Educational Administration: Theory and Practice
2024, 30(1), 3252-3263
ISSN: 2148-2403
https://kuey.net/

Research Article

On Text Datasets For Stress Detection: A Comprehensive Analysis And Future Ideas

Kalyani Pendke[1*], Charu Goel[2], Mayuri Digalwar[3]

[1*]Indian Institute of Information Technology, Nagpur, Maharashtra, India. dtea20cse004@iiitn.ac.in [0000000194039720]
[2]Indian Institute of Information Technology, Nagpur, Maharashtra, India. charugoel@iiitn.ac.in [0000000223393551]
[3]Indian Institute of Information Technology, Nagpur, Maharashtra, India. mayuri@iiitn.ac.in [0000000211628085]

***Corresponding Author:** Kalyani Pendke
*Indian Institute of Information Technology, Nagpur, Maharashtra, India. dtea20cse004@iiitn.ac.in [0000000194039720]

Citation*:* Kalyani Pendke et al. (2024), On Text Datasets For Stress Detection: A Comprehensive Analysis And Future Ideas, *Educational Administration: Theory and Practice*, *30*(1), 3252-3263
Doi: 10.53555/kuey.v30i1.3884

ARTICLE INFO

ABSTRACT

Human health depends on the early identification of psychological stress. Stress can be detected by analyzing the text written by humans. Currently, a limited number of text datasets exist for stress detection. This article gives a detailed analysis of the base papers of ten text datasets used for emotion and stress detection and will be helpful to researchers in selecting the best datasets for their research work. Datasets chosen for this analysis are Dreaddit, TensiStrength, SMHD, VENT, GoEmotion, ISEAR, EmoInt, EmoBank, and TEC. Data annotation is an important task in the preparation of a text dataset. In this article four techniques for getting an inter-rater agreement are discussed, namely Pearson Correlation Coefficient, Spearman Correlation Coefficient, Krippendorff's alpha coefficient, and Fleiss' Kappa. After doing a detailed analysis of the existing datasets, the Dreaddit dataset seems to be the best option for research work on stress detection. Also, the Pearson correlation Coefficient technique provides the best results for getting an inter-rater agreement. This article concludes with a discussion of several unresolved challenges and potential research directions for text-based identification of stress.

Keywords: Emotion Detection · Stress · NLP · BERT · Sentiment Analysis · Affective computing.

1 Introduction

Most of us experience Psychological Stress in our day-to-day lives. It is of two types. The first one, called good stress, also referred to as "Eustress", inspires a person to accomplish his or her targets. The second one, called negative stress, commonly referred to as "Distress", demotivates a person to finish their responsibilities [1]. Psychological stress adversely affects human health. Hence, early identification of psychological stress is very crucial. For the rest of this paper, we will be referring to "Psychological Stress" as "Stress". People under stress experience certain emotions such as anger, sadness, and embarrassment [2]. Therefore, it is very important to recognize the emotions causing stress in humans. To detect stress through text data, knowledge of psychological stress and Natural Language Processing (NLP) is a must.

The human body gives certain responses to difficult tasks or problems. We can call these responses as "stress". "Stressor" is one of the important concepts related to stress, which is responsible for causing stress in an individual. A stressor can be a human being, position, or circumstance [4]. Negative emotions in humans are very crucial. They are crucial in the bidirectional relationship of stress with disease [2]. Some of the negative emotions, like anger, sadness, and embarrassment, can cause diseases in humans. Therefore, it's important to recognize the negative emotions that lead to stress in people [5].

Nowadays, people express themselves through social networking platforms. A huge amount of data from social networking sites is available, which can be used to identify stress in humans. Many people have mental health issues these days. Transcripts and other clinical notes from interviews with such people are available. Proper analysis of such a clinical text is an urgent need of the hour. Artificial Intelligence (AI) has several branches, one of which is NLP, which is widely used to perform a detailed analysis of text from social networking websites and clinical data of people with mental health issues. NLP approaches are important for sentiment analysis, information extraction, emotion recognition, and mental health monitoring. [6]. NLP techniques can also be used for stress detection using text data.

This article surveys previous research done in the area of stress recognition using text data. A major concern in text data analysis for stress detection is the availability of datasets. Very few text datasets exist for stress detection. The researchers either use an existing dataset or create their own dataset. In both of these cases, the detailed knowledge of the dataset includes the process of gathering text data, annotating the data, and utilizing a range of deep learning (DL) and machine learning (ML) methods for dataset validation, etc. The present article throws light on the various existing text datasets, whose base papers have been analyzed in detail. For researchers who want to create their own dataset, the main hurdle is annotating the text data. Hence, we also included the four techniques for getting inter-rater agreement. A detailed analysis of the existing text dataset and knowledge of techniques of inter-rater agreement will enable other researchers to choose the most appropriate dataset for their investigation.

To the best of our understanding, only two text datasets are available for stress detection: one is Dreaddit and the other is TensiStrength. In this article, we have selected these two datasets for stress detection, and the remaining eight datasets, i.e., SMHD, VENT, GoEmotion, ISEAR, EmoInt, EmoBank, and TEC, are used for emotion detection, which can further be utilized for stress detection. When humans are stressed, they experience negative emotions like anger, fear, and sadness [7]. And when they are not stressed, they experience emotions like joy and happiness. Stress and emotions are intertwined. By considering this relationship, one can prepare a new text dataset by selecting the text related with thesefeelings [5]. Therefore, it is critical to examine the widely accessible text dataset in order to detect emotions.

Scholars working on text-based stress recognition have done a survey of existing techniques for detection of stress. However, to the best of our understanding, a survey of the base papers of the text dataset has never been carried out previously, which highlights the novelty of the work done in this research article.

