliable stress detection system.

[1] https://github.com/senticnet/stress-detection

2) To save the annotation cost and enable flexibility to the future updates, we design effective fully automated methods to denoise and annotate the unlabelled corpus by using a combination of rule-based and PLM-based systems. Additionally, we enquire whether existing systems trained on sentiment polarity classification can be applied to detect stress from the text or not. Extensive experiments show that benchmark results are obtained by using BERT-based models via transfer learning, and there is still plenty of room for improvement.

The remainder of the paper is arranged as follows: Section II provides a discussion on the existing studies done in this direction; Section III lists the details on our datasets, including dataset preparation and preprocessing using the automated denoising and annotation methods; Section IV compares the different methods on our datasets for stress detection; finally, Section V lists concluding remarks and future directions.

II. Background

The concept of stress detection of text data is emerging as an important subject in the literature, owing to its value in analyzing mental as well as physical health of the individuals. It is defined as the recognition of emotions which are closely related to the state of stress/depression, such as anxiety, confusion, annoyance, etc. This task has been deeply studied from the perspective of bio-signals using machine learning algorithms [13]. However, text-based data obtained from messaging services and/or social media serve as an equally important source for analyzing stress in individuals. Several methods have been devised for stress detection which use textual data as input. For example, Thelwall [14] proposed a lexical approach for stress analysis, where direct and indirect expressions for both stress and relaxation are detected by using a rule-based approach. The rules are defined separately for different categories and genres, which are detected using a set of keywords. The tweets are sourced from Twitter separately for each of the defined categories. However, this approach does not provide a robust and state-of-the-art benchmark performance as compared to solutions formulated using machine learning algorithms. Lin et al. [15] designed a three-level framework for stress detection from cross-media microblog data, by extracting middle-level representations based on psychological and art theories and employing a deep sparse neural network to learn the stress categories by incorporating the cross-media attributes. Lin et al. [16] proposed to extract discriminative hand-crafted statistical features and high-level semantic features, and used a hybrid model combining multi-task learning with convolutional neural network (CNN) to identify the stressor subjects and stressor events of the given social post to measure a users stress level from his/her social media data. It is evident that the dataset preparation in previous methods is done quite rudimentarily, and requires human intervention in some way or the other. For example, Lin et al. [16] uses Sina Weibo to construct a stress classification model, in which keyword patterns belonging to different stress categories were determined using a rule-based approach. This was followed by manual annotation of data using the extracted keyword patterns. Turcan and McKeown [17] constructed a dataset from Reddit, named Dreadit, for the purpose of stress classification. A number of stress related subreddits are used for building the corpus. However, the annotation of data is done manually as such, and no automated methods are applied for the same. In a similar direction, Mauriello et al. [18] constructed a dataset for detecting stressor categories from SMS-like messages. In this work also, data annotation was done in a manual fashion by using consolidated stressor categories derived manually from the Holmes and Rahe stress scale.

Stress detection on a textual dataset can be performed by a number of algorithms, which may be studied under three basic heads - lexicon-based methods [19], embedding-based methods [20], [21], and PLMs-based methods [22]. The lexicon-based methods are highly efficient and simple to implement, however, they feature inferior classification performance [23]. In comparison, embedding-based methods which employ the use of a shallow deep neural network (DNN) feature better performance and higher generalization performance. The best metrics are given by the PLMs, which are pre-trained in large-scale corpus and generally beneficial for downstream natural language processing tasks [24], [25]. PLMs can form large scale dependencies over the input text to comprehensively understand the underlying sentiment and return highly accurate predictions.

Our work alleviates the shortcomings of the above mentioned works on stress detection, especially with regards to dataset construction. We a transfer learning based approach to annotate data examples automatically, without any manual intervention. The use of Transformer-based models [26] for annotation not only provides robustness to our annotation strategy, but also yields better results when compared to rigid rule based processes as used in the above works.

III. Dataset Construction

In this section, we first introduce the overview of our datasets, then detail the process of dataset construction, including data collection, data preprocessing and data annotation. Figure 1 shows the work flow for dataset construction.

