LAXARY: A Trustworthy Explainable Twitter Analysis Model for Post-Traumatic Stress Disorder Assessment

[1]Mohammad Arif Ul Alam, [2]Dhawal Kapadia
[1]*Department of Computer Science, University of Massachusetts Lowell*
[2]*IQVIA, Manhattan, Newyork*
mohammadariful_alam@uml.edu,dhawalkapadia8@gmail.com

Abstract—**Veteran mental health is a significant national problem as large number of veterans are returning from the recent war in Iraq and continued military presence in Afghanistan. While significant existing works have investigated twitter posts-based Post Traumatic Stress Disorder (PTSD) assessment using blackbox machine learning techniques, these frameworks cannot be trusted by the clinicians due to the lack of clinical explainability. To obtain the trust of clinicians, we explore the big question, can twitter posts provide enough information to fill up clinical PTSD assessment surveys that have been traditionally trusted by clinicians? To answer the above question, we propose, LAXARY (Linguistic Analysis-based Exaplainable Inquiry) model, a novel Explainable Artificial Intelligent (XAI) model to detect and represent PTSD assessment of twitter users using a modified Linguistic Inquiry and Word Count (LIWC) analysis. First, we employ clinically validated survey tools for collecting clinical PTSD assessment data from real twitter users and develop a PTSD Linguistic Dictionary using the PTSD assessment survey results. Then, we use the PTSD Linguistic Dictionary along with machine learning model to fill up the survey tools towards detecting PTSD status and its intensity of corresponding twitter users. Our experimental evaluation on 210 clinically validated veteran twitter users provides promising accuracies of both PTSD classification and its intensity estimation. We also evaluate our developed PTSD Linguistic Dictionary's reliability and validity.**

Index Terms—**Post Traumatic Stress Disorder, Twitter Analysis, Explainable AI, Trustworthy AI**

I. Introduction

Combat veterans diagnosed with PTSD are substantially more likely to engage in a number of high risk activities including engaging in interpersonal violence, attempting suicide, committing suicide, binge drinking, and drug abuse [1]. Despite improved diagnostic screening, outpatient mental health and inpatient treatment for PTSD, the syndrome remains treatment resistant, is typically chronic, and is associated with numerous negative health effects and higher treatment costs [3]. As a result, the Veteran Administration's National Center for PTSD (NCPTSD) suggests to reconceptualize PTSD not just in terms of a psychiatric symptom cluster, but focusing instead on the specific high risk behaviors associated with it, as these may be directly addressed though behavioral change efforts [1]. Consensus prevalence estimates suggest that PTSD impacts between 15-20% of the veteran population which is typically chronic and treatment resistant [1]. The PTSD patients support programs organized by different veterans peer support organization use a set of surveys for local weekly assessment to detect the intensity of PTSD among the returning veterans. However, recent advanced evidence-based care for PTSD sufferers surveys have showed that veterans, suffered with chronic PTSD are reluctant in participating assessments to the professionals which is another significant symptom of war returning veterans with PTSD. Several existing researches showed that, twitter posts of war veterans could be a significant indicator of their mental health and could be utilized to predict PTSD sufferers in time before going out of control [16]–[20], [22], [23]. However, all of the proposed methods relied on either blackbox machine learning methods or language models based sentiments extraction of posted texts which failed to obtain acceptability and trust of clinicians due to the lack of their explainability.

In the context of the above research problem, we aim to answer the following **research questions**

- Given clinicians have trust on clinically validated PTSD assessment surveys, can we fill out PTSD assessment surveys using twitter posts analysis of war-veterans?
- If possible, what sort of analysis and approach are needed to develop such XAI model to detect the prevalence and intensity of PTSD among war-veterans only using the social media (twitter) analysis where users are free to share their everyday mental and social conditions?
- How much quantitative improvement do we observe in our model's ability to explain both detection and intensity estimation of PTSD?

In this paper, we propose LAXARY, an explainable and trustworthy representation of PTSD classification and its intensity for clinicians.

The **key contributions** of our work are summarized below,

- The novelty of LAXARY lies on the proposed clinical surveys-based PTSD Linguistic dictionary creation with words/aspects which represents the instantaneous perturbation of twitter-based sentiments as a specific pattern and help calculate the possible scores of each survey question.
- LAXARY includes a modified LIWC model to calculate the possible scores of each survey question using PTSD Linguistic Dictionary to fill out the PTSD assessment

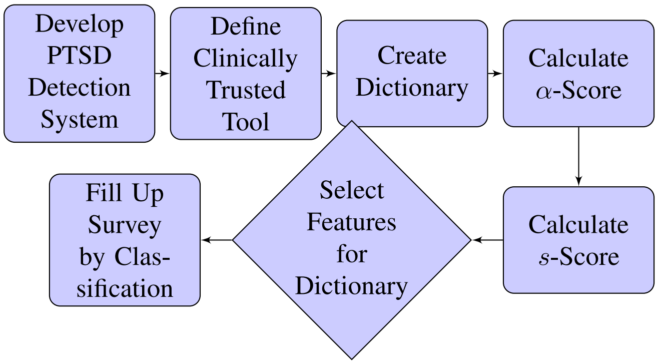
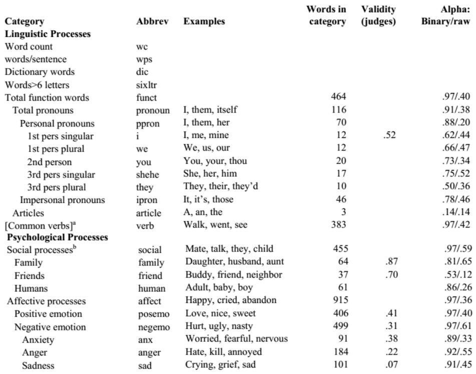