Stress detection through text is an upcoming research area in which many researchers are working. One of the initial problem they usually face is proper selection of the dataset. This article will help them to know the various aspects of the existing text datasets and various techniques for getting inter-rater agreement. The methodology adapted for the selection of papers for this survey is as follows: Articles from numerous scientific databases were explored in detail. The databases include Springer, PubMed, IEEE Xplore, and other reputed journals. The keywords used for searching papers include stress, anxiety, depression, text datasets, NLP, etc. The selection criteria for shortlisting the papers include stress detection through textual data and the methodology mentioned in the paper.

The current paper is structured in seven sections. An Introduction is covered in the first section. Second section discusses a detailed survey of existing datasets used for stress and emotion detection. The third section deals with the utilization of datasets, while the fourth describes techniques for inter-rater agreement. Linguistic Inquiry and Word Count (LIWC), the most widely used emotion feature extraction tool, is further examined in the fifth section. The sixth section discusses key research challenges in stress detection through text, and the last section concludes the paper.

Our review answers the following three questions:
(i) Which datasets are suitable for text analysis-based stressand emotion detection?
(ii) Which are the various techniques used for getting inter-rater agreement during the annotation of a text dataset through text analysis?
(iii) What are the key research challenges in stress detection through text?

2 Detailed Analysis of Existing Datasets Used for Stress and Emotion Detection

This segment summarizes the information about current datasets for text databased stress and emotion detection, by giving a detailed analysis of the base papers for these datasets. The parameters considered are details about datasets, process for annotation of the datasets, application of ML and DL techniques on datasets, future scope, and emotions considered in the datasets.

Figure 1 below describes the various sources available for downloading text datasets for stress and emotion detection from text.

2.1 Dreaddit Dataset

In 2019, Turcan et al. [8] developed the dataset called the “Dreaddit dataset”. They collected 190000 Reddit posts from five distinct Reddit sections namely abuse, anxiety, financial, PTSD, and social.

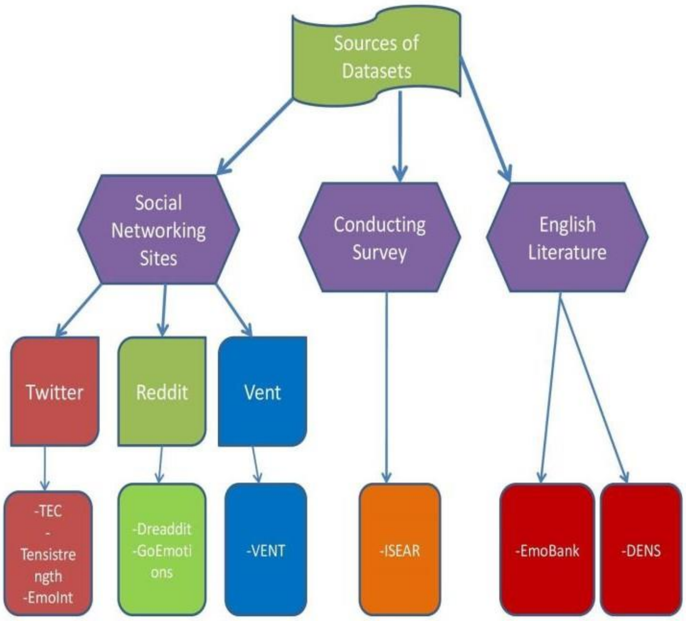

Fig.1: Sources of Datasets

They took the help of Amazon Mechanical Turk crowd workers to label the posts. The final dataset consisted of 3553 posts out of which 2838 posts are used for training posts while 715 posts are used for testing. Every post in the dataset is annotated by at least five raters. The labels for each segment were finalized by considering the highest vote of the raters. Raters' agreement was calculated using the Fleiss Kappa method [9], giving the value of k as 0.47 on all labeled data.

For the calculation of the results in [8], they used CNN, Gated Recurrent Neural Network (GRNN), N-Gram, LR, and Base version of BERT algorithms. Out of all these algorithms, the BERT-base model achieved the greatest values of 0.7518 for precision, 0.8699 for recall, and 0.8065 for F1-score. The emotion labels used for datasets were "stress" and "not stress". The author's future strategy calls for contextualizing stress and providing justification using the contextual features of the text.

2.2 TensiStrength Tool's Dataset

In 2017, Thelwall et al. [10] created a dataset for the "TensiStrength" tool. By analyzing tweets, this tool determined how much stress and relaxation were present in them. It improved an algorithm called "Sentistrength", which was previously used to measure the degree of relaxation and stress in sentences [11]. TensiStrength employed a different set of rules and a technique that uses lexicons in order to recognize the presence of symptoms of stress and relaxation in the sentence. Tweets for the dataset were collected from different sources for one month, i.e., July 2015. These tweets were rated by three separate raters. Interrater reliability scores were calculated using the Krippendorff score [12] technique. Algorithms used were J48 Tree, AdaBoost, JRip Rules, NB, Decision Table, LR, and SVM. The parameters usedfor comparison were Pearson's correlations and Mean Absolute Deviation (MAD). Two labels given by this toolwere "stressed" and "non-stressed".

2.3 SMHD Dataset

In 2018, Cohan et al. [13] discussed a new dataset called as "SMHD (Self-reported Mental Health Diagnoses) dataset" which consist of posts from Reddit website. The dataset contains user labels showing the mental health condition(s) linked with each user. Also, they had collected each user's Reddit posts for the duration of January 2006 and December 2017 (included). Mental health conditions considered in the paper were schizophrenia spectrum disorders (schizophrenia), anxiety disorders (anxiety), bipolar disorders (bipolar), obsessive-compulsive disorders (OCD), depressive disorders (depression), feeding and eating disorders (eating), autism spectrum disorders (autism), and disorders which are related to stress and trauma. Various algorithms such as LR, XGBoost, Linear SVM, Supervised Fast Text, and CNN were used to check the performance of the model. The highest precision of 100 percent is achieved by LR and SVM. The highest recall value of 57.95 percent was achieved by CNN. The Supervised Fast Text method produced a value of 57.60 for the F-score, which was the highest among other methods.

