Annotating Panic in Social Media using Active Learning, Transformers and Domain Knowledge

Sandra Mitrović
IDSIA - USI/SUPSI
Lugano, Switzerland
sandra.mitrovic@idsia.ch

Fabio Frisone
Università Cattolica del Sacro Cuore
Milan, Italy
fabio.frisone@unime.it

Suryam Gupta
SVNIT
Surat, India
i19ma038@amhd.svnit.ac.in

Chiara Lucifora
University of Bologna
Bologna, Italy
chiara.lucifora@unibo.it

Dragana Čarapić
University of Montenegro
Podgorica, Montenegro
draganacarapic09@gmail.com

Carlo Schillaci, Samuele Di Giovanni
SUPSI
Lugano, Switzerland
{name}.{surname}@student.supsi.ch

Ayushi Singh
SVNIT
Surat, India
i19ma009@amhd.svnit.ac.in

***Abstract*—Nowadays, researchers unanimously agree on the undeniable importance of mental health. However, the literature related to tracking mental disorders in textual content from social media platforms is heavily inclined towards specific problems. In particular, panic disorder/panic attacks are heavily understudied in the current literature and the relevant resources are missing. Therefore, in this work we focus on collecting an annotated dataset. To this end, in order to mitigate the annotation effort by selectively annotating unlabeled data, we propose an active-learning based approach with uncertainty sampling supported by contextualized (Transformer-based) representations, symptomatic and psychometric features and domain expertise. Our evaluation demonstrates the efficiency of the proposed approach both in terms of classification accuracy and predictions confidence. Our contribution to the research community is an annotated dataset of 13,036 tweets that distinguishes between personal panicking experiences such as panic attacks, other panic-related content and completely panic-unrelated content hoping that it will foster research on the topic.**

***Index Terms*—Data acquisition, Active Learning, Mental disorders, Transformers, Uncertainty, Classification algorithms, Natural language processing, Machine learning**

I. Introduction

Mental disorders have significant impact on the society, not only because they affect a large portion of the population but also due to profound consequences they have for individuals, various aspects of their life, their surroundings and the community in general [1]. In particular, monitoring mental health in social media, is of utmost importance given their growing popularity and the increasing rate at which their users share details from their private lives, their thoughts and feelings. Moreover, some social media users post their content almost in real-time. As a result, analysis of user-generated content can provide immediate insight into users' mental health.

The current state of the literature on tracking mental disorders in social media reflects the importance of the topic, but these studies are heavily biased towards specific mental disorders such as depression [2], [3], anorexia [4], bipolar disorder, ADHD and PTSD [5]. Panic disorder in particular is remarkably understudied in the current literature. Furthermore, current studies tend to neglect occasional, less frequent episodes related to mental health. On the contrary, we deem that these episodes, such as panic attacks, could also take a toll on "officially-healthy" person's life quality, e.g. experiencing anticipatory anxiety in fear that the panic attack will reoccur [6], post-panic attack exhaustiveness [7], avoidance behavior [8] to circumvent what might be perceived as a likely problematic situation.

Although psychological studies differ on the etiology of panic disorder/attacks, distinguishing between psychic [9] and organic [10], [11] causes, they undoubtedly reveal the importance of prompt acting when it comes to problem identification. According to the psychic perspective, the reason why the panic attack keeps recurring and worsening is the fact that the conditioning related to stimulus, imagination, and response becomes fixed in the mind, which is capable of not only imagining but also actualizing the fear of death [9]. On the other hand, according to the organic perspective, which focuses primarily on the functioning of the brain, after particularly stressful and emotionally intense periods, the brain might be able not only to interpret even neutral signals as danger, but also to signal alarm in situations other than those actually associated with dangerous situations [10], [11].

Framing the phenomenon of panic attack can be complex. The Diagnostic and Statistical Manual of Mental Disorders defines it as "a sudden surge of intense fear or intense discomfort that reaches a peak within minutes" [12]. However, not all definitions equate panic attack with fear. For example, some psychology dictionaries describe a panic attack as an acute episode of anxiety, characterized by emotional tension and unbearable fear that prevents adequate organization of thought and action [13]. Furthermore, individuals who experience this disorder have an altered vegetative experience, in which they may have, for example, palpitations, pounding heart, accelerated heart rate, sweating, trembling, smothering, and chest pain [12]. In addition, panic can be associated with the phenomena of derealization (when one experiences feelings of unreality) or depersonalization (when one feels detached from oneself) [14]. Additional difficulty when it comes to recognizing panic in social media content, is that