Develop PTSD Detection System
Define Clinically Trusted Tool
Create Dictionary
Calculate α-Score
Calculate s-Score
Select Features for Dictionary
Fill Up Survey by Classification

Fig. 1. Overview of our framework



Category	Abbrev	Examples	Words in category	Validity (judges)	Alpha: Binary/raw
Linguistic Processes					
Word count	wc				
words/sentence	wps				
Dictionary words	dic				
Words>6 letters	sixltr				
Total function words	funct		464		.97/.40
Total pronouns	pronoun	I, them, itself	116		.91/.38
Personal pronouns	ppron	I, them, her	70		.88/.20
1st pers singular	i	I, me, mine	12	.52	.62/.44
1st pers plural	we	We, us, our	12		.66/.47
2nd person	you	You, your, thou	20		.73/.34
3rd pers singular	shehe	She, her, him	17		.75/.52
3rd pers plural	they	They, their, they'd	10		.50/.36
Impersonal pronouns	ipron	It, it's, those	46		.78/.46
Articles	article	A, an, the	3		.14/.14
[Common verbs][a]	verb	Walk, went, see	383		.97/.42
Psychological Processes					
Social processes[b]	social	Mate, talk, they, child	455		.97/.59
Family	family	Daughter, husband, aunt	64	.87	.81/.65
Friends	friend	Buddy, friend, neighbor	37	.70	.53/.12
Humans	human	Adult, baby, boy	61		.86/.26
Affective processes	affect	Happy, cried, abandon	915		.97/.36
Positive emotion	posemo	Love, nice, sweet	406	.41	.97/.40
Negative emotion	negemo	Hurt, ugly, nasty	499	.31	.97/.61
Anxiety	anx	Worried, fearful, nervous	91	.38	.89/.33
Anger	anger	Hate, kill, annoyed	184	.22	.92/.55
Sadness	sad	Crying, grief, sad	101	.07	.91/.45

Fig. 2. WordStat dictionary sample

surveys which provides a practical way not only to determine fine-grained discrimination of physiological and psychological health markers of PTSD without incurring the expensive and laborious in-situ laboratory testing or surveys, but also obtain trusts of clinicians who are expected to see traditional survey results of the PTSD assessment.

- Finally, we evaluate the accuracy of LAXARY model performance and reliability-validity of generated PTSD Linguistic Dictionary using real twitter users' posts. Our results show that, given normal weekly messages posted in twitter, LAXARY can provide very high accuracy in filling up surveys towards identifying PTSD ($\approx 96\%$) and its intensity (≈ 1.2 mean squared error).

II. Overview

Fig. I shows a schematic representation of our proposed model. It consists of the following logical steps: (i) Develop PTSD Detection System using twitter posts of war-veterans(ii) design real surveys from the popular symptoms based mental disease assessment surveys; (iii) define single category and create PTSD Linguistic Dictionary for each survey question and multiple aspect/words for each question; (iv) calculate α-scores for each category and dimension based on linguistic inquiry and word count as well as the aspects/words based dictionary; (v) calculate scaling scores (s-scores) for each dimension based on the α-scores and s-scores of each category based on the s-scores of its dimensions; (vi) rank features according to the contributions of achieving separation among categories associated with different α-scores and s-scores; and select feature sets that minimize the overlap among categories as associated with the target classifier (SGD); and finally (vii) estimate the quality of selected features-based classification for filling up surveys based on classified categories i.e. PTSD assessment which is trustworthy among the psychiatry community.

III. Related Works

Twitter activity based mental health assessment has been utmost importance to the Natural Language Processing (NLP) researchers and social media analysts for decades. Several studies have turned to social media data to study mental health, since it provides an unbiased collection of a person's language and behavior, which has been shown to be useful in diagnosing conditions. [5] used n-gram language model (CLM) based s-score measure setting up some user centric emotional word sets. [6] used positive and negative PTSD data to train three classifiers: (i) one unigram language model (ULM); (ii) one character n-gram language model (CLM); and 3) one from the LIWC categories α-scores and found that last one gives more accuracy than other ones. [7] used two types of s-scores taking the ratio of negative and positive language models. Differences in language use have been observed in the personal writing of students who score highly on depression scales [16], forum posts for depression [17], self narratives for PTSD ([18], [19]), and chat rooms for bipolar [20]. Specifically in social media, differences have previously been observed between depressed and control groups (as assessed by internet-administered batteries) via LIWC: depressed users more frequently use first person pronouns ([22]) and more frequently use negative emotion words and anger words on Twitter, but show no differences in positive emotion word usage ([23]). Similarly, an increase in negative emotion and first person pronouns, and a decrease in third person pronouns, (via LIWC) is observed, as well as many manifestations of literature findings in the pattern of life of depressed users (e.g., social engagement, demographics) ([25]). Differences in language use in social media via LIWC have also been observed between PTSD and control groups ([26]).

All of the prior works used some random dictionary related to the human sentiment (positive/negative) word sets as category words to estimate the mental health but very few of them addressed the problem of explainability of their solution to obtain trust of clinicians. Islam et. al proposed an explainable topic modeling framework to rank different mental health features using Local Interpretable Model-Agnostic Explanations and visualize them to understand the features involved in mental health status classification using the [2] which fails to provide trust of clinicians due to its lack of interpretability in clinical terms. In this paper, we develop LAXARY model where first we start investigating clinically validated survey tools which are trustworthy methods of PTSD assessment among clinicians, build our category sets based on the survey questions and use these as dictionary words in terms of first person singular number pronouns aspect for next level LIWC algorithm. Finally, we develop a modified LIWC algorithm to estimate survey scores (similar to sentiment category scores of naive LIWC) which is both explainable and trustworthy to clinicians.