2.4 VENT Dataset

In 2021, Malko et al. [14] discussed the "VENT dataset". The dataset was prepared by using self-annotated posts (vents) from the VENT website. For the preparation of the dataset, the authors took snapshots of more than 100 million messages. All these snapshots were collected from the developers of VENT. For collected messages, the tags have been given by the original writer of the messages. The labels of the VENT dataset are aligned in two-level pecking order. There are eighty-five emotion categories at the most fundamental level, and the author has divided them into five groups: affective states, dates, groups of people, character/role/imaginary content, and miscellaneous. Out of these five groups, affective states contain nine categories, namely: creativity, affection, fear, anger, positivity, feelings, sadness, happiness, surprise. The Dates group contains forty-six categories related to seasonal events and dates, such as ramadan, paralympics, autumn, etc.

Groups of people consist of thirteen categories, including Pride'18 Women HM, etc. Imaginary/Character/Role content group holds seven categories associated to imaginary topics and fictional ones, such as Star Wars, Vampire. The miscellaneous group carries ten categories of various types, for example, gaming, candy, etc. Users can always access the nine categories connected to affective states while paying for all the remaining categories. After the fundamental level, the next level consists of 1187 labels. The authors have selected labels related to the main emotions including affection, fear, surprise, anger, happiness, sadness.

For the purpose of finding the relation between text and labels in the above mentioned "VENT dataset", the authors conducted four types of analysis, including qualitative analysis, vocabulary-based analysis, emoji based analysis, and text-to-label machine learning classifier analysis. For the last type of analysis, they have selected the BERT base model. The above-mentioned six emotion labels were used to calculate the performance parameters including recall, precision and F1-score. Out of all the emotion labels, the affection emotion label had the highest value for recall (0.65), precision (0.62), and F1-score (0.63). The authors discovered a strong correlation between the writers' self-assigned labels and the emotional states that were conveyed in the texts. The future scope discussed by the authors was to deal with the mixed emotions expressed in the message. The dataset in [14] is not publicly available due to ethical and privacy concerns.

2.5 GoEmotions Dataset

In 2020, Demszky et al. [15] developed a dataset with the name "GoEmotions". It includes 58,000 comments from Reddit website which was labeled as "neutral" or belonging to "27 different emotion classes".

The emotions considered in the above mentioned dataset were approval, amusement, admiration, anger, relief, annoyance, curiosity, desire, caring, disappointment, disapproval, disgust, confusion, excitement, embarrassment, gratitude, fear, love, grief, optimism, nervousness, sadness, surprise, pride, realization, remorse, and joy. This dataset was manually annotated. Three raters were assigned for each example. By using interrater correlation, the authors calculated rater agreement for each emotion. For every rater r belonging to R, the authors calculated the Spearman correlation. This correlation is calculated between judgments given by r and the average of judgments given by other raters. This correlation is calculated for all examples for which r has given the rating [16]. The authors used Principal Preserved Component Analysis (PPCA) to acquire a more nuanced understanding of rater agreement and the emotion space's hidden structure [17].

The authors in [15] used different algorithms like BERT, BiLSTM, etc. for the calculation of the results. Out of the algorithms used, the model that uses BERT yields an average value of 0.46 as a F1 score.

2.6 DENS Dataset

In 2019, Liu et al. [18] discussed a dataset known as DENS. The full form of this dataset is the "Dataset for Emotions of Narrative Sequences". For doing emotion analysis in several classes, the authors used detailed narratives of English language. They collected data, including various classic literature and modern online narratives, available on the two sources, Project Gutenberg and Wattpad, respectively. Amazon Mechanical Turk was used to perform the annotation. There were three different annotators for each paragraph. Only paragraphs that received a majority of the annotators' approval were considered valid. In mathematical terms, this is similar to obtaining a value of k equal to 0.4 for the Fleiss Kappa method [9].

The emotions discussed in the dataset [18] were neutral, fear, sadness, anger, joy, love, anticipation, surprise, and disgust. For the calculation of the result, they calculated the micro-F1 score of the various models, such as linear SVM, TF-IDF, Depeche Emotion lexicons, NRC Emotion lexicons, Hierarchical RNN (HRNN), Bi-directional RNN with Self-Attention, and BERT model which was properly fine-tuned. The BERT model achieved a value of 60.4 for the micro-F1 score. This value was the highest among all the above-mentioned models. The future scope given by the authors is to use intelligence for emotion analysis.

2.7 ISEAR Dataset

The Swiss National Centre of Competence in Research created the ISEAR dataset by conducting a survey [19]. The survey was headed by Wallbott and Scherer. It comprises different tags for emotions which includes: sadness, joy, guilt, fear, anger, disgust, and shame. A questionnaire regarding their experiences and feelings toward a specific event was given to about three thousand applicants from different cultural backgrounds. Their responses were collected to prepare this dataset. This dataset contains 7666 final sentences labeled with specific emotions. The distribution of the dataset was as follows: sadness-1096, joy-1094, fear-1095, guilt-1093, anger1096, shame-1096, disgust-1096.