TABLE I
EXAMPLES OF TWEETS AND THEIR RELATIVE INTERPRETATION

	Original tweet text	**Our comment (category)**
a)	*"Am I the only one that's about to have a nervous breakdown? This sh*t is absolutely crazy and some people are acting as if it's a joke! While my a** up here bout to have a panic attack reading sh*t F*CK IM TIRED OF THIS SH*T!!!! "* *"I have had 5 panic attacks TODAY thinking I am going to die of #coronavirus can anyone help me what can I do I am so scared I stay in + only go out for food shop but still petrified @DrAmirKhanGP @DrRanj"*	*Person panicking (PP)*
b)	*"Consumer panic buying creates supply chain "bullwhip effect" – Read Sarah Rathke's commentary: https://t.co/tUQsYwIBBC#supplychain #legal #retail #COVID19 https://t.co/EjZRoxXTfS"* *"1hr before PM Lee's public address about taking additional steps to curb Covid19 in SG, the panic buying resumes. Panic buying is a selfish act. Shop sensibly please! #Covid_19 #Singapore"*	*Pointing at panic buying (PO)*
c)	*"I think #publicpressure is being manufactured by #media- this information due to its wild inaccuracies will only lead to #panic #covid19 https://t.co/enNpWdACek"* *"#Covid19. It's bad, for sure. But the media wants you to believe it's the apocalypse, to keep you glued to live news, which is their only means of competing with ad free streaming services. Buy into the caution, but not the panic. 80,000 USA flu deaths in 2017/18 winter btw"* *"To our media friends all over the world, you have a responsibility to not fuel fear, panic & anxiety. As we report the number of covid19 infections and deaths, please also report the number of people who have recovered. So many people have gone into depression or even worse."*	*Claiming that media is spreading panic (PO)*
d)	*"This new situation can be overwhelming, stressful, and intimidating, but don't panic! We have some tips that can keep you crushing it during the COVID-19 crisis.https://t.co/uBfEAunGyI#COVID19 @jcu #StayAtHome #FlattenTheCurve"* *"WHEN ALL HOPE IS LOST, DON'T GIVE UP! - Short CLIP I am sending love to all the people suffering from mental health problems during the #Coronavirus #lockdown. Don't fear reaching out for help if things go downhill for you . #Itsokaynottobeokay #covid19 https://t.co/wOBccwGtSJ"*	*Person's plea not to panic and/or to ask for help (PO)*
e)	*"The captain of a U.S. Navy aircraft carrier facing a growing outbreak of the #coronavirus on his ship was fired on Thursday by Navy leaders who said he created a panic by sending his memo pleading for help to too many people. https://t.co/43O12O8hVb"* *"I feel half this scare about #coronavirus is created by not-so-experts trying to be one.. half-baked data scientists trying to model the number of deaths and inducing panic in people. I am seeing that some influencers are busy sharing scientific papers on viruses as well."*	*Pointing at someone else panicking (PO)*
f)	*"Scammers are using your fear of #COVID19 for their own gain. They're promising secret cures and pretending to be charities. Phony news on social media is causing more panic. #ThinkBeforeYouClick and before giving anyone personal/financial information. https://t.co/ii8UPvg7UI"* *"It looks like fraudsters are trying to take advantage of #CoronavirusPandemic economic panic via fake celebrity pages. Here is a fraudulent page of @TheEllenShow https://t.co/iLywuzdOaO"*	*Abusing panic to commit fraud (PO)*
g)	*"There's a malicious VN flying around WhatsApp about the Corona virus, 5 G end of the world bullshit ...why are people like this , spreading fear and panic everywhere ..#COVID19 #LaCasaDePapel4 #COVID19Pandemic"* *"Received countless conspiracy theory WhatsApp messages linking 5G to #Coronavirus! Can people stop forwarding unsubstantiated messages and read this, because your actions are creating panic, leading to abuse of engineers and damage to our mobile network. https://t.co/XTeL1cOEnZ"*	*Pointing at fake news / conspiracy theories (PO)*
h)	*"I'm sorry if you have trouble breathing... you're probably gonna die, hate to break it to you.. #coronavirus"* *"So when it rains do the remains get washed into a subwater system feeding wells, lakes, rivers and streams? Always worry about pandemic remains...and mass graves. New York City considers mass grave in park for coronavirus victims #COVID2019 #COVID19 https://t.co/kSmVSTodzo"* *"I have a student in the Guangdong province of China who reported to me that Africans are being targeted and forced into 14-day quarantine in hotels. What is going on? Anyone knows more? @msnbc @cnn #pandemic #panic #COVID19 #coronavirus #China #AfricansinChina"*	*Inciting panic (PO)*
i)	*"Effective immediately, we're expanding #COVID19AB testing to anyone in Alberta who has fever, cough, shortness of breath, a runny nose or a sore throat. #ableg #COVID19 https://t.co/dAoQmKo4Yv"* *"Scary to think UK now has 987 more deaths than China, where #coronavirus started, and 861 more deaths than Iran, where mass graves where shown by satellite. Government should be ashamed and held responsible for these completely avoidable deaths #COVID2019 #COVID19"*	*Panic-unrelated content (UN)*

panic-related content is not necessarily related to panic attack (see Table I b)-h)). For example, during COVID-19 Twitter users were posting a lot about "panic-buying" (Table I b)) and media/politicians provoking panic (Table I c)); some were trying to calm down other people trying to stop massive panic (Table I d)), others were pointing at people panicking (Table I e)), spreading fake news/conspiracy theories (Table I g)) or even attempting potential frauds (Table I f)), while some

users, unfortunately, were even trying to provoke panic (see Table I h)). Moreover, panic attack can occur either because of real danger or because of an inner emotional tension that contributes to events being perceived as threatening. This makes even manual annotation difficult, since just based on the written text, an annotator cannot be sure if the author's perception is real or imaginary. Nevertheless, detecting panic attacks is crucial both in everyday life and even more in turbulent times such as the recent COVID-19 pandemics, where even people with no previous history of panic disorders have faced such issues.