IV. Demographics of Clinically Validated PTSD Assessment Tools

There are many clinically validated PTSD assessment tools that are being used both to detect the prevalence of PTSD and its intensity among sufferers. Among all of the tools, the most popular and well accepted one is Domain-Specific Risk-Taking (DOSPERT) Scale [8]. This is a psychometric scale that assesses risk taking in five content domains: financial decisions (separately for investing versus gambling), health/safety, recreational, ethical, and social decisions. Respondents rate the likelihood that they would engage in domain-specific risky activities (Part I). An optional Part II assesses respondents' perceptions of the magnitude of the risks and expected benefits of the activities judged in Part I. There are more scales that are used in risky behavior analysis of individual's daily activities such as, The Berlin Social Support Scales (BSSS) [9] and Values In Action Scale (VIAS) [10]. Dryhootch America [11], [14], a veteran peer support community organization, chooses 5, 6 and 5 questions respectively from the above mentioned survey systems to assess the PTSD among war veterans and consider rest of them as irrelevant to PTSD. The details of dryhootch chosen survey scale are stated in Table I. Table!II shows a sample DOSPERT scale demographic chosen by dryhootch. The threshold (in Table I) is used to calculate the risky behavior limits. For example, if one individual's weekly DOSPERT score goes over 28, he is in critical situation in terms of risk taking symptoms of PTSD. Dryhootch defines the intensity of PTSD into four categories based on the weekly survey results of all three clinical survey tools (DOSPERT, BSSS and VIAS)

- *High risk PTSD*: If one individual veteran's weekly PTSD assessment scores go above the threshold for all three PTSD assessment tools i.e. DOSPERT, BSSS and VIAS, then he/she is in high risk situation which needs immediate mental support to avoid catastrophic effect of individual's health or surrounding people's life.
- *Moderate risk PTSD*: If one individual veteran's weekly PTSD assessment scores go above the threshold for any two of the three PTSD assessment tools, then he/she is in moderate risk situation which needs close observation and peer mentoring to avoid their risk progression.
- *Low risk PTSD*: If one individual veteran's weekly PTSD assessment scores go above the threshold for any one of the three PTSD assessment tools, then he/she has light symptoms of PTSD.
- *No PTSD*: If one individual veteran's weekly PTSD assessment scores go below the threshold for all three PTSD assessment tools, then he/she has no PTSD.

Tool	D	B	V
questions	8	3	5
chosen	5	6	5
total points	35	18	25
threshold	28	13	15

TABLE I
Dryhootch chosen PTSD assessment surveys (D: DOSPERT, B: BSSS and V: VIAS) demographics

Questions
1: Betting a day's income at the horse races
2: Drinking heavily at a social function
3: Disagreeing with an authority figure on a major issue
4: Engaging in unprotected sex
5: Leaving your young children alone at home while running an errand
Answers (Scores)
1: Extremely Unlikely
2: Moderately Unlikely
3: Somewhat Unlikely
4: Not Sure
5: Somewhat Likely
6: Moderately Likely
7: Extremely Likely
0: Skip Question

TABLE II
Sample Dryhootch chosen questions from DOSPERT

V. Twitter-based PTSD Detection

To develop an explainable model, we first need to develop twitter-based PTSD detection algorithm. In this section, we describe the data collection and the development of our core LAXARY model.

A. Data Collection

We use an automated regular expression based searching to find potential veterans with PTSD in twitter, and then refine the list manually. First, we select different keywords to search twitter users of different categories. For example, to search self-claimed diagnosed PTSD sufferers, we select keywords related to PTSD for example, post trauma, post traumatic disorder, PTSD etc. We use a regular expression to search for statements where the user self-identifies as being diagnosed with PTSD. For example, Table IV shows a self-identified tweet posts. To search veterans, we mostly visit to different twitter accounts of veterans organizations such as ”MA Women Veterans @WomenVeterans”, ”Illinois Veterans @ILVetsAffairs”, ”Veterans Benefits @VAVetBenefits” etc. We define an inclusion criteria as follows: **one twitter user will be part of this study if he/she describes himself/herself as a veteran in the introduction and have at least 25 tweets in last week**. After choosing the initial twitter users, we search for self-identified PTSD sufferers who claim to be diagnosed with PTSD in their twitter posts. We find 685 matching tweets which are manually reviewed to determine if they indicate a genuine statement of a diagnosis for PTSD. Next, we select the username that authored each of these tweets and retrieve last