2.8 EmoInt Dataset

In 2017, Mohammad et al. [20] developed “EmoInt”. This dataset was employed in the competition known as "WASSA2017 Shared Task on Emotion Intensity." In order to identify the emotional intensities, the authors created a dataset. They found that emotional hashtags in tweets are used to convey strong emotions. The dataset consists of tweets labelled for various emotion intensities. These emotions comprise sadness, anger, joy, and fear. In order to get true fine-grained ratings and improve uniformity in the annotation process, the authors employed the Best-Worst Scaling (BWS) method. Individual feature sets used by authors were character N-Grams (CN), Word N-Grams (WN), Word Embedding (WE), All Lexicons (L), and individual lexicons. They calculated the Pearson Correlation r by trying various combinations of above mentioned individual feature sets. Out of all the possible combinations, the combination Word Embedding (WE) + All Lexicons (L)achieved the highest values of 0.64 for anger, 0.65 for joy, 0.71 for sadness and the combination of Word N- Grams (WN) + Word Embedding (WE) + All Lexicons (L) gained highest value of 0.65 for fear emotion.

2.9 EmoBank Dataset

In 2017, Buechel et al. [21] discussed the “EmoBank Dataset”. This dataset consists of 10,000 English sentences from newspapers, travel guides, blogs, letters, news headlines, fiction, and essays. The authors annotated the data and placed it in the Valence-Arousal-Dominance (VAD) format. This dataset broadly consisted of two sub-datasets: SemEval-2007 (SE-07) and MASC. MASC is the Sub-Corpus of the American National Corpus which is manually annotated. They selected the domain of news headlines and collected 1092 filtered headlines for the dataset from SE-07. From the MASC, they selected domains of newspapers (1,314), travel guides (919), blogs (1,336), letters (1,413), essays (1,135), and fiction (2,753). The dataset was annotated by using CROWDFLOWER (CF) crowd workers. For calculating Inter-Annotator Agreement (IAA) they selected MAE (Mean Absolute Error) or Pearson’s correlation coefficient. The emotions discussed in the dataset were fear, sadness, anger, disgust, surprise, and happiness.

2.10 Twitter Emotion Corpus Dataset

Mohammad et al. [22] proposed the TEC (Twitter Emotion Corpus) Dataset in the year of 2012. TEC was built by collecting more than 20,000 tweets from the Twitter website by using emotion-word hashtags. The authors mainly aimed at proving the correlation between self-annotation hashtags and the hashtags of trained judges. The authors conducted various experiments on the TEC dataset. The first experiment was “Is it possible for a classifier to pick up on emotion hashtags?”. For this experiment, they received an F-score of 49.9 percent. The second experiment was “Can TEC improve emotion classification in a new domain?”. For this experiment,they received an F-score of 40.1 percent. The six emotions discussed in the dataset were sadness, joy, anger, fear, surprise, and disgust. In the future, authors would collect tweets with hashtags for other emotions.

2.11 Overall Analysis of Datasets

Of the ten datasets mentioned above, Dreaddit [8] and TensiStrength [10] are primarily utilized for stress detection. The SMHD dataset [13] is used to solve mental health problems. The remaining seven datasets [14, 15, 18, 20–23] are used for detecting emotions from text. Nonetheless, stress in a person is clearly indicated by the presence of three negative emotions: fear, anger, and sadness [7]. One can prepare their own dataset by combining the text associated with these emotions. And this dataset can further be used for stress detection. Dataset annotation is an important task. Most of the above authors used manual annotators to annotate data. In order to get IIA, they have used various mathematical functions such as Spearman correlation, Principal preserved Component Analysis, Pearson correlation, and Krippendorff coefficient. Various ML techniques, like SVM, and DL techniques, like BiLSTM, GRNN, CNN, etc., have been used to get results on the datasets. The BERT model served as the standard model for most researchers’ work. A detailed analysis of the above datasets will be useful for researchers working in the domains of stress detection and emotion detection from text. Table 1 discusses a few significant datasets.

Table 1: Tabular analysis of the available datasets which uses text for stress/emotion identification

Dataset	Details	Techniques used	Annotation	Link
Dreaddit [8]	Consists of 190K posts from five categories (abuse, anxiety, financial, PTSD, social) of Reddit communities.	BERT-base model achieved Precision (0.7518), Recall (0.8699), F-score (0.8065)	Fleiss Kappa method	http://www.cs.columbia.edu/~eturc an data/dreaddit zip.

TensiStrength [10]	Constructed by fetching the tweets from Twitter.	Pearson correlations and MAD, AdaBoost, NB, DT, J48 Tree, JRip Rules, LR,SVM	Krippendorff Coefficient	http://sentistrength.wlv.ac.uk/ TensiStrength.html
GoEmotion [15]	Consist of manually annotated 58k English Reddit comments	The model based on BERT obtained an average F1-score of .46	Principal Preserved Component Analysis	https://github.com/googleresearch/ googleresearch/tree/master/goemotions
DENS [18]	Collection of English long form narratives used for multi-class emotion analysis	An average micro-F1 score of 60.4 was attained by the BERT model.	Fleiss Kappa method	Send Data request on academic dataset @wattpad.com
Vent[14]	Consist of 10,48,576 vents(posts) from Vent website	BERT model with emotion labels Affection achieved Precision (.62), Recall (.65), Fscore (.63)	–	The dataset is not publicly available

3 Utilization of Datasets

This Section discusses the usage of each of the 10 mentioned datasets (in Section 2) by researchers.

For this research work, we have selected citations from each dataset as one of the metrics to determine its utilization. In order to find the citations for datasets, we have used the Google Scholar website. On the 24th day of October 2023 we calculated the total citations of each of the 10 datasets mentioned inSection 2 and have listed them here in Table 2.