A. *Overview of our datasets*

We construct four datasets in total, two for each social media platform (Reddit and Twitter). Each of these datasets have unique attributes which demand their separate analysis. The labels for each of these datasets are binary in nature, where a value of “0” denotes stress negative examples, and a value of “1” denotes stress positive examples. Table I reports the statistical attributes about these datasets. The details of each dataset is given as following:

1) **Reddit Title**: This dataset consists of the titles of the posts collected from both stress and non-stress related subreddits. This dataset can be seen as analogous to the Twitter dataset, as it is marked by small text lengths. It thoroughly balanced for the different predictive classes,

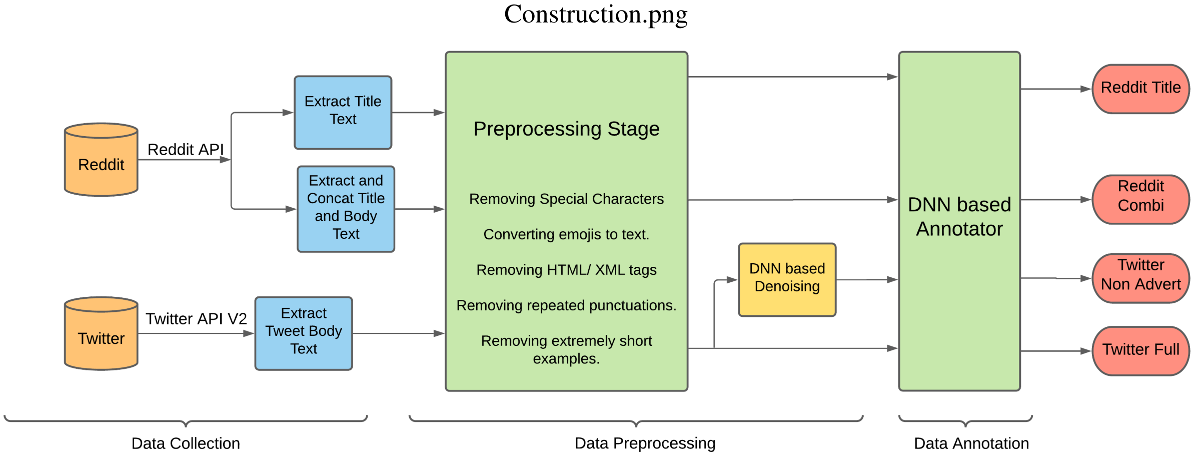

Fig. 1. Flow diagram of the dataset construction process. There are mainly three stages, i.e., data collection, data processing, and data annotation.

which entails the generation of high accuracy models with minimal bias towards a certain class. The examples are represented from a period staring from September 2019 to September 2021.

2) **Reddit Combi**: This dataset concatenates both the title and the body text of the posts extracted from subreddits related to stress, as well as from subreddits entailing a positive emotion. It entails examples having a long text length, and are in general more descriptive than the Reddit Title dataset. This dataset is however unbalanced in nature, with Stress Negative articles being in minority as compared to Stress Positive examples. The reason for this is that the Stress Negative articles were sourced from subreddits relating to happiness and other positive emotions. Thus, given the nature of these subreddits, most of the articles comprised of non-text data such as images, GIFs, videos, etc. Hence, this dataset can be used as a benchmark tool for approaches defined either for the Reddit Title dataset, or for testing models which can effectively capture the underlying understanding of extremely long texts.
3) **Twitter Full**: The Twitter Full dataset comprises of tweets taken from both stress and non-stress related hashtags. The tweets extracted belong to a period between September 2019 to September 2021. This dataset has not been processed for noise removal, which include advertisements and marketing data. Hence, this dataset is a better representation of the real-world data.
4) **Twitter Non-Advert**: This dataset has been derived from the Twitter Full dataset, by using a PLM-based denoising method for removal of advertisement and other irrelevant noisy examples. Though not an accurate representation of the real world data, it provides for relatively clean examples for accurate analysis of methods defined for stress classification from textual data. The denoising technique is discussed in Section III-C.

TABLE I
STATISTICS ABOUT THE CONSTRUCTED DATASETS. # DENOTES NUMBER.

Platform	Dataset	Avg.# char.	# Vocab.	# Positives	# Negatives
Reddit	**Title**	93	13706	2745	2811
	Combi	908	33764	2745	859
Twitter	**Full**	182	39465	4534	4366
	Non-Advert	177	14283	1268	783

B. Data collection

The examples for the four datasets are scraped from Twitter and Reddit using their respective APIs[2]. The examples collected belong to a period between September 2019 and September 2021. Hence, our datasets in principle effectively capture the general trends of stress induced by the Covid-19 pandemic across the world. We separately list out the methodologies followed for data collection from the two sources.