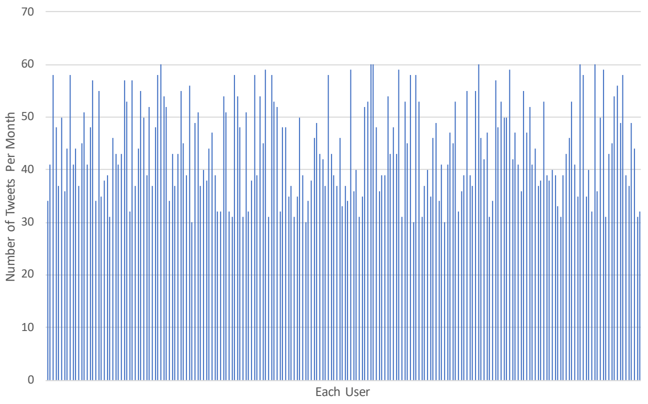


Fig. 3. Each 210 users' average tweets per month

week's tweets via the Twitter API. We then filtered out users with less than 25 tweets and those whose tweets were not at least 75% in English (measured using an automated language ID system.) This filtering left us with 305 users as positive examples. We repeated this process for a group of randomly selected users. We randomly selected 3,000 twitter users who are veterans as per their introduction and have at least 25 tweets in last one week. After filtering (as above) in total 2,423 users remain, whose tweets are used as negative examples developing a 2,728 user's entire weeks' twitter posts where 305 users are self-claimed PTSD sufferers. We distributed Dryhootch chosen surveys among 1,200 users (305 users are self claimed PTSD sufferers and rest of them are randomly chosen from previous 2,423 users) and received 210 successful responses. Among these responses, 92 users were diagnosed as PTSD by any of the three surveys and rest of the 118 users are diagnosed with NO PTSD. Among the clinically diagnosed PTSD sufferers, 17 of them were not self-identified before. However, 7 of the self-identified PTSD sufferers are assessed with no PTSD by PTSD assessment tools. The response rates of PTSD and NO PTSD users are 27% and 12%. In summary, we have collected one week of tweets from 2,728 veterans where 305 users claimed to have diagnosed with PTSD. After distributing Dryhootch surveys, we have a dataset of 210 veteran twitter users among them 92 users are assessed with PTSD and 118 users are diagnosed with no PTSD using clinically validated surveys. The severity of the PTSD are estimated as Non-existent, light, moderate and high PTSD based on how many surveys support the existence of PTSD among the participants according to dryhootch manual [11], [14].

B. Pre-processing

We download 210 users' all twitter posts who are war veterans and clinically diagnosed with PTSD sufferers as well which resulted a total 12,385 tweets. Fig 3 shows each of the 210 veteran twitter users' monthly average tweets. We categorize these Tweets into two groups: Tweets related to work and Tweets not related to work. That is, only the Tweets that use a form of the word work* (e.g. work,worked, working, worker, etc.) or job* (e.g. job, jobs, jobless, etc.) are identified as work-related Tweets, with the remaining categorized as non-work-related Tweets. This categorization method increases the likelihood that most Tweets in the work group are indeed talking about work or job; for instance, Back to work. Projects are firing back up and moving ahead now that baseball is done. This categorization results in 456 work-related Tweets, about 5.4% of all Tweets written in English (and 75 unique Twitter users). To conduct weekly-level analysis, we consider three categorizations of Tweets (i.e. overall Tweets, work-related Tweets, and non work-related Tweets) on a daily basis, and create a text file for each week for each group.

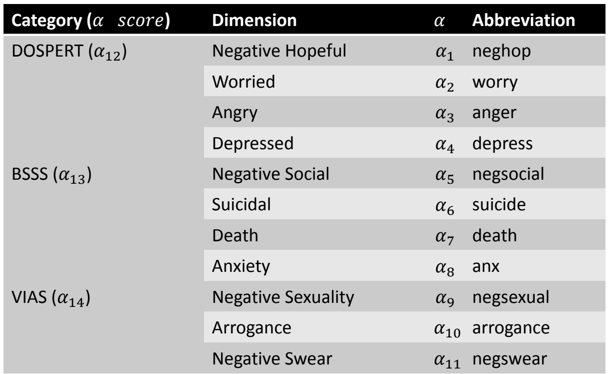
C. PTSD Detection Baseline Model

We use Coppersmith proposed PTSD classification algorithm to develop our baseline blackbox model [7]. We utilize our positive and negative PTSD data (+92,-118) to train three classifiers: (i) unigram language model (ULM) examining individual whole words, (ii) character n-gram language model (CLM), and (iii) LIWC based categorical models above all of the prior ones. The LMs have been shown effective for Twitter classification tasks [5] and LIWC has been previously used for analysis of mental health in Twitter [6]. The language models measure the probability that a word (ULM) or a string of characters (CLM) was generated by the same underlying process as the training data. We first train one of each language model (clm^+ and ulm^+) from the tweets of PTSD users, and another model (clm^- and ulm^-) from the tweets from No PTSD users. Each test tweet t is scored by comparing probabilities from each LM called $s - score$

$$s = \frac{lm^+(t)}{lm^-(t)} \tag{1}$$

A threshold of 1 for $s - score$ divides scores into positive and negative classes. In a multi-class setting, the algorithm minimizes the cross entropy, selecting the model with the highest probability. For each user, we calculate the proportion of tweets scored positively by each LIWC category. These proportions are used as a feature vector in a loglinear regression model [12]. Prior to training, we preprocess the text of each tweet: we replace all usernames with a single token (USER), lowercase all text, and remove extraneous whitespace. We also exclude any tweet that contained a URL, as these often pertain to events external to the user.

We conduct a LIWC analysis of the PTSD and non-PTSD tweets to determine if there are differences in the language usage of PTSD users. We applied the LIWC battery and examined the distribution of words in their language. Each tweet was tokenized by separating on whitespace. For each user, for a subset of the LIWC categories, we measured the proportion of tweets that contained at least one word from that category. Specifically, we examined the following nine categories: first, second and third person pronouns, swear, anger, positive emotion, negative emotion, death, and anxiety words. Second person pronouns were used significantly less



Category (α score)	Dimension	α	Abbreviation
DOSPERT (α_{12})	Negative Hopeful	α_1	neghop
	Worried	α_2	worry
	Angry	α_3	anger
	Depressed	α_4	depress
BSSS (α_{13})	Negative Social	α_5	negsocial
	Suicidal	α_6	suicide
	Death	α_7	death
	Anxiety	α_8	anx
VIAS (α_{14})	Negative Sexuality	α_9	negsexual
	Arrogance	α_{10}	arrogance
	Negative Swear	α_{11}	negswear

Fig. 4. Category Details

often by PTSD users, while third person pronouns and words about anxiety were used significantly more often.