Table 2: Year wise Citations of Datasets

Year wise Citations of Datasets						
Dataset	Total Citation till 24 Oct 2023	2019	2020	2021	2022	2023
VENT	8	0	0	4	1	3
DENS	34	1	9	9	8	6
Dreaddit	75	0	3	12	33	27
TensiStrength	82	6	20	10	15	15
SMHD	118	4	15	26	39	35
EmoBank	221	35	35	41	37	38
EmoInt	283	43	37	48	38	30
TEC	304	53	34	22	22	27
GoEmotion	408	2	15	75	148	161
ISEAR	1351	73	74	106	85	80

The complete citation count for the dataset dated October 24, 2023 is displayed in Figure 2, and Figure 3 gives the year-wise citations of the dataset.

Total citations received by datasets are as follows: for VENT (8), DENS (34), Dreaddit (75), TensiStrength (82), SMHD (118), EmoBank (221), EmoInt (283), TEC (304), GoEmotion (408), and ISEAR (1351). Among all the datasets, the ISEAR dataset has received the highest citations at 1351. It was the oldest dataset among all the remaining datasets, and it is highly used for emotion detection tasks. The second- highest citations received by the GoEmotion dataset, i.e., 408. Since its inception, the GoEmotion dataset has been the favorite choice of researchers for emotion detection and stress detection.

The ISEAR dataset received the highest citations for all three years, i.e., 2019, 2020, and 2021. In 2019, the citations were 73; in 2020, the citations were 74; and in 2021, the citations were 106. The GoEmotion dataset received the highest citation in 2022 and 2023. In 2022, the citations were 148, and in 2023, the citations were 161.

Every dataset is important and is used for a specific purpose. By calculating the citation, we will get some idea about the number of researchers utilizing the dataset for their work. Some datasets like VENT, DENS, and Dreaddit were introduced in recent years, and hence they have a lower number of citations.

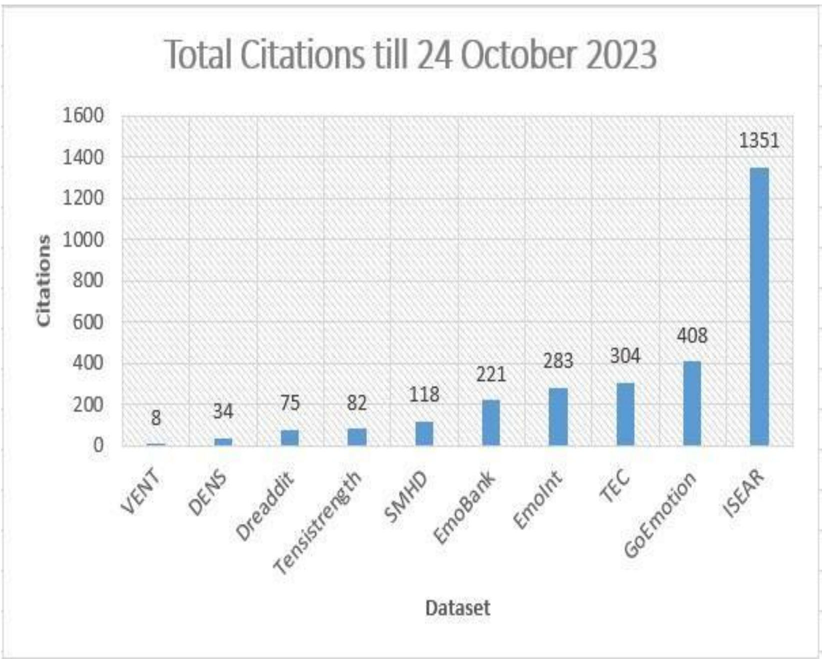

Fig.2: Total Citations of Dataset dated 24 October 2023

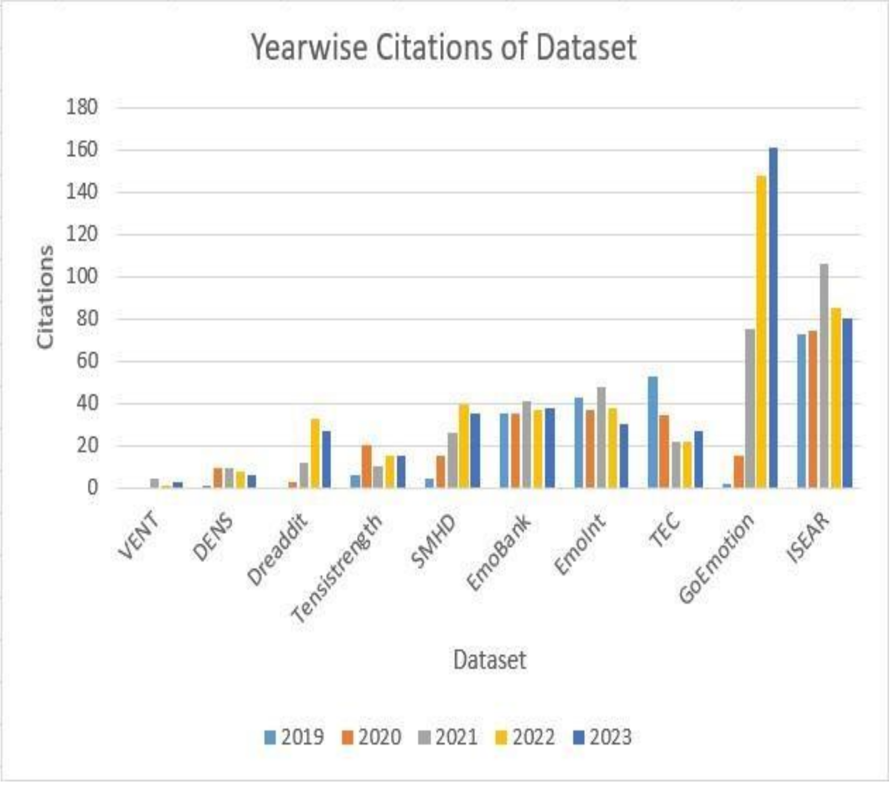

Fig.3: Year wise Citations of Dataset

The remaining datasets, i.e., TensiStrength, SMHD, EmoBank, EmoInt, TEC, GoEmotion, and ISEAR, have had a good number of citationsin the last five years. This indicates that these datasets are highly useful for emotion detection and stress detection tasks.