1) **Reddit**: For building the Reddit-based datasets, the Reddit API is used via PRAW wrapper classes. For collecting the Stress Positive examples, the subreddits[3] r\Stressed, r\Stress, r\Depressing, r\Depression and r\MentalHealth are used, whereas for Stress Negative examples, r\Happy, r\MadeMeSmile, r\MakeMeSmile and r\Wholesome are used. The examples are collected from each of the subreddits by sorting the examples by New, Hot, Top and Rising. From each of the scraped examples, the title and body text were extracted and two separate corpus were created - Reddit Title having only the title text, and Reddit Combi having the title and body text concatenated together.
2) **Twitter**: The Twitter corpus was constructed by using Twitter API V2. The tweets were collected between the period of September 2019 and September 2021. Stress Positive examples were sourced from

[2] Twitter API V2; Reddit API

[3] Subreddits are Reddit-specific equivalent of subforums, and are preceded by the 'r\' sequence.

the tweets containing the hashtags #Stress, #Stressed, #Tired, #FeelingUseless, #MentalHealthMatters, #MentalHealth, #FeelingStressed, #IamStressed, #Fatigued and #PandemicBlues, whereas the Stress Negative examples were sourced from tweets containing the hashtags #Happiness, #Happy, #Delighted, #Joy and #Blessed. From each of the scraped examples, only the tweet body text was extracted for creating the corpus.

It must be noted that the characteristics of the text data collected from Twitter and Reddit vary substantially from each other. While the examples collected from Reddit were descriptive, lengthy, and had a proper semantic structure, the examples collected from Twitter were short, abrupt, and in general lacked a proper grammatical structure.

C. Data Preprocessing

The data preprocessing stage is kept common for both the Reddit and the Twitter corpus. In this stage, all HTML tags, links, URLs, phone numbers and special unicode characters are removed. Also removed are repeated occurrences of punctuations and whitespaces. The emojis are converted to their text equivalent by using the Emojify library. Furthermore, extremely short examples containing less than ten characters are pruned. For the Twitter corpus, twitter specific symbols such as RT, QT, # are also removed. However, we did not remove the hashtag text from the examples owing to the fact that they also form an integral part of the overall content of the said example. Additionally, we use a PLM-based denoising approach wherein we finetune a DistilBERT model [27], which is obtained by the distillation of the base BERT model [22].

In the distillation process, the base BERT model is used to train a smaller model via the student-teacher training approach [28]. The end result is a much more lighter and efficient model which retains much of the characteristics of the base (teacher) model. The dataset used for fine-tuning the DistilBERT model [29] comprises of two separate corpus containing the clickbait/advertisement data and the clean data separately. This finetuned model is then applied to the existing Twitter corpus for removing noisy examples, leading to the creation of Twitter Non Advert dataset as defined above. This denoising method was not used for the Reddit corpus since on manual inspection it was revealed that the Reddit data was relatively free from noisy examples. Furthermore, the preprocessing methodology described above removed most of the few noisy examples that were contained in Reddit.

D. Data Annotation

In the data annotation stage, we test out three PLM models trained on the task of emotion classification for automating this process. Given the fact that stress is often a culmination of one or more base emotions, we aim to use the emotion classes closely related to stress for labelling the examples in our corpus as Stress Positives, i.e., all examples which are classified with the either of the emotion classes related to stress are labelled as Stress Positives.

The information about the PLM models along with emotion classes used for identifying Stress Positive examples is listed as follows:

1) **BERT trained on the Twitter Emotion dataset (Model A)**. This model uses the pretrained BERT-base-uncased checkpoint which is finetuned on the Twitter Emotion dataset [30]. The Twitter Emotion dataset contains tweets labelled with six basic emotions - sadness, joy, love, anger, fear and surprise. For the purpose of stress annotation, the classes belonging to the emotions - sadness, anger and fear are taken to be Stress Positives, while the other emotion classes are considered to be Stress Negatives.
2) **RoBERTa trained on the Twitter Emotion dataset (Model B)**. This model uses the RoBERTa-base [31] checkpoint finetuned on the Twitter Emotion dataset. The annotation methodology remains the same as in Model A, since the dataset used for fine-tuning this model is the same as in Model A.
3) **RoBERTa trained on the GoEmotions dataset (Model C)**. Using the pretrained RoBERTa-base model as the base, this model uses the GoEmotions dataset for the purpose of fine-tuning. The GoEmotions dataset [32] is a manually annotated corpus of English comments sourced from Reddit. Each example is labelled with 27 emotion categories. In our data annotation task, we use the emotion classes - anger, annoyance, confusion, disappointment, disapproval, disgust, embarrassment, fear, grief, nervousness, sadness and remorse for labelling the examples in our datasets as Stress Positives, whereas examples belonging to the remaining emotion classes are labelled as Stress Negatives.