VI. LAXARY: Explainable PTSD Detection Model

The heart of LAXARY framework is the construction of PTSD Linguistic Dictionary. Prior works show that linguistic dictionary based text analysis has been much effective in twitter based sentiment analysis [21], [27]. Our work is the first of its kind that develops its own linguistic dictionary to explain automatic PTSD assessment to confirm trustworthiness to clinicians.

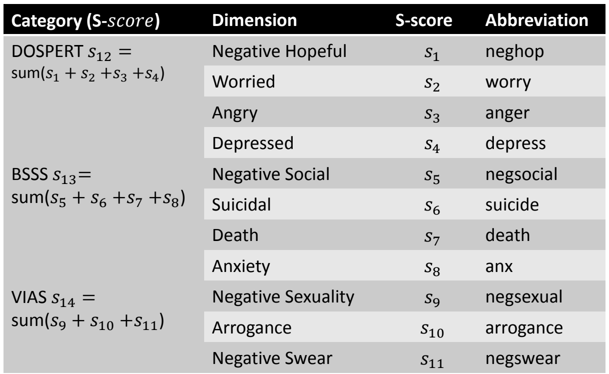
A. *PTSD Linguistic Dictionary Creation*

We use LIWC developed WordStat dictionary format for our text analysis [15]. The LIWC application relies on an internal default dictionary that defines which words should be counted in the target text files. To avoid confusion in the subsequent discussion, text words that are read and analyzed by WordStat are referred to as target words. Words in the WordStat dictionary file will be referred to as dictionary words. Groups of dictionary words that tap a particular domain (e.g., negative emotion words) are variously referred to as subdictionaries or word categories. Fig 2 is a sample WordStat dictionary. There are several steps to use this dictionary which are stated as follows:

Pronoun selection: At first we have to define the pronouns of the target sentiment. Here we used first person singular number pronouns (i.e., I, me, mine etc.) that means we only count those sentences or segments which are only related to first person singular number i.e., related to the person himself.

Category selection: We have to define the categories of each word set thus we can analyze the categories as well as dimensions' text analysis scores. We chose three categories based on the three different surveys: 1) DOSPERT scale; 2) BSSS scale; and 3) VIAS scale.

Dimension selection: We have to define the word sets (also called dimension) for each category. We chose one dimension for each of the questions under each category to reflect real survey system evaluation. Our chosen categories are state in



Category (S-*score*)	Dimension	S-score	Abbreviation
DOSPERT s_{12} = sum($s_1 + s_2$ +s_3 +s_4)	Negative Hopeful	s_1	neghop
	Worried	s_2	worry
	Angry	s_3	anger
	Depressed	s_4	depress
BSSS s_{13}= sum($s_5 + s_6$ +s_7 +s_8)	Negative Social	s_5	negsocial
	Suicidal	s_6	suicide
	Death	s_7	death
	Anxiety	s_8	anx
VIAS s_{14} = sum($s_9 + s_{10}$ +s_{11})	Negative Sexuality	s_9	negsexual
	Arrogance	s_{10}	arrogance
	Negative Swear	s_{11}	negswear

Fig. 5. S-score table details

Algorithm 1 S-Score Calculation Algorithm

```
procedure S-SCORE($\alpha, \max\_\alpha\_of\_category, \#scale$)
    $inter=\frac{\max\_\alpha\_of\_category}{number\_of\_scale})$

    for all i in 1 to $number\_of\_scale$ do
        if $i \times inter \leq \alpha \leq (i+1) \times inter$ then
            s_score =i
        end if
    end for
    return s_score =i
end procedure
```

Fig 4.

Score calculation α-score: α-scores refer to the Cronbach's alphas for the internal reliability of the specific words within each category. The binary alphas are computed on the ratio of occurrence and non-occurrence of each dictionary word whereas the raw or uncorrected alphas are based on the percentage of use of each of the category words within texts.

B. *Psychometric Validation of PTSD Linguistic Dictionary*

After the PTSD Linguistic Dictionary has been created, we empirically evaluate its psychometric properties such as reliability and validity as per American Standards for educational and psychological testing guideline [13]. In psychometrics, reliability is most commonly evaluated by Cronbach's alpha, which assesses internal consistency based on inter-correlations and the number of measured items. In the text analysis scenario, each word in our PTSD Linguistic dictionary is considered an item, and reliability is calculated based on each text file's response to each word item, which forms an N(number of text files) × J(number of words or stems in a dictionary) data matrix. There are two ways to quantify such responses: using percentage data (uncorrected method), or using "present or not" data (binary method) [15]. For the uncorrected method, the data matrix comprises percentage values of each word/stem are calculated from each text file. For the binary method, the data matrix quantifies whether or not a word was used in a text file where ”1” represents yes and ”0” represents no. Once the data matrix is created, it is used to calculate Cronbach’s alpha based on its inter-correlation matrix among the word percentages. We assess reliability based on our selected 210 users’ Tweets which further generated a 23,562 response matrix after running the PTSD Linguistic Dictionary for each user. The response matrix yields reliability of .89 based on the uncorrected method, and .96 based on the binary method, which confirm the high reliability of our PTSD Dictionary created PTSD survey based categories. After assessing the reliability of the PTSD Linguistic dictionary, we focus on the two most common forms of construct validity: convergent validity and discriminant validity [24]. Convergent validity provides evidence that two measures designed to assess the same construct are indeed related; discriminate validity involves evidence that two measures designed to assess different constructs are not too strongly related. In theory, we expect that the PTSD Linguistic dictionary should be positively correlated with other negative PTSD constructs to show convergent validity, and not strongly correlated with positive PTSD constructs to show discriminant validity. To test these two types of validity, we use the same 210 users’ tweets used for the reliability assessment. The results revealed that the PTSD Linguistic dictionary is indeed positively correlated with negative construct dictionaries, including the overall negative PTSD dictionary (r=3.664,p<.001). Table III shows all 16 categorical dictionaries. These results provide strong support for the measurement validity for our newly created PTSD Linguistic dictionary.