All of the above factors reinforce the limitation of resources to study this problem, as no annotated datasets are currently available. Thus, making use of supervised machine learning approaches is impossible. Therefore, in this project, we aim to provide a relevant annotated dataset originating from social media. Since human annotation process is costly, tedious and time consuming, and especially as social media volumes are immense, our goal is to train a machine learning classifier that will serve as data labeller, based on as little human annotation effort as possible. To this end, we propose an active learning-based [15] approach supported by carefully devised features and domain expertise. We specifically resort to active learning as it proved to be beneficial for addressing problems with the shortage of labelled examples [16]–[18]. More concretely, it permits for a classifier to actively choose the most informative examples for growing its training set. Typically, these examples significantly enhance classification performance, as also demonstrated in case of text classification [16], [19]. As most informative examples we consider those where the model obtains the most uncertain predictions and we ask the annotators to help the model by labelling those examples in the next round of annotation. Given that panic is closely related to fear and anxiety and could be easily confused with any of the two, to focalize the dataset and improve classifier potential, apart from domain expertise we also resort to contextualized (Transformer-based) representations [20]–[22], symptomatic and psychometric features [23], [24] that encode various psychological aspects hidden in the textual content. We apply our method to a COVID-19 Kaggle tweet dataset. With only 353 tweets labelled in two annotation rounds by our six annotators, we demonstrate that with proposed approach model's performance improves across rounds and that our model is able to obtain certain predictions for more than 12,600 tweets. In the end, we contribute the research community with annotated dataset of more than 13,000 annotated tweets (human annotations + certain model predictions) together with the conducted workflow[1].

II. Related Work

In the current literature, detecting panic has mainly been investigated in the context of public gatherings and crowded situations, based on videos [25]–[27] and images [28] tracking facial expressions and (unusual) people movements.

[1]https://github.com/SandraMNE/AnnotatingPanic

To the best of our knowledge, very few works study panic using textual sources [29], [30]. The aim of [29] was oriented towards understanding which features and metrics are relevant for assessing the panic potential of a message. Therefore, human annotators were engaged to label social media messages (tweets) with respect to how likely a particular message could incite panic. Apart from that, our paper differs in features and classifier used, as well as the annotation methodology where we used various elements of active learning. On top of this, our collected dataset is in English, and we will make it publicly available to the research community.

On the contrary, in [30], an unsupervised approach for panic detection was suggested. The paper, however, considers panic in general, without differentiating between panic categories. In addition, the paper mostly focuses on handling negation using contextual valence shifters for better sentiment classification.

III. Research Objectives and Preliminary Assumptions

As already mentioned panic can occur and be referred to in different contexts in social media posts, which might both be closely related to panic attack or completely orthogonal to it. On the other hand, a user might post a content full of fear and anxiety that is not necessarily related to panic. For example, during COVID-19 pandemics, many people tweeted about the number of recovered and dead, expressing concern and fear but without actually panicking (see Table I i)). Therefore, we decided to distinguish among three classes:

- Class *"Person Panicking"* (denoted as "PP"): a content whereby an individual describes either a personal experience of a panic situation or witnessing a panic attack-like experience;
- Class *"Panic - Other"* (denoted as "PO"): any content which is somehow related to panic but which does not fall under "PP";
- Class *"Unrelated"* (denoted as "UN"): any content not relating to panic, no matter how fearful, anxious or negative it is.

The above classes are set as annotation target classes. We consider "PP" as our main target class and our main objective is to annotate as many tweets categorized with "PP" as possible. We also aim at avoiding multi-labelling: a tweet should belong to a single class.

Given the classes under consideration, we pose an assumption that some characteristics of posted content might be a necessary (but not also a sufficient) condition for class discrimination. As such, we assume that:

H1 The content is more likely to belong to "PP" or "PO" class if it contains the keyword "panic";

H2 The content is more likely to belong to "PP" class if it contains:
   a) first-person point of view;
   b) physical symptoms that usually co-occur with panic attacks (albeit this can be deceptive, see Table I i));
   c) is negative and refers in some way to a disease, a mental state etc.;

H3 The content is more likely to belong to "UN" or "PO" class if it contains an URL link, since personal testimony in panic attack "mode" would not require an external content (and would not provide enough patience and reasoning capability to include a visual illustration of personal condition, e.g. a user attaching an illustrative meme).

Please note that we do not aim to identify tweets of panic-disorder diagnosed patients, but rather any occurrence of panic attack(-like) experience (even in a healthy individual).

IV. Panic Annotation Framework

This section presents the proposed annotation framework (see Figure 1), based on active learning paradigm. It represents an iterative process consisting of four main pillars: candidates selection, (human) annotation process with majority voting as post-processing step, model building, and certainty/uncertainty sampling depending whether the selected stopping criteria is satisfied or not. Active learning help us mitigate the annotation effort by selectively annotating unlabeled data, using most informative examples.