For the purpose of validating the performance of the three models listed above, we create gold datasets separately for the datasets built from Twitter and Reddit. These gold datasets consist of high quality examples which are annotated manually. For the Reddit corpus, the size of the gold dataset is 234 examples, while the Twitter corpus contains 120 examples. Table II lists the metrics obtained for the aforementioned three models on these gold datasets. It is evident from the metrics in Table II that Model B, which represents RoBERTa-base model finetuned on Twitter Emotions dataset, provides the best performance among all the three models. However, the annotation performance on the Reddit corpus is inferior to the Twitter corpus.

To improve the performance of Model B, we use the thresholding strategy in which we define specific limits for the confidence level of the dominant emotion class predicted by the models. Only if the dominant class is predicted with a confidence level above the threshold is when the corresponding example is labelled as Stress Positive. We used three thresholding levels at 90%, 80% and 50% of the confidence level to test out our approach, and we found that using the thresholding level of 90% yields the maximum improvement in the annotation performance of Model B.

TABLE II
PERFORMANCE OF DIFFERENT PLMS USED FOR ANNOTATION.

Dataset	Measure	Model A	Model B	Model C
Reddit Title	Accuracy	78.11	77.25	78.11
	F1	80.00	79.05	78.84
Reddit Body	Accuracy	75.11	78.54	79.82
	F1	78.52	80.77	79.83
Twitter	Accuracy	90.00	91.67	85.00
	F1	91.04	92.65	85.71

TABLE III
USING THRESHOLDING FOR ROBERTA MODEL TRAINED ON TWITTER EMOTIONS DATASET (MODEL B)

Dataset	Measure	Threshold		
		>50%	>80%	>90%
Reddit Title	Accuracy	77.25	80.69	80.26
	F1	79.05	83.87	84.24
Reddit Body	Accuracy	78.11	78.97	78.54
	F1	80.60	82.81	82.88
Twitter	Accuracy	91.67	93.33	93.33
	F1	92.65	94.29	94.37

The metrics obtained under the three confidence levels are provided in Table III. Though the performance of our approach obtained on the gold dataset is remarkable for the Twitter corpus, there is a decrease in the metrics obtained for the Reddit corpus. On inquiring into the plausible reasons for this performance reduction, we made a number of observations with regards to the nature of data.

For the Reddit Title gold dataset, we found that the length of the title text was quite small, especially when compared to the Twitter gold dataset. This observation is confirmed from Table I, where the average example length in the Twitter dataset is almost double than that of the Reddit Title dataset. Hence, it is possible that the Transformer-based models used for annotation are not being able to capture the underlying emotion of the text effectively. On the other hand, the Reddit Combi gold dataset suffers from a similar issue. The length of the text in the gold examples was found to be very large, which is responsible for the reduced performance of our approach. This is due to the fact that the Transformer - based models are hugely bottlenecked by their memory and size limitations, which prevents them from practically processing long texts with the same efficiency.

However, given that our approach is totally automated and requires no manual intervention in any form, the results yielded on the gold datasets are quite satisfactory for practical purposes. Our approach overcomes the hassles and issues involved with manual annotation, and effectively eliminates the presence of the labelling bias. The labelling bias, which is often a cause of concern due to the variations in manual annotation, is eliminated via this approach since a single model is used for annotating the entire dataset. Thus, based on the results obtained with different models and thresholding limits, we choose Model B for automated annotation our four datasets, with a threshold limit of 90%. Additionally, it is to be noted from Table I that the size of Reddit Combi and Reddit Non-Advert dataset is lower than the other two datasets. Nonetheless, this does not lead to any impact in the quality of our automated annotation process.

TABLE IV
COMPARISON WITH OTHER DATASETS. *Lan.* DENOTES LANGUAGE, I.E., ZH (CHINESE) AND EN (ENGLISH).

Dataset	Source	Lan.	Size	Balanced	Annotation
Lin et al. [15]	Sina Weibo	zh	57785	Partially	Rule-based
Lin et al. [16]	Sina Weibo	zh	2000	No	Manual
Thelwall [14]	Twitter	en	3066	No	Manual
Turcan et al. [17]	Reddit	en	3553	Yes	Manual
Ours					
Reddit Title	Reddit	en	5556	Yes	Automated
Reddit Combi	Reddit	en	3604	No	Automated
Twitter Full	Twitter	en	8900	Yes	Automated
Twitter Non-Advert	Twitter	en	2051	Partially	Automated

E. Data Verification

To verify the quality of annotation performed by our automated method, we randomly sampled 200 examples from the entire corpus for each of the datasets and employed 5 manual (human) annotators to cross-verify the same. On an average, the labelling accuracy was reported to be around 94% across the four datasets, which substantiates the high quality of annotation provided by our method.