#	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
r	4.5	5.2	5.2	4.9	3.2	4.1	3.7	4.2	3.7	5.1	4.2	4.6	3.9	3.8	4.0	4.3

TABLE III

Validity assessment of PTSD Linguistic Dictionary for each question. Top row represents the question number and for all of the cases, p<.001

C. Feature Extraction and Survey Score Estimation

We use the exact similar method of LIWC to extract α-scores for each dimension and categories except we use our generated PTSD Linguistic Dictionary for the task [15]. Thus we have total 16 α-scores in total. Meanwhile, we propose a new type of feature in this regard, which we called scaling-score (s-score). s-score is calculated from α-scores. The purpose of using s-score is to put exact scores of each of the dimension and category thus we can apply the same method used in real weekly survey system. The idea is, we divide each category into their corresponding scale factor (i.e., for DOSPERT scale, BSSS scale and VIAS scales) and divide them into 8, 3 and 5 scaling factors which are used in real survey system. Then we set the s-score from the scaling factors from the α-scores of the corresponding dimension of the questions. The algorithm is stated in Figure VI-A. Following Fig VI-A, we calculate the s-score for each dimension. Then we add up all the s-score of the dimensions to calculate cumulative s-score of particular categories which is displayed in Fig 5. Finally, we have total 32 features among them 16 are α-scores and 16 are s-scores for each category (i.e. each question). We add both of α and s scores together and scale according to their corresponding survey score scales using min-max standardization. Then, the final output is a 16 valued matrix which represent the score for each questions from three different Dryhootch surveys. We use the output to fill up each survey, estimate the prevalence of PTSD and its intensity based on each tool’s respective evaluation metric.

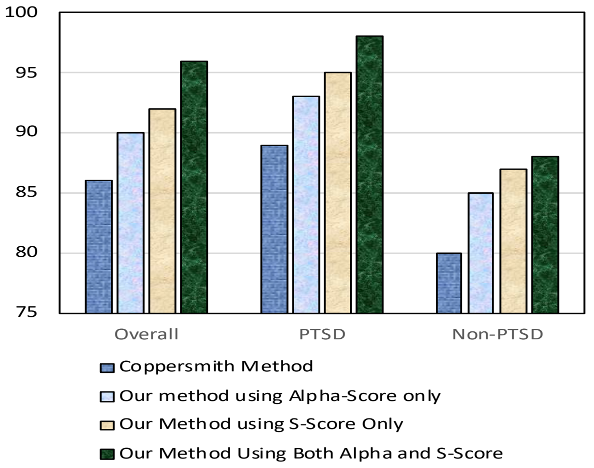


Fig. 6. Comparisons between Coppersmith et. al. and our method

VII. Experimental Evaluation

To validate the performance of LAXARY framework, we first divide the entire 210 users’ twitter posts into training and test dataset. Then, we first developed PTSD Linguistic Dictionary from the twitter posts from training dataset and apply LAXARY framework on test dataset.

In loving memory my mom, she was only 42, I was 17 and taken away from me. I was diagnosed with having P.T.S.D *LINK So today I started therapy, she diagnosed me with anorexia, depression, anxiety disorder, post traumatic stress disorder and wants me to @USERNAME The VA diagnosed me with PTSD, so I cant go in that direction anymore I wanted to share some things that have been helping me heal lately. I was diagnosed with severe complex PTSD and... LINK*

TABLE IV

Example of PTSD user’s twitter post

Class	**TP**	**FP**	**Prec**	**Rec**	**F**	**MCC**	**ROC**	**PRC**
PTSD	0.99	0.32	0.97	0.99	0.98	0.76	0.83	0.96
No-PTSD	0.67	0.07	0.90	0.67	0.77	0.76	0.83	0.63
Overall PTSD	0.96	0.30	0.96	0.96	0.96	0.76	0.83	0.94

TABLE V

LAXARY model based classification details

A. Results

To provide an initial results, we take 50% of users’ last week’s (the week they responded of having PTSD) data

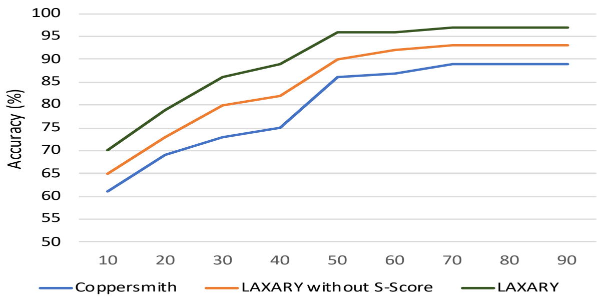


Fig. 7. Percentages of Training dataset and their PTSD detection accuracy results comparisons. Rest of the dataset has been used for testing

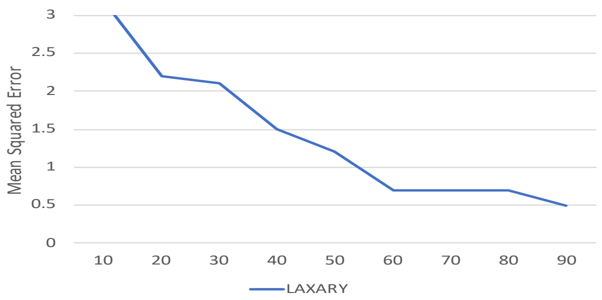


Fig. 8. Percentages of Training dataset and their Mean Squared Error (MSE) of PTSD Intensity. Rest of the dataset has been used for testing