As a preparatory step, given the volume of social media data, we perform featurization for panic annotation, exploiting psychometric and symptomatic features. Based on these, we generate a focalized dataset which serves as a base for candidate selection.

Candidates selection refers to extracting a sample of observations to be annotated. This serves to restrict the dataset size and reduce the annotation effort. To this end, different strategies could be devised, ranging from a random to a more informative sampling. The former is more straightforward but might be less favorable in the case of excessive class imbalance. The latter entails a more problem-appropriate sampling, potentially involving domain knowledge as well, and is consequently more likely to capture the target class but unfortunately can also lead to sample bias.

The annotation process involves the efforts of human annotator(s) to label the observations from the candidates dataset, according to the given guidelines.

The obtained annotated dataset is used to construct predictive model in a supervised manner. The model not only infers the labels for the remaining set of instances, but also outputs the confidence score for each prediction. Based on the latter, uncertainty score per observation is calculated. It serves as a criterion for preparing the basis for the next iteration candidate selection step.

V. Featurization for Panic Annotation

We translate the characteristics mentioned in III into features[2] that we will use for dataset preprocessing and model building.

The "panic" keyword indicator variable $\mathbb{1}_{\exists "panic"}$, equaling 1 if the term "panic" (case-insensitive) occurs in the text and 0 otherwise, is used to denote the keyword presence.

[2]The terms "feature" and "variable" will be used interchangeably in what follows.

Next, we focus on the symptomatology of panic attack, that is, we intend to look for mention of the physical symptoms accompanying panic attack. The idea behind such approach is that the user might complain about experiencing physical symptoms characteristic of a panic attack, without actually explicitly mentioning the keyword "panic". To collect physical symptoms, relevant panic-related medical literature was consulted. Lexicons identified in three previous related works [8], [31], [32], selected as they appeared to be the most comprehensive, were unified into a single list[3] of panic symptoms $\mathcal{S}$. Same as before, an indicator variable $\mathbb{1}_{\exists s}$ is assigned to each physical symptom s, $s \in \mathcal{S}$. The issue with physical symptoms characterizing panic attacks is that they are not exclusively linked to panic. Specifically, among the symptoms mentioned in the above papers we could observe that many of these could be observed with other diseases, as, for example, palpitations, shortage of breath, sweating, etc.

To identify whether the content refers to the first-person perspective, we introduce an indicator variable that captures personal pronoun, and refers, in particular, to the first person singular.

Finally, for encoding more sophisticated information from psychological perspective (including emotions- and health/illness-related information), we resort to psychometric features. Psychometry emerged from the evidence that words hold a significant psychological value [24]. It aims to translate and quantify psychological and psychosocial constructs into observable variables [23]. To extract psychometric features, we exploit the newest edition of Linguistic Inquiry and Word Count (LIWC) [24], a well-known text analysis application widely used in the research community. Since panic can easily be confused with both fear and anxiety, we decided to not take into account a single emotion-related information (although LIWC-22 has *emo_anx* as a variable), but instead, to consider only if the text emotion and tone are negative (*emo_neg* and *tone_neg*). On top of these, we utilize two health-related psychometric variables: *illness* and *mental*. Furthermore, even though first-person point of view could be easily captured using natural language processing Part-of-Speech tagging, we opt for using LIWC-22 readily available *i* variable for this purpose, increasing the number of used LIWC-22 variables to five (5). Each of these variables (denoted as $LIWC_x$, with x being the name of the original LIWC-22 variable) takes non-negative continuous values with 0 denoting the absence of the observed construct (e.g. $LIWC_{illness} = 0$ means that the text does not refer to an illness).

VI. Data

Initial Dataset We start with publicly available Kaggle dataset on COVID-19 -related tweets[4]. The dataset is gathered based on COVID-containing hashtags and consists of 8,642,360 tweets out of which 4,827,372 in English.

[3]The complete list of the extracted symptoms will be included with the code.

[4]https://www.kaggle.com/datasets/smid80/coronavirus-covid19-tweets-early-april

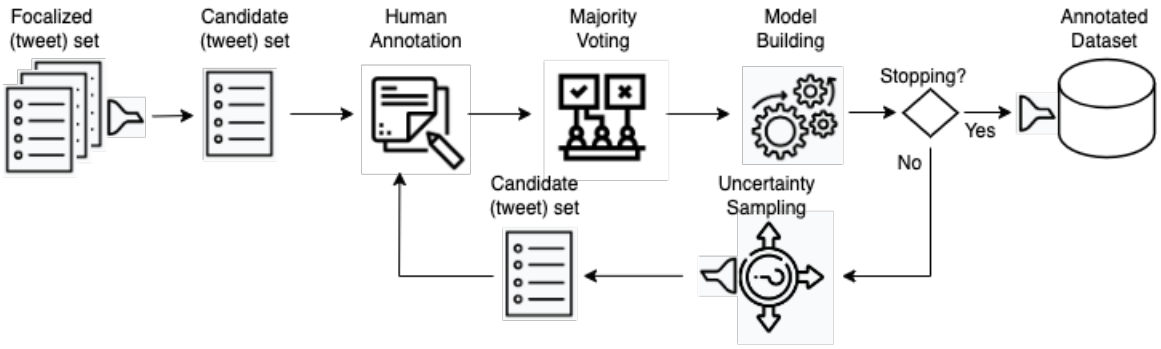

Fig. 1. The illustration of panic annotation workflow.