Table IV compares our datasets with previous works. While our datasets are the first ones to use an automated annotation strategy, Reddit Title and Twitter Full in particular provide a large labelled balanced dataset for stress detection compared to the existing datasets. Furthermore, our pre-processing strategy is relatively more sophisticated, which has been discussed at length in the above sections.

IV. EXPERIMENTS

We compare the performance of different methods in our datasets for stress detection and explore the possibility of the applicability of the classifiers trained on sentiment analysis.

A. Baselines and Experiment Setup

We test the performance of several lexicon, embedding and PLM-based classifiers on our datasets.

a) ***Lexicon-based methods:*** They are based on statistical rules to predict the emotion of the input text. For the purpose of stress detection, we treat the "Neutral" and "Negative" emotions as Stress Positives, whereas the "Positive" emotion is accorded a Stress Negative label. Since the lexical-based methods do not require any training and can be used "off-the-shelf", the metrics are obtained by taking the datasets in their entirety; no splits are considered. The used lexicons are as following:

- **SenticNet** [19] is a commonsense knowledge base for sentiment analysis which leverages neurosymbolic AI to form the dependency relations within the different parts of the input text for modeling the sentiment.

TABLE V

ACCURACY AND F1 SCORE FOR DIFFERENT MODELS ON THE TWITTER AND REDDIT DATASETS. PLM DENOTES PRE-TRAINED LANGUAGE MODEL.

S.N.	**Type**	**Method**	**Reddit**				**Twitter**			
			Title		**Combi**		**Full**		**Non-Advert**	
			Acc.	**F1**	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**
1	Lexicon-based Methods	SenticNet [19]	78.99	79.76	80.02	78.76	77.46	79.51	79.33	80.94
2		SO-CAL [33]	75.25	77.25	84.05	77.25	72.71	69.55	68.06	70.51
3		AFFIN [34]	73.74	75.04	75.80	75.04	77.32	74.31	78.22	80.14
4		TensiStrength [14]	71.04	76.30	81.02	76.30	73.87	80.12	73.62	76.21
5		VADER [35]	77.57	77.92	74.50	77.92	75.09	70.63	74.16	75.71
6		SentiWordNet [36]	64.15	64.03	67.62	64.03	61.22	53.91	53.10	50.57
7	Embedding-based Methods	FastText [37]	94.06	93.92	94.04	93.92	84.33	85.52	84.40	88.28
8		Glove [38]	90.10	89.80	89.46	89.80	76.12	76.51	75.85	81.00
9		Word2Vec [39]	91.81	91.60	88.07	91.60	78.43	79.33	80.00	84.70
10	PLM-based Methods	RoBERTa [31]	93.76	96.52	94.72	96.52	**91.13**	**91.11**	**88.32**	**90.12**
11		DistilBERT (w/o Summarization) [27]	**98.20**	**98.15**	**97.64**	**98.15**	87.36	88.25	86.83	89.20
12		DistilBERT (w/ Summarization)	-	-	92.65	95.06	-	-	-	-

- **SO-Cal** [33] (Semantic Orientation Calculator) is a tool for extracting the underlying sentiment from the text and according it a polarity on the basis of the opinion expressed by it towards the main subject matter.
- **Afinn** [34] uses affective word list and sentiment lexicons for scoring each word in the input text for valence, and provides a net score which represents the nature of the polarity present in it.
- **TensiStrength** [14] is built upon the target of identifying the strength of stress and relaxation from short texts. It uses two independent ranking parameters - Relaxation and Stress, and provides a score for each of them.
- **VADER** [35] uses a list of lexical features created via a combination of qualitative and quantitative methods specifically tuned towards identifying the sentiment in short social media texts.
- **SentiWordNet** [36] uses WordNet synsets to mark an input text using varying degrees of positivity, negativity and neutrality, which are sourced from Princeton's WordNet Gloss corpus.