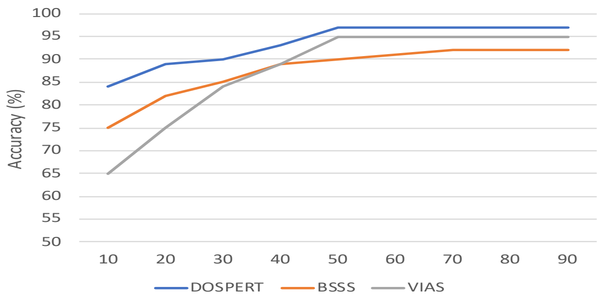


Fig. 9. Percentages of Training dataset and their Accuracies for each Survey Tool. Rest of the dataset has been used for testing

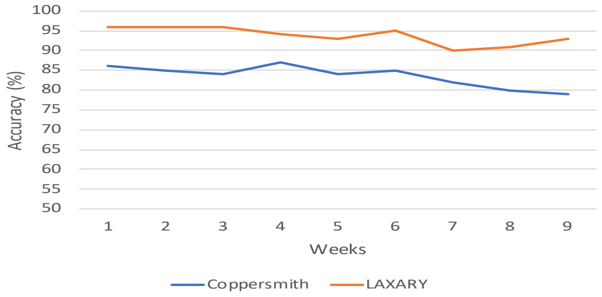


Fig. 10. Weekly PTSD detection accuracy change comparisons with baseline model

to develop PTSD Linguistic dictionary and apply LAXARY framework to fill up surveys on rest of 50% dataset. The distribution of this training-test dataset segmentation followed a 50% distribution of PTSD and No PTSD from the original dataset. Our final survey based classification results showed an accuracy of 96% in detecting PTSD and mean squared error of 1.2 in estimating its intensity given we have four intensity, No PTSD, Low Risk PTSD, Moderate Risk PTSD and High Risk PTSD with a score of 0, 1, 2 and 3 respectively. Table V shows the classification details of our experiment which provide the very good accuracy of our classification. To compare the outperformance of our method, we also implemented Coppersmith et. al. proposed method and achieved an 86% overall accuracy of detecting PTSD users [7] following the same training-test dataset distribution. Fig 6 illustrates the comparisons between LAXARY and Coppersmith et. al. proposed method. Here we can see, the outperformance of our proposed method as well as the importance of $s-score$ estimation. We also illustrates the importance of $\alpha-score$ and $S-score$ in Fig 7. Fig 7 illustrates that if we change the number of training samples (%), LAXARY models outperforms Coppersmith et. al. proposed model under any condition. In terms of intensity, Coppersmith et. al. totally fails to provide any idea however LAXARY provides extremely accurate measures of intensity estimation for PTSD sufferers (as shown in Fig 8) which can be explained simply providing LAXARY model filled out survey details. Table V shows the details of accuracies of both PTSD detection and intensity estimation. Fig 9 shows the classification accuracy changes over the training sample sizes for each survey which shows that DOSPERT scale outperform other surveys. Fig 10 shows that if we take previous weeks (instead of only the week diagnosis of PTSD was taken), there are no significant patterns of PTSD detection.

VIII. Challenges and Future Work

LAXARY is a highly ambitious model that targets to fill up clinically validated survey tools using only twitter posts. Unlike the previous twitter based mental health assessment tools, LAXARY provides a clinically interpretable model which can provide better classification accuracy and intensity of PTSD assessment and can easily obtain the trust of clinicians. The central challenge of LAXARY is to search twitter users from twitter search engine and manually label them for analysis. While developing PTSD Linguistic Dictionary, although we followed exactly same development idea of LIWC WordStat dictionary and tested reliability and validity, our dictionary was not still validated by domain experts as PTSD detection is highly sensitive issue than stress/depression detection. Moreover, given the extreme challenges of searching veterans in twitter using our selection and inclusion criteria, it was extremely difficult to manually find the evidence of the self-claimed PTSD sufferers. Although, we have shown extremely promising initial findings about the representation of a blackbox model into clinically trusted tools, using only 210 users' data is not enough to come up with a trustworthy model.

Moreover, more clinical validation must be done in future with real clinicians to firmly validate LAXARY model provided PTSD assessment outcomes. In future, we aim to collect more data and run not only nationwide but also international-wide data collection to establish our innovation into a real tool. Apart from that, as we achieved promising results in detecting PTSD and its intensity using only twitter data, we aim to develop Linguistic Dictionary for other mental health issues too. Moreover, we will apply our proposed method in other types of mental illness such as depression, bipolar disorder, suicidal ideation and seasonal affective disorder (SAD) etc. As we know, accuracy of particular social media analysis depends on the dataset mostly. We aim to collect more data engaging more researchers to establish a set of mental illness specific Linguistic Database and evaluation technique to solidify the genralizability of our proposed method.

IX. Conclusion

To promote better comfort to the trauma patients, it is really important to detect Post Traumatic Stress Disorder (PTSD) sufferers in time before going out of control that may result catastrophic impacts on society, people around or even sufferers themselves. Although, psychiatrists invented several clinical diagnosis tools (i.e., surveys) by assessing symptoms, signs and impairment associated with PTSD, most of the times, the process of diagnosis happens at the severe stage of illness which may have already caused some irreversible damages of mental health of the sufferers. On the other hand, due to lack of explainability, existing twitter based methods are not trusted by the clinicians. In this paper, we proposed, LAXARY, a novel method of filling up PTSD assessment surveys using weekly twitter posts. As the clinical surveys are trusted and understandable method, we believe that this method will be able to gain trust of clinicians towards early detection of PTSD. Moreover, our proposed LAXARY model, which is first of its kind, can be used to develop any type of mental disorder Linguistic Dictionary providing a generalized and trustworthy mental health assessment framework of any kind.