Data preprocessing We focus on English tweets and additionally filter all tweets that are either replies to the original tweets (variable *"reply_to_user_id"* contains non missing value) or are quotes of another tweet (variable *"is_quote"* is True). We thus end up with 3,506,040 tweets. The original Kaggle dataset is already free from retweets but contains number of retweets for each tweet.

Focalizing the dataset Selecting candidates for annotation is essential as annotating more than 3.5 millions of tweets would be a tremendous effort, if not almost impossible. A straightforward approach could be to rely on extracting tweets containing "panic" keyword. However, we found that "panic" keyword occurs in only slightly more than 8.1K tweets (out of 3.5M tweets in total, so around 0.23%). One the one hand, 8.1K tweets is still too many to annotate. On the other hand, if we resorted to random sampling of tweets, given that the number of tweets with "panic" keyword is too low (only 0.23%), we might have ended up in the situation where no tweet is annotated as panic-related (that is, classified as either of "PO" or "PP", but especially "PP" as our main target class). We, therefore, further narrow down the obtained dataset, performing a filtering as follows. We keep a tweet if:

- each of considered LIWC features is positive ($\forall x, x \in \{1, ..5\}: \ LIWC_x > 0$) OR
- "panic" keyword is present ($\mathbb{1}_{\exists \text{“}panic\text{”}} = 1$) OR
- at least one of the physical symptoms typical for the panic attack is present ($\sum_{s \in \mathcal{S}} \mathbb{1}_{\exists s} > 0$)

After filtering, we obtain our focalized dataset of 16540 tweets.

VII. Candidate Selection for Annotating Panic

Our focalized dataset is still too large for manual annotation. Choosing observations that are to be annotated could be done in different ways in active learning setup. One possible and frequently used option is pool-based sampling where observations are selected from a large pool of unlabeled data. We follow this direction and subsample the large unannotated set of observations to a quite smaller one, exploiting the same panic-related features and psychometric features considered for constructing focalized dataset and assumptions discussed in III. More specifically, we impose that either one of the following two criteria (at observation level) is satisfied:

1) all selected LIWC features are positive and at least one of the (panic keyword, panic physical symptoms) is present: $(\forall x: LIWC_x > 0) \wedge ((\mathbb{1}_{\exists \text{“}panic\text{”}} = 1) \vee (\sum_{s \in \mathcal{S}} \mathbb{1}_{\exists s} > 0))$

2) at least one of the selected LIWC features is not present but both panic keyword and panic physical symptoms are present: $(\exists x: LIWC_x = 0) \wedge (\mathbb{1}_{\exists \text{“}panic\text{”}} = 1) \wedge (\sum_{s \in \mathcal{S}} \mathbb{1}_{\exists s} > 0)$

This procedure is devised in a tailored way to retain as many as possible panic-related tweets, keeping the size of candidate dataset reasonable for annotating efforts.

VIII. Human Annotation Process

Actual annotation, that is, assigning tweets to our target classes was done manually by human annotators[5]. Although the number of annotation rounds could be arbitrarily high in the active learning setup, in order to reduce the load on human annotators, we required only two annotation rounds for training purposes.

Six annotators were involved in the annotation process. Among these, three annotators have expertise in the relevant domains, these being clinical psychology, cognitive science and English language and literature, respectively. Other three annotators have good command of English language but come from geographically distant locations and culturally different backgrounds, to confront comprehension of written texts from different cultural perspectives. As one of these three (non-domain expert) annotators later became involved in the model generation, he was substituted by another annotator in the subsequent annotation round, in order to avoid potential influencing/bias on annotations.

After collecting the output from human annotators, we check the Inter-Annotator Agreement (IAA) which measures the level of agreement between annotators. Given that we have more than two annotators, instead of Cohen's kappa [33] score, we use Fleiss' kappa [34] score, which quantifies how well the class assignments agree over a group of multiple annotators. With six annotators, it is reasonable to expect that unanimous

[5]Annotation using ChatGPT was attempted several months upon its introduction, but to no avail: on a small sample of tweets that a human would easily assign to different classes, ChatGPT was always assigning the same class (PP).

voting would be hard to reach, hence we opt for resolving disagreements using majority voting.

IX. Predictive Model Building

As mentioned, human annotator input is first reconciled with majority voting. During this process, tweets with no majority agreement are discarded. The remaining tweets are used for building a predictive model in order to classify the rest of our focalized dataset into our three target classes.