*b) **Embedding-based methods**:* We represent input text using word embeddings and use a shallow neural network for stress classification. Only the two-layer neural networks are set as trainable. The model is trained using Adam optimizer and cross-entropy loss with an early-stopping callback to avoid overfitting. The four datasets constructed in this study are randomly split into training, validation and test splits with a percentage of 60%-20%-20%. The metrics reported in our study are obtained from the test splits of the datasets. We employ the following pre-trained word embeddings:

- **FastText** [37] leverages character n-grams to generalize well to the rare words and the out-of-vocabulary words. We use the 100-dimensional version in our experiments.
- **GloVe** [38] uses both global and local information to learn word embeddings. We use the 300-dimensional embeddings trained on the Common Crawl corpus (glove.840B.300d) for two Reddit datasets, and use the 200-dimensional embeddings trained on 2 billion tweets (glove.twitter.27B) for two Twitter-based datasets.
- **Word2Vec** [39] is one of most popular word embedding methods. We use its 300-dimensional CBOW version which is trained on Google News.

*c) **PLM-based methods**:* Recently, pre-trained language models have achieved significant achievements in text classification and sentiment analysis. Similarly, the datasets are split into training, validation and tests sets in ratios of 60-20-20. The PLM models are finetuned with a fully connected classification layer using the training set. The metrics obtained are representative of the test set. We use the following PLMs:

- **RoBERTa** [31] is the robustly optimized version of the base BERT model. It achieved state-of-the-art performance in various tasks.
- **DistilBERT** [27]: DistilBERT is a lightweight version of the base BERT model, with a substantial reduction in the model complexity and number of trainable parameters.

They are finetuned using the HuggingFace library [40], and all the parameters during the model training process have been kept in their default setting.

B. Results

Table V reports the performance of different methods on four datasets in terms of F1 scores and accuracy. It can be noticed that the lexicon-based methods achieve inferior performance with regards to embedding-based methods and PLM-based methods. This is because lexicon-based methods mainly rely on word-level sentiment polarity, and cannot capture the underlying semantic information of the input text. Barring SenticNet in Table V, all other methods are based upon polarity detection and classification, which are accompanied with inferior results. Hence, this indicates that stress detection is not synonymous with sentiment polarity detection; in fact stress may not always necessarily carry a negative emotion in context [41].

TABLE VI
Examples taken from the Reddit and Twitter datasets. A label of 1 denotes Stress Positive examples, and a label of 0 denotes Stress Negative examples

Platform	Text	Label
Reddit	My worst enemy is myself when it comes to stress Yes, I’m stuck in a job I hate and I don’t have the option of leaving at all but I am my own worst enemy here. I constantly stress about the small things and its really affecting my capacity to perform. The days feel like years, and I’m always at the verge of breaking after one or two days. Please teach me methods to overcome this. I tried reframing my perspective but in the moment I still stress.	1
	Anyone else have semi-suicidal thoughts on a regular basis? I’m not saying gun to my head thoughts. I mean like, driving down the road and once in a while my brain says “just turn into the opposite lane. It’ll look like an accident” but then I just shake it off. Or shaving my legs and think ”just cut, bleed out, it’s so easy” but I never actually act on these things. They’ve just become a part of my life. I feel like my depression is almost personified in a weird cloud that follows me around with bad intentions that I have to constantly argue with. It’s exhausting.	1
	I just squatted 305lbs for the first time! I have a personal goal of squatting 315 lbs for six reps. Two months ago I could only do 195lbs. Tonight I just did 305! AAAAAAAAAA I’m so excited!!!! I just needed to share this with more people but I am STOKED!! Just 10lbs from my goal! I’m going to go for it Thursday!	0
	Finally got all my certifications to take a step forward in becoming a local94 engineer! I never would’ve thought I could be able to do something like this and I really can’t believe I was able to get so close to becoming an engineer	0
Twitter	When your 4 year old suddenly has fever for days and tummy trouble. stress!	1
	Lying in bed, trying to sleep. Brain riddled with anxieties about my kids, my marriage, myemployment situation. Bed time does not bring comfort to me. stressed Anxiety trials life can be hard.	1
	When that smile is bright cuz you have a man in your life that brightens your day happy smile	0
	Hope this makes someone smile like it did me. Happy	0

TABLE VII
Results using different summarizer models for DistilBERT on Reddit Combi dataset

Methods	**T5**	**BART**	**Pegasus**
Accuracy	92.65	91.12	84.88
F1	95.06	93.74	88.91

The performance of embedding-based methods is satisfactory, with the model based on FastText embeddings leading the models based on GloVe and Word2Vec. The probable reason is that FastText can alleviate the out-of-vocabulary problem to some extent. However, an innate disadvantage of using embedding-based models is that they cannot capture long term dependencies within the text, which leads to reduction in overall metrics when compared with the PLM-based methods.