References

[1] Hartl, TL., Rosen, C., Drescher, K., Lee, TT., Gusman, F. Predicting High-Risk Behaviors in Veterans With Posttraumatic Stress Disorder. The Journal of Nervous and Mental Disease. 2005;193(7):464-472.

[2] Ex-Twit: Explainable Twitter mining on health data T Islam - arXiv preprint arXiv:1906.02132, 2019 - arxiv.org

[3] Rosenheck, R., Frisman L., Kasprow, W. Improving access to disability benefits among homeless persons with mental illness: an agency-specific approach to services integration. American Journal of Public Health. 1999;89(4):524-528.

[4] Sayers, SL., Farrow, VA., Ross, J., Oslin, DW. 2009. Family problems among recently returned military veterans referred for a mental health evaluation. J. Clin Psych.;70(163-170).

[5] Bergsma, S.; McNamee, P.; Bagdouri, M.; Fink, C.; and Wilson, T. 2012. Language identification for creating language-specific Twitter collections. In ACL Workshop on Language in Social Media.

[6] De Choudhury, M.; Gamon, M.; Counts, S.; and Horvitz, E. 2013. Predicting depression via social media. In Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM).

[7] Coppersmith, G.; Dredze, M.; and Harman, C. 2014. Quantifable mental health signals in Twitter.

[8] Blais, A.R., Weber, E. U. (2006) A Domain-Specific Risk-Taking (DOSPERT) scale for adult populations. Judgment and Decision Making, 1, 33-47.

[9] Schulz, U., Schwarzer, R. (2003). Soziale Unterstzung bei der Krankheitsbewltigung. Die Berliner Social Support Skalen (BSSS) [Social support in coping with illness: The Berlin Social Support Scales (BSSS)]. Diagnostica, 49, 73-82.

[10] Macdonald, Bore, Munro. (2008). Values in action scale and the Big 5: An empirical indication of structure. Journal of Research in Personality, 42, 787799

[11] Rizwana Rizia, Zeno Franco, Katinka Hooyer, Nadiyah Johnson, A. B. M. Kowser Patwary, Golam Mushih Tanimul Ahsan, Bob Curry, Mark Flower, Sheikh Iqbal Ahamed: iPeer: A Sociotechnical Systems Approach for Helping Veterans with Civilian Reintegration. ACM DEV 2015: 85-93

[12] Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; and Duchesnay, M. P. E. 2011. scikit-learn: Machine learning in Python. The Journal of Machine Learning Research 12:2825 2830.

[13] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014).Standards foreducational and psychological testing. Washington, DC: AERA.

[14] http://dryhootch.org/

[15] Pennebaker, J. W., Chung, C. K., Ireland, M. E., Gonzales, A. L., Booth, R. J. (2007). The development and psychometric properties of LIWC2007. Austin,

[16] Stephanie S. Rude, Eva-Maria Gortner, and James W. Pennebaker. 2004. Language use of depressed and depression-vulnerable college students. Cognition & Emotion, 18(8):11211133, December

[17] Nairan Ramirez-Esparza, Cindy K. Chung, Ewa Kacewicz, and James W. Pennebaker. 2008. The psychology of word use in depression forums in English and in Spanish: Testing two text analytic approaches. In Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM).

[18] Wendy DAndrea, Pearl H. Chiu, Brooks R. Casas, and Patricia Deldin. 2011. Linguistic predictors of post-traumatic stress disorder symptoms following 11 September 2001. Applied Cognitive Psychology, 26(2):316323, October.

[19] Jennifer Alvarez-Conrad, Lori A. Zoellner, and Edna B. Foa. 2001. Linguistic predictors of trauma pathology and physical health. Applied Cognitive Psychology, 15(7):S159S170.

[20] Adam D. I. Kramer, Susan R. Fussell, and Leslie D. Setlock. 2004. Text analysis as a tool for analyzing conversation in online support groups. In Proceedings of the ACM Annual Conference on Human Factors in Computing Systems (CHI).

[21] van der Lee, Chris and van der Zanden, Tess and Krahmer, Emiel and Mos, Maria and Schouten, Alexander, Automatic identification of writers' intentions: Comparing different methods for predicting relationship goals in online dating profile texts, Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)

[22] Cindy Chung and James Pennebaker. 2007. The psychological functions of function words. Social communication, pages 343359.

[23] Minsu Park, Chiyoung Cha, and Meeyoung Cha. 2012. Depressive moods of users portrayed in Twitter. In Proceedings of the ACM SIGKDD Workshop on Healthcare Informatics (HI-KDD).

[24] Campbell, D.T., & Fiske, D. (1959). Convergent and discriminant validation by the multitraitmultimethod matrix. Psychological Bulletin, 56, 81105.

[25] Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013b. Predicting postpartum changes in emotion and behavior via social media. In Proceedings of the ACM Annual Conference on Human Factors in Computing Systems (CHI), pages 32673276. ACM.

[26] Glen A. Coppersmith, Craig T. Harman, and Mark Dredze. 2014. Measuring post traumatic stress disorder in Twitter. In Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM).

[27] Amanda Andrei, Alison Dingwall, Theresa Dillon, Jennifer Mathieu: Developing a Tagalog Linguistic Inquiry and Word Count (LIWC) 'Disaster' Dictionary for Understanding Mixed Language Social Media: A Work-in-Progress Paper. LaTeCH@EACL 2014: 91-94