Model target per tweet is determined by majority vote class, while for the model input we exploit:

- $LIWC_x$ features (denoted together as *LIWC*, 5 in total)
- panic keyword indicator $\mathbb{1}_{\exists\, \text{“}panic\text{”}}$ (denoted as *P*)
- physical symptoms indicator $\sum_{s \in \mathcal{S}} \mathbb{1}_{\exists s}$ (marked as *S*)
- URL presence indicator $\mathbb{1}_{\exists\, URL}$ (denoted as *U*), equaling 1 if tweet text contains an URL and 0 otherwise
- contextual embeddings (denoted as *Emb*, 768 features) of tweet texts, obtained using Huggingface Twitter4SSE model [22] which was trained using Sentence Transformers [35] and initialized with the BERTweet model [21], already trained on tweet corpora (including tweets from COVID-19 period). Tweet texts are previously preprocessed to eliminate (uninformative) URLs.

Gradient Boosting is chosen for a classifier, since it is well-known to outperform other non-deep learning methods, such as logistic regression and Random Forest. Unfortunately, small training set size impedes the usage of deep architectures.

Once the model is trained (and evaluated) on small set of annotated tweets, it is applied on the large set of unannotated tweets. To obtain more robust evaluation (especially given small training size), we repeat the procedure *n* times[6].

X. Uncertainty Sampling

Once the model inference on large unannotated corpus is completed, the observations whose corresponding predictions the model is the most uncertain about should be sent for the next round of annotation. In the active learning terminology, this is referred to as uncertainty sampling. Since uncertainty could be measured using different approaches, we consider the two approaches most appropriate for our objective. As mentioned in III, we target at each tweet being exclusively (and undoubtedly) assigned to a single class, hence the model should confidently decide in favor of one particular class. Given that our model outputs three probability scores (one per each class) per run (*k*), we translate the above requirement into two uncertainty criteria c_1^k and c_2^k defined as:

$$c_1^k(x) = \begin{cases} 1, & \text{if } \mathrm{P}^k(\hat{y}_{max}|x) < th_{c_1} \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where $P^k(\hat{y}_{max}|x)$ denotes the most confident probability score for the observation x within the k^{th} run, and

$$c_2^k(x) = \begin{cases} 1, & \text{if } \mathrm{P}^k(\hat{y}_{max}|x) - P^k(\hat{y}_{max-1}|x) < th_{c_2} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where $P^k(\hat{y}_{max}|x)$ and $P^k(\hat{y}_{max-1}|x)$ denote the most confident and the second most confident score for the observation x within the k^{th} run, respectively.

In other words, with the first criterion, we consider the observation as uncertain if its most confidently predicted class ($P^k(\hat{y}_{max}|x)$) is predicted with probability that is less than a given threshold th_{c_1}. This criterion is very intuitive and logically similar to well established least confidence sampling [18] in active learning, except that the latter (as its name says) calculates the inverse and performs ranking of the observations instead of fixing the predetermined threshold.

With the second criterion, we consider the observation as uncertain if the difference between the two most confident predictions is less than a given threshold th_{c_2}. This approach is again very intuitive and very similar to active learning margin of confidence sampling [36], except that again, instead of exact threshold ranking is used.

Finally, we proclaim the observation as uncertain $uncert(x) = 1$ if either of the two conditions c_1^k and c_2^k is satisfied in at least half number of runs, that is:

$$uncert(x) = \begin{cases} 1, & \text{if} \sum_{k=0}^{n} \mathbb{1}_{c_1^k(x)=1} \geqslant \frac{n}{2} \text{ or } \sum_{k=0}^{n} \mathbb{1}_{c_2^k(x)=1} \geqslant \frac{n}{2} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

Observations marked as uncertain serve as input to the next manual annotation round.

XI. Generating Final Dataset

The stopping criterion for our iterative workflow could be selected in many different ways: we might decide to stop only after obtaining specific classification accuracy, or we might stop based on the number of annotation rounds (our case for the sake of reducing annotation effort) or we might continue until we obtain somewhat balanced distribution between classes. Regardless of which stopping criteria we opt for, we select for the final dataset only the observations where the model is absolutely certain about corresponding prediction. We consider a prediction as certain (according to model) if it is certain within each of *n* performed runs and we define the prediction as certain within a run if and only if, none of the criteria c_1 and c_2 for that prediction are satisfied within a run.

XII. Annotation Evaluation

A. Round 1 Evaluation

Human Annotation Using the two criteria explained in VII, we obtained 213 tweets. These were were sent to annotators to kick-start the annotation round 1.

We obtained low annotator agreement in the round 1. This is partially a consequence of somewhat misaligned annotation guidelines (with respect to the objective stated in III). More

[6]The procedure is repeated with different (but saved) random seeds to ensure reproducibility of results.

TABLE II
ANNOTATORS AGREEMENT RESULTS.