The best performance metrics are given by PLM-based models. Even amongst them, DistilBERT is able to provide better metrics than RoBERTa for the Reddit dataset. This can be alluded to the fact that RoBERTa is quite a large model in terms of parameters and complexity, and hence it requires a much larger corpus to be comprehensively finetuned to yield state-of-the-art results. Owing to this reason, RoBERTa is able to outperform DistilBERT on the Twitter datasets.

As an addition, we try to improve the performance of the DistilBERT model on the Reddit Combi dataset by using abstractive summarization, which captures the entire content of the input text and represents it in a concise format. To this end, we apply three well Transformer based models trained on the summarization task on the Reddit Combi dataset, since it is comprised of examples which have a large character length. From Table VII, it can be seen that T5 [42] summarization provides better metrics than BART [43] and Pegasus [44].

However, we were not able to match the performance metrics obtained without using the summarization model. This proves that that summarizer models were not able to capture the underlying context of the sentences with a high precision. It is observed that the performance of all the methods is lower on the Twitter datasets when compared to the Reddit ones. This disparity can be pointed out to the different patterns in writing styles followed in both the platforms. We list out some examples collected from Reddit and Twitter in table VI.

It showed that the articles posted on Reddit tend to be more comprehensive and have a better semantic structure, with a proper flow of ideas. On the other hand, articles on Twitter tend to be short, cryptic and intermixed with emojis and hashtags, and they have a random flow of ideas within the entire text. Thus, Reddit provides better quality text articles as compared to Twitter. In fact, this reason can also substantiate the fact that the performance gains obtained by using DistilBERT or RoBERTa are comparatively less for the two Twitter datasets when compared to the other methods.

V. Conclusion and Future Works

The purpose of this study is to initiate significant studies to detect stress levels in individuals from their activities in the digital world. We construct four datasets from Reddit and Twitter for stress detection, and automate the annotation process by using a Transformer-based model via transfer learning, without any human intervention. Furthermore, we conducted a comparative study on these datasets by using several methods based on lexicons, embeddings, and PLMs. In the future, we will construct databases and design models for fine-grained stress detection task.

References

[1] Z. Wang, S. Ho, and E. Cambria, "A review of emotion sensing: Categorization models and algorithms," *Multimedia Tools and Applications*, vol. 79, pp. 35 553–35 582, 2020.

[2] M. Grassi, E. Cambria, A. Hussain, and F. Piazza, "Sentic web: A new paradigm for managing social media affective information," *Cognitive Computation*, vol. 3, no. 3, pp. 480–489, 2011.

[3] Y. Susanto, A. Livingstone, B. C. Ng, and E. Cambria, "The hourglass model revisited," *IEEE Intelligent Systems*, vol. 35, no. 5, pp. 96–102, 2020.

[4] H. Lin, J. Jia, J. Qiu, Y. Zhang, G. Shen, L. Xie, J. Tang, L. Feng, and T. Chua, "Detecting stress based on social interactions in social networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 9, pp. 1820–1833, 2017.

[5] H. Selye, "What is stress," *Metabolism*, vol. 5, no. 5, pp. 525–530, 1956.

[6] G. Napoli, "Stress and depressive symptoms among italian mental health nurses during the covid-19 pandemic, a cross-sectional study," *Archives of Psychiatric Nursing*, 2021.

[7] S. Ji, S. Pan, X. Li, E. Cambria, G. Long, and Z. Huang, "Suicidal ideation detection: A review of machine learning methods and applications," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 1, pp. 214–226, 2021.

[8] S. Ji, X. Li, Z. Huang, and E. Cambria, "Suicidal ideation and mental disorder detection with attentive relation networks," *Neural Computing and Applications*, vol. 34, 2022.

[9] G. Giannakakis, D. Grigoriadis, K. Giannakaki, O. Simantiraki, A. Roniotis, and M. Tsiknakis, "Review on psychological stress detection using biosignals," *IEEE Transactions on Affective Computing*, 2019.

[10] J. Zhai and A. Barreto, "Stress detection in computer users based on digital signal processing of noninvasive physiological variables," in *Proceedings of International Conference of the IEEE engineering in medicine and biology society (EMBC)*, 2006, pp. 1355–1358.

[11] C. Barrie and A. Frey, "Faces in the crowd: Twitter as alternative to protest surveys," *PloS one*, vol. 16, no. 11, 2021.

[12] S. C. Finlay, "Age and gender in reddit commenting and success," *Journal of Information Science Theory and Practice*, vol. 2, no. 3, pp. 18–28, 2014.

[13] S. S. Panicker and P. Gayathri, "A survey of machine learning techniques in physiology based mental stress detection systems," *Biocybernetics and Biomedical Engineering*, vol. 39, no. 2, pp. 444–469, 2019.