4 Techniques for inter rater agreement

During the analysis of datasets, one crucial step is the annotation of datasets. Most authors used human annotators for annotating their texts. In general, two to five human annotators are used by authors to annotate the same text. In order to finalize the gold label for the text, they have used various mathematical techniques for inter-rater agreement.

Some important techniques for getting inter-rater agreement are discussed below:

4.1 Pearson Correlation Coefficient (PCC)

One can use the Pearson correlation method to determine the linear relationship between two variables. Karl Pearson created it in the late 19th century. It is mainly used to calculate the strength of the correlation and thedirection in which the correlation progresses. The value of PCC (r) ranges from -1 to +1. An "r" value of -1 depicts a perfect negative correlation, 0 depicts no correlation and +1 depicts a perfect positive correlation between the given variables [24].

The value of PCC is computed by the formula:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

where,

$\overline{x}$ and $\overline{y}$ represent the mean values of the two variables, and x_i and y_i are the values of the individual variables [44].

4.2 Spearman Correlation Coefficient(SCC)

The Spearman Correlation Coefficient, essentially thought of as a rank-based variation of PCC, can be utilized for variables with non-normal distributions and non-linear relationships. Additionally, it can be applied to studies of ordinal features as well as continuous data. When working with data that might not meet Pearson correlation's linearity requirements, Spearman correlation can be helpful. Like the PCC, the value of SCC (ρ) also ranges between -1 to +1. A ρ value of -1 depicts a perfect negative correlation, 0 depicts no correlation and +1 depicts a perfect positive correlation between the variables [25].

The value of SCC is computed by the formula:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

where,

n is the number of samples that are available and d_i are the pairwise distances of the ranks of the variables x_i and y_i.

4.3 Krippendorff's Alpha Coefficient (KAC)

KAC is a measure in Statistics which is used to assess the consistency or accordance between several raters or coders. It was developed by the scientist Klaus Krippendorff. The agreement that has been observed between raters as well as the expected agreement that might happen by chance are both taken into account by Krippendorff's alpha. It suggests a single value that measures the degree of accordance between the raters or coders over and beyond what would be predicted by chance. KAC (α) also ranges from -1 to 1. A higher α value signifies a greater agreement among the raters. A α value of 0 indicates agreement no better than random chance and 1 indicates perfect accordance. A negative α value indicates that there is low reliability or agreement among the raters [12]. Researchers and analysts can utilize Krippendorff's alpha, which offers a more thorough measurement than basic agreement percentages, to assess the agreement or reliability of human coding or annotation processes.

The value of KAC is computed by the formula:

$$\alpha = 1 - \frac{Do}{De}$$

where,

Do = The observed disagreement among raters. It measures how often the raters disagree on their assignments, and De = The expected disagreement, representing the level of disagreement that would occur by chance.

4.4 Fleiss' Kappa

Joseph L. Fleiss invented Fleiss' Kappa in 1971. This statistical measure is used to evaluate the degree of agreement between multiple raters or observers when categorizing items into discrete groups. It is an advancement of Cohen's Kappa, which measures the degree of agreement between two raters. For scenarios involving three or more raters, Fleiss' Kappa is used [9]. The value of Fleiss' Kappa (k) is given by the formula:

$$k = \frac{Po - Pe}{1 - Pe}$$

where,

Pe is the expected proportion of agreement by chance and Po is the observed proportion of agreement.

k can have a value between -1 and 1. Perfect agreement among raters is indicated by a value of k = 1, whereas the observed agreement is no better than chance is indicated by a value of k = 0. The negative value of k shows the systematic disagreement among raters.

5 Detailed Analysis of Linguistic Inquiry and Word Count (LIWC)

There are various tools and language resources available for extracting emotions from the text. Tools available are the Valence Aware Dictionary for Sentiment Reasoning [VADER] [26], LIWC [27], NRC Emotion Lexicon [28], Affective Norms for English Words [ANEW] [29], TextBlob [30], EMOTIVE [31], SenticWordNet [32], Sentistrength [33], SenticNet [34] etc. Out of these tools, LIWC is used by most of the researchers. In the remaining part of this section, we will be discussing LIWC in more detail.

LIWC [27] is a very popular application for doing text analysis. This tool accepts the user data and then compares all the words in the user data with the large number of LIWC's inbuilt dictionaries, producing results that determine the percentage of user words that fall into each category of LIWC. There are a total of five versions of the LIWC application. The versions are LIWC, the second version (LIWC2001), the third version (LIWC2007), the fourth version (2015), and now the fifth version (LIWC-22). The latest version, i.e., LIWC22, updated the original dictionary and software with the intention to enhance the analysis of text data.

The important part of LIWC-22 text analysis technique lies in its dictionary. Over 12,000 words, word stems, sentences, and several emoji's make up the internal dictionary of LIWC-22. They have redesigned the overall architecture of the dictionary by splitting the main categories into two parts, i.e., the "Basic" and "Expanded" super-categories.

"Basic" categories include four main additions, i.e., determiners, cognition, affect, and social behaviors. Determiners are one of the types of parts of speech used by linguists to describe words that come before nouns that specify a quantity (the second rule, five toys). The new generic category of cognition represents many ways in which people think or relate to their thinking. Affect category is changed in order to overcome the two flaws of previous versions. The first flaw was the variation in usage of emotion language from the time of the first version of LIWC, whereas the second one was that there was no clear distinction between "emotion words" and "sentiment." Social Behaviors category improves the previous "social processes" variable by adding more words associated with prosocial behaviors, politeness, interpersonal conflict, moralization, and communication.