Round #	Total Tweets	Fleiss' kappa	Unanimous agreement				Majority Voting				
			PP	*PO*	*UN*	*Agreement # / %*	*PP*	*PO*	*UN*	*NoMajority*	*Agreement # / %*
Round 1	213	0.2787	24	12	0	36 / 16.9	94	59	7	53	160 / 75.12
Round 2	308	0.1163	0	9	4	13 / 4.22	1	101	91	115	193 / 62.66

specifically, annotators were given a freedom to assign a tweet to more than one class and/or to skip the instance in case of doubt. This led to quite some noise which, obviously, had to be cleaned. Nevertheless, as can be seen in the Table II, the number of tweets with unanimous agreement was quite modest, hence as anticipated, we performed additional preprocessing applying majority voting class assignment. Even though the obtained Fleiss' kappa score was only 0.2787 which, according to Landis and Koch [37] assessment is still considered fair agreement, after majority voting we obtained 160 tweets that we could further use for building our multiclass classifier.

A few examples where majority voting did not result in agreement could be seen in Table IV.

Model Evaluation We slightly diverge from classical active learning setup and instead of using the complete annotated dataset (160 tweets) for training, we also use it for model evaluation, dividing train and test into stratified splits using 70/30 ratio. The LIWC features are standardized to zero mean and unit standard deviation using StandardScaler method of Scikit-learn [38] (performed better than normalization). We use XGBoost [39] library for model implementation.

The model performance in terms of F1-score could be seen to the left side of Table III (Round 1). The best performing features are *Emb+LIWC* for all three F1-scores and we see that adding additional features on top of these does not help. Also, as expected due to class imbalance, macro F1-score is remarkably lower than its weighted/micro counterparts.

Checking the 20 most important features of XGBoost classifier we noted that $LIWC_i$ feature is ranked as second most important feature in all the runs, which undoubtedly contributes to identification of "PP" class given that they relate to personal content.

Uncertainty Sampling We empirically choose $th_{c_1} = 0.51$ since we would like that the most confident score is very dominant compared to the other two, and in the worse case, at least slightly higher than half. Additionally, we empirically choose $th_{c_2} = 0.2$, with the motivation that if the first criterion is not satisfied, that is, $P^k(\hat{y}_{max}|x) > 0.51$ but still very low, e.g. < 0.6, that at least the difference between the highest two probabilities is pronounced.

As we considered $n = 10$ runs for our predictive model, according to our uncertainty definition in Equation 3, we should check if at least $n/2 = 5$ times an observation satisfies either of the two conditions. Number of observations satisfying criterion c_1 at least $n/2 = 5$ times is 99, while 297 observations satisfy criterion c_2 at least $n/2 = 5$ times. As 88 of these observations satisfy both conditions at least $n/2 = 5$ times, hence, in total 308 observations satisfy either of the two at least $n/2 = 5$ times. These were used as input for the second round of annotation.

B. Round 2 Evaluation

Human Annotation Annotators obtained 308 tweets that the model was uncertain about. Even though the annotator agreement was again very low (the unanimous agreement was met on only 13 tweets and the obtained Fleiss' kappa score was even worse than in the first round (only 0.1163, see Table II), after majority voting adjustment we obtained 193 tweets that we could add to previously acquired 160 to retrain our multiclass classifier. Additional beneficial side effect of round 2 annotation is that the distribution of classes became skewed in the direction of UN class. Although this still did not make our classes balanced, it remarkably reduced the disproportion.

Model Evaluation At round 2, we repeat the exact same procedure for the model training and evaluation as at round 1, except that instead of 160 tweets we consider 353 (=160+193) tweets. The results obtained are presented to the right side of the Table III (Round 2). We can see that in terms of weighted, micro and macro F1-score, the round 2 model outperforms the round 1 model for 8.2%, 6.1% and as much as 56.8%. This proves that our active learning based approach works. It is to be noticed that the best performing features for the round 2 model are *Emb+LIWC+P+S*, different from *Emb+LIWC* for the round 1 model.

Checking the classifier 20 most important features we noted that $LIWC_i$ and $LIWC_{mental}$ features are always ranked among the most important features in all the runs.

Uncertainty Sampling following the same thresholds as at round 1, we obtain only 12 observations satisfying criterion c_1 at least $n = 5$ times and only 122 observations satisfy criterion c_2 at least $n = 5$ times. In total only 122 observations satisfy either of the two at least $n = 5$ times. It is worth noticing that model uncertainty decreased as compared to the round 1. In theory, these 122 observations should be further sent to annotators for the third round of annotation, but we have decided not to continue with annotations beyond this point.

C. Final Dataset

Other than uncertain predictions within round 2, we check for the certain ones. We find that 12,683 tweets have certain predictions (out of 16,187). These tweets, together with 160 and 193 tweets annotated in the first and second round, respectively, constitute final version of the released dataset (final class distribution can be seen in Table V).

TABLE III
CLASSIFICATION RESULTS PER ANNOTATION ROUND (AVERAGED ACROSS 10 RUNS EACH).