[14] M. Thelwall, "Tensistrength: Stress and relaxation magnitude detection for social media texts," *Information Processing & Management*, vol. 53, no. 1, pp. 106–121, 2017.

[15] H. Lin, J. Jia, Q. Guo, Y. Xue, J. Huang, L. Cai, and L. Feng, "Psychological stress detection from cross-media microblog data using deep sparse neural network," in *Proceedings of the 2014 IEEE International Conference on Multimedia and Expo (ICME)*, 2014, pp. 1–6.

[16] H. Lin, J. Jia, L. Nie, G. Shen, and T.-S. Chua, "What does social media say about your stress?" in *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, 2016, pp. 3775–3781.

[17] E. Turcan and K. McKeown, "Dreaddit: A reddit dataset for stress analysis in social media," in *Proceedings of the 2019 International Workshop on Health Text Mining and Information Analysis LOUHI@EMNLP*, 2019, pp. 97–107.

[18] M. L. Mauriello, T. Lincoln, G. Hon, D. Simon, D. Jurafsky, and P. Paredes, *SAD: A Stress Annotated Dataset for Recognizing Everyday Stressors in SMS-like Conversational Systems*, 2021.

[19] E. Cambria, Q. Liu, S. Decherchi, F. Xing, and K. Kwok, "Senticnet 7: A commonsense-based neurosymbolic ai framework for explainable sentiment analysis," in *Proceedings of the Language Resources and Evaluation Conference, LREC*, 2022.

[20] Q. Liu, H. Huang, Y. Gao, X. Wei, Y. Tian, and L. Liu, "Task-oriented word embedding for text classification," in *Proceedings of the International Conference on Computational Linguistics (COLING)*, 2018, pp. 2023–2032.

[21] Q. Liu, J. Lu, G. Zhang, T. Shen, Z. Zhang, and H. Huang, "Domain-specific meta-embedding with latent semantic structures," *Information Sciences*, vol. 555, pp. 410–423, 2021.

[23] V. Bonta and N. K. N. Janardhan, "A comprehensive study on lexicon based approaches for sentiment analysis," *Asian Journal of Computer Science and Technology*, vol. 8, no. S2, pp. 1–6, 2019.

[22] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, 2019, pp. 4171–4186.

[24] M. Ge, R. Mao, and E. Cambria, "Explainable metaphor identification inspired by conceptual metaphor theory," in *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI) Conference on Artificial Intelligence*, 2022.

[25] Q. Liu, X. Geng, H. Huang, T. Qin, J. Lu, and D. Jiang, "Mgrc: An end-to-end multigranularity reading comprehension model for question answering," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of Annual Conference on Neural Information Processing Systems*, 2017, pp. 5998–6008.

[27] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," 2020.

[28] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015.

[29] A. Chakraborty, B. Paranjape, S. Kakarla, and N. Ganguly, "Stop clickbait: Detecting and preventing clickbaits in online news media," in *Proceedings of the IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)*, 2016, pp. 9–16.

[30] E. Saravia, H.-C. T. Liu, Y.-H. Huang, J. Wu, and Y.-S. Chen, "CARER: Contextualized affect representations for emotion recognition," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018, pp. 3687–3697.

[31] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019.

[32] D. Demszky, D. Movshovitz-Attias, J. Ko, A. S. Cowen, G. Nemade, and S. Ravi, "Goemotions: A dataset of fine-grained emotions," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics, ACL*, 2020, pp. 4040–4054.

[33] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Computational linguistics*, vol. 37, no. 2, pp. 267–307, 2011.

[34] F. Å. Nielsen, "A new anew: Evaluation of a word list for sentiment analysis in microblogs," vol. 718, pp. 93–98, 2011.

[35] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the International Association for the Advancement of Artificial Intelligence (AAAI) Conference on Web and Social Media*, vol. 8, no. 1, 2014.

[36] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining."

[37] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," pp. 427–431, 2017.

[38] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.

[39] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proceedings of Annual Conference on Neural Information Processing Systems*, 2013, pp. 3111–3119.

[40] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Huggingface's transformers: State-of-the-art natural language processing," 2020.

[41] S. Folkman, "The case for positive emotions in the stress process," *Anxiety, stress, and coping*, vol. 21, no. 1, pp. 3–14, 2008.

[42] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," pp. 140:1–140:67, 2020.

[43] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," pp. 7871–7880, 2020.

[44] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization," pp. 11 328–11 339, 2020.