LIWC-22 has removed some categories from previous versions. They are interrogatives (where, who, what), comparison words (best, after, greater), certain low-base-rate punctuation (semicolons, colons, quotation marks, parentheses, dashes), and relativity (motion words, space, sum of time). The reasons for their exclusion are their rare use by most of the researchers, low internal solidity, and continually low base rates.

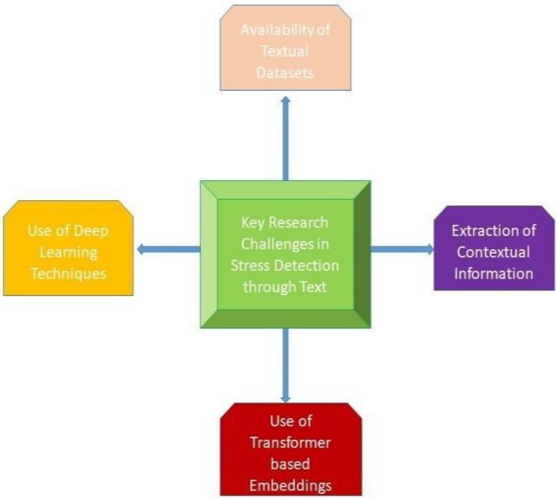
Availability of Textual Datasets
Use of Deep Learning Techniques
Key Research Challenges in Stress Detection through Text
Extraction of Contextual Information
Use of Transformer based Embeddings

Fig.4: Key Research Challenges in Stress Detection through Text

6 Key Research Challenges in Stress Detection through Text

From the survey of recent research on stress detection, it can be inferred that a very small amount of work has been done in the area of stress identification using text data and that it has a huge scope for research. Figure 4 describes the key research challenges in stress detection using text data.

The survey of papers on stress and emotion detection through text suggests that there are two phases of research: language representation and classification. In the first phase, the extraction of dependent information is an important factor to consider. Once the contextual information is successfully extracted, it will help to improve the classification accuracy of the second phase [35,36]. Transformer based embedding improves the accuracy of contextual information extraction [37,39]. There are some limitations to the use of transformer-based embedding. Sometimes, it may not be suitable for small networks. Increased complexity and a lack of vocabulary are some of the other problems of transformer-based embedding [40,41].

A limited number of text datasets exist for stress detection. There is a great need to create a large and balanced dataset for detecting stress through text. There are datasets based on lexicons that are used to identify emotions in text. The improvement of such datasets for stress detection could be a great field of research. A large number of emotion-tagged assets in various languages are required [42]. Languages such as Spanish, Hindi, and French can be used to create emotion-labeled resources, and it can improve the research in these languages. Emoticons are used to express the intensity of emotions. There is a great need for research in analyzing the use of emoticons with emotions.

Most researchers use social networking websites to collect their datasets. People use casual language when writing posts on social networking websites. The writing styles of people need to be analyzed to learn the context of information. DL is a promising but difficult technique for retaining syntactic structures and word order. The use of DL techniques for emotion extraction through text is very complex. More exploration is required to use DL techniques in the field of emotion detection through text [43].

7 Conclusion

Stress detection in humans is of utmost importance as far as building a good and healthy society is concerned. Human-written texts are themselves helpful in detecting stress. This article elaborates very well on the work of researchers who experimented with the detection of stress using text data. To apply DL algorithms, we require datasets in large quantities. Currently, large text-based datasets for stress detection are not available, and therefore there is a great need to prepare such datasets. The articles presented herewith discuss, at large, the base papers of existing text-based datasets for emotion and stress detection. This information will be helpful for upcoming researchers to prepare balanced datasets for their research work. Various other challenges and key research directions have also been discussed in this article. The challenges mentioned above can be overcome by using the latest DL techniques with NLP, resulting in the detection of the stress of the stressed person.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Rudland, J.R., Golding, C., Wilkinson, T.J.: The stress paradox: how stress can be good for learning. Medical education 54(1), 40–45 (2020)
2. Levenson, R.W.: Stress and illness: a role for specific emotions. Psychosomatic medicine 81(8), 720 (2019)
3. Scherer, K.R.: What are emotions? and how can they be measured? Social science information 44(4), 695–729 (2005)
4. Monroe, S.M., Slavich, G.: Psychological stressors: overview. Stress: Concepts, cognition, emotion, and behavior, 109–115 (2016)
5. Pendke, K., Digalwar, M., Goel, C.: Identification and analysis of emotions from the text for stress detection. In: Proceedings of the 2023 Fifteenth International Conference on Contemporary Computing, pp. 623–630 (2023)
6. Zhang, T., Schoene, A.M., Ji, S., Ananiadou, S.: Natural language processing applied to mental illness detection: a narrative review. NPJ digital medicine 5(1), 1–13 (2022)
7. Turcan, E., Muresan, S., McKeown, K.: Emotion-infused models for explainable psychological stress detection. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 2895–2909 (2021)
8. Turcan, E., McKeown, K.: Dreaddit: A reddit dataset for stress analysis in social media. arXiv:1911.00133(2019)
9. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychological bulletin 76(5), 378 (1971)
10. Thelwall, M.: Tensistrength: Stress and relaxation magnitude detection for social media texts. Information Processing Management 53(1), 106– 121 (2017)
11. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., Kappas, A.: Sentiment strength detection in short informal text. Journal of the American society for information science and technology 61(12), 2544–2558 (2010)
12. Krippendorff, K.: Content analysis: An introduction to its methodology (2nd Thousand Oaks. CA: Sage Publications (2004)
13. Cohan, A., Desmet, B., Yates, A., Soldaini, L., MacAvaney, S., Goharian, N.: Smhd: a large-scale resource for exploring online language usage for multiple mental health conditions. arXiv:1806.05258 (2018)

14. Malko, A., Paris, C., Duenser, A., Kangas, M., Molla, D., Sparks, R., Wan, S.: Demonstrating the reliability of self-annotated emotion data. In: Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access, pp. 45–54 (2021)
15. Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., Ravi, S.: Goemotions: A dataset of fine-grained emotions. arXiv:2005.00547 (2020)
16. Delgado, R., Tibau, X.-A.: Why cohen's kappa should be avoided as performance measure in classification.PloS one 14(9), 0222916 (2019)
17. Cowen, A.S., Laukka, P., Elfenbein, H.A., Liu, R., Keltner, D.: The primacy of categories in the recognitionof 12 emotions in speech prosody across two cultures. Nature human behaviour 3(4), 369–382 (2019)
18. Liu, C., Osama, M., De Andrade, A.: Dens: A dataset for multi-class emotion analysis. arXiv:1910.11769 (2019)
19. Scherer, K.R., Wallbott, H.G.: Evidence for universality and cultural variation of differential emotion response patterning. Journal of personality and social psychology 66(2), 310 (1994)
20. Mohammad, S.M., Bravo-Marquez, F.: Wassa-2017 shared task on emotion intensity. arXiv:1708.03700 (2017)
21. Buechel, S., Hahn, U.: Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pp. 578–585 (2017)
22. Mohammad, S.: emotional tweets. In: * SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012), pp. 246–255 (2012)
23. Li, Y., Kazameini, A., Mehta, Y., Cambria, E.: Multitask learning for emotion and personality detection. arXiv:2101.02346 (2021)
24. Pearson, K.: Contributions to the mathematical theory of evolution. Philosophical Transactions of the Royal Society of London. A 185, 71–110 (1894)
25. Spearman, C.: The proof and measurement of association between two things. (1961)
26. Hutto, C., Gilbert, E.: Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 8, pp. 216–225 (2014)
27. Boyd, R.L., Ashokkumar, A., Seraj, S., Pennebaker, J.W.: The development and psychometric properties of liwc-22. Austin, TX: University of Texas at Austin, 1–47 (2022)
28. Mohammad, S.M., Turney, P.D.: Crowdsourcing a word–emotion association lexicon. Computational intelligence 29(3), 436–465 (2013)
29. Bradley, M.M., Lang, P.J.: Affective norms for English words (anew): Instruction manual and affective ratings. Technical report, Technical report C-1, the center for research in psychophysiology . . . (1999)
30. Govindasamy, K.A., Palanichamy, N.: Depression detection using machine learning techniques on twitter data. In: 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), pp.960–966 (2021). IEEE
31. Sykora, M., Jackson, T., O'Brien, A., Elayan, S.: Emotive ontology: Extracting fine-grained emotions fromterse, informal messages (2013)
32. Baccianella, S., Esuli, A., Sebastiani, F., et al.: Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: Lrec, vol. 10, pp. 2200–2204 (2010)
33. Thelwall, M., Buckley, K., Paltoglou, G.: Sentiment strength detection for the social web. Journal of the American Society for Information Science and Technology 63(1), 163–173 (2012)
34. Cambria, E., Speer, R., Havasi, C., Hussain, A.: Senticnet: A publicly available semantic resource for opinion mining. In: 2010 AAAI Fall Symposium Series (2010)
35. Salido Ortega, M.G., Rodr´ıguez, L.-F., Gutierrez-Garcia, J.O.: Towards emotion recognition from contextual information using machine learning. Journal of Ambient Intelligence and Humanized Computing 11(8), 3187– 3207 (2020)
36. Acheampong, F.A., Wenyu, C., Nunoo-Mensah, H.: Text-based emotion detection: Advances, challenges, and opportunities. Engineering Reports 2(7), 12189 (2020) 37. Allouch, M., Azaria, A., Azoulay, R., Ben-Izchak, E., Zwilling, M., Zachor, D.A.: Automatic detection of insulting sentences in conversation. In: 2018 IEEE International Conference on the Science of Electrical Engineering in Israel (ICSEE), pp. 1–4 (2018). IEEE
37. Huang, C., Trabelsi, A., Za¨ıane, O.R.: Ana at semeval-2019 task 3: Contextual emotion detection in conversations through hierarchical lstms and bert. arXiv:1904.00132 (2019)
38. Huang, Y.-H., Lee, S.-R., Ma, M.-Y., Chen, Y.-H., Yu, Y.-W., Chen, Y.- S.: Emotionx-idea: Emotion bert– an affectional model for conversation. arXiv:1908.06264 (2019)
39. Joselson, N., Hall´en, R.: Emotion classification with natural language processing (comparing bert and bi-directional lstm models for use with twitter conversations) (2019)
40. Yu, Z., Wang, Y., Liu, Z., Cheng, X.: Emotionx-antenna: An emotion detector with residual gru and text cnn. Technical report, tech. rep., Technical report (2019)

41. Ahmad, Z., Jindal, R., Ekbal, A., Bhattachharyya, P.: Borrow from rich cousin: transfer learning for emotion detection using cross lingual embedding. Expert Systems with Applications 139, 112851 (2020)
42. Ligthart, A., Catal, C., Tekinerdogan, B.: Systematic reviews in sentiment analysis: a tertiary study. Artificial intelligence review 54(7), 4997–5053 (2021)
43. Zhang, B., Yu, J., Chen, W., Liu, H., Li, H., Guo, H.: Experimental Study on Bond Performance of NC-UHPC Interfaces with Different Roughness and Substrate Strength. Materials, 16(7), 2708 (2023).