Features	Round 1			Round 2		
	Weighted F1 *(avg ± std)*	***Micro F1*** *(avg ± std)*	***Macro F1*** *(avg ± std)*	***Weighted F1*** *(avg ± std)*	***Micro F1*** *(avg ± std)*	***Macro F1*** *(avg ± std)*
LIWC	0.763±0.023	0.775±0.024	0.585±0.093	0.756±0.035	0.759±0.033	0.745±0.037
LIWC+P+S	0.748±0.023	0.756±0.025	0.529±0.049	0.854±0.024	0.855±0.024	0.852±0.027
Emb	0.825±0.041	0.842±0.042	0.566±0.029	0.785±0.026	0.787±0.026	0.784±0.027
Emb+P+S	0.825±0.040	0.842±0.041	0.567±0.028	0.903±0.039	0.903±0.039	0.902±0.039
Emb+LIWC	0.851±0.034	0.869±0.036	0.588±0.025	0.833±0.037	0.835±0.035	0.833±0.037
Emb+LIWC+U	0.851±0.034	0.869±0.036	0.588±0.025	0.835±0.038	0.837±0.037	0.835±0.039
Emb+LIWC+P+S	0.846±0.040	0.8625±0.043	0.583±0.030	0.921±0.033	0.922±0.032	0.922±0.032
Emb+LIWC+P+S+U	0.825±0.040	0.842±0.041	0.567±0.028	0.915±0.030	0.915±0.030	0.915±0.030

TABLE IV
EXAMPLES OF TWEETS WITH NO MAJORITY VOTE AND THEIR RELATIVE HUMAN ANNOTATIONS.

Round #	Tweet text	Class: # of annotations
R1	*"Personally, I don't have a problem with #StayAtHome Before #coronavirus I never left the house. Outside is overrated! #anxiety #OCD #agoraphobia #recluse"*	*PP:1, PO:1, PP+PO:1, UN:3*
	"I've just discovered this if you are a #vulnerable person. Unfortunately it doesn't include people with #mentalillness. This could include #anxiety and #agoraphobia which makes it extremely difficult to get out. Why? #coronavirus https://t.co/pGlQJYvu7D"	*PP:1, PO:2, UN:3*
	"Why did I start watching #pandemic? I feel stupid. Thanks @netflix #paranoia #netflix #coronavirus #CoronavirusUSA #panic"	*PP:1, PO:2, UN:3*
	"I think I relate to this because coping with anxiety is finally just enough. Everyone is dealing with the same panic now + staying home. Whereas on a normal day I'd be coping with anxiety + normal life functions. #coronavirus #anxietydisorder #Mentalhealth https://t.co/y4dBrzSqXv"	*PP:2, PO:2, PP+PO:1, UN:1*
R2	*"Coping with stress and anxiety amid coronavirus outbreak https://t.co/Rt9r1Rb1f9 #MentalHealth #COVID19 #Anxiety #Stress - Check out my chat with Good Day Rochester! @foxrochester"*	*PP:1, PO:2, UN:2*
	"This is the best layout of the money and motives behind the #COVID19 panic I have seen to date. Tel your friends! https://t.co/sD55Lo1I96"	*PO:3, UN:2 (?:1)*
	"Evidences are being looked at all the time about mask. Maybe I will just panic order myself mask rather than wait but I want to save lives and protect the NHS. #DailyBriefing #PressConference #coronavirus"	*PP:3, PO:3*
	"#California The stress of #coronavirus can affect your mind, body, spirit, and relationships. I have specialized expertise in treating #stress and #trauma, I've opened up a few new slots in my tele-health practice for patients throughout CA https://t.co/SyI4Cu9f7e #anxiety https://t.co/toKlU1Y1km"	*PO:3, UN:3*

TABLE V
CLASS DISTRIBUTION IN THE FINAL DATASET. "MV" STANDS FOR "MAJORITY VOTING", "RX" FOR THE ROUND #X.

Class	MV R1	MV R2	R2 model unanimous	Total # (%)
PP	94	1	934	1029 (7.89)
PO	59	101	6995	7155 (54.89)
UN	7	91	4754	4852 (37.22)
Total	160	193	12683	13036 (100)

D. Evaluation on a Random Sample

In the end, we asked annotators to label another set of 161 randomly selected tweets, to be used for evaluation purposes exclusively. We again received low unanimous agreement and resorted to majority voting scheme. Out of 161 tweet, with majority voting we get agreement on 109 tweets, with the following distribution: 35, 69 and 5 per "UN", "PO" and "PP" class, respectively. Figure 2 shows how well human annotations (to the left) agree with model annotations (to the right). In particular, we check how many times (out of total 10 runs), model confidently predicts a particular label. For example, "PO (M-7x)" means that the model 7 out of 10 times gives certain prediction in favor of class "PO". The green color denotes the alignment between human and model label, while red color denotes disagreement. The stronger nuances of color indicate the stronger model confidence (higher number of runs that the model is certain about prediction). Hence, the more confidently the model agrees with human label, the darker the nuance of green color gets, and similarly, the more confidently the model misclassifies the human label, the stronger the red nuance is (to "penalize" the error). We can observe that most of the tweets of "PO" (47) and "UN" (24) classes are correctly classified by model, although there are 9 tweets of "PO" class which are misclassified as "UN" with high confidence and all 5 tweets of "PP" class are misclassified.

XIII. DISCUSSION AND LIMITATIONS

We would like to emphasize that some choices were made due to the characteristics of the data at hand. For example, even though we would like to ensure robustness using cross validation, due to the small number of observations belonging to "UN" class in the first round (only 7) we had to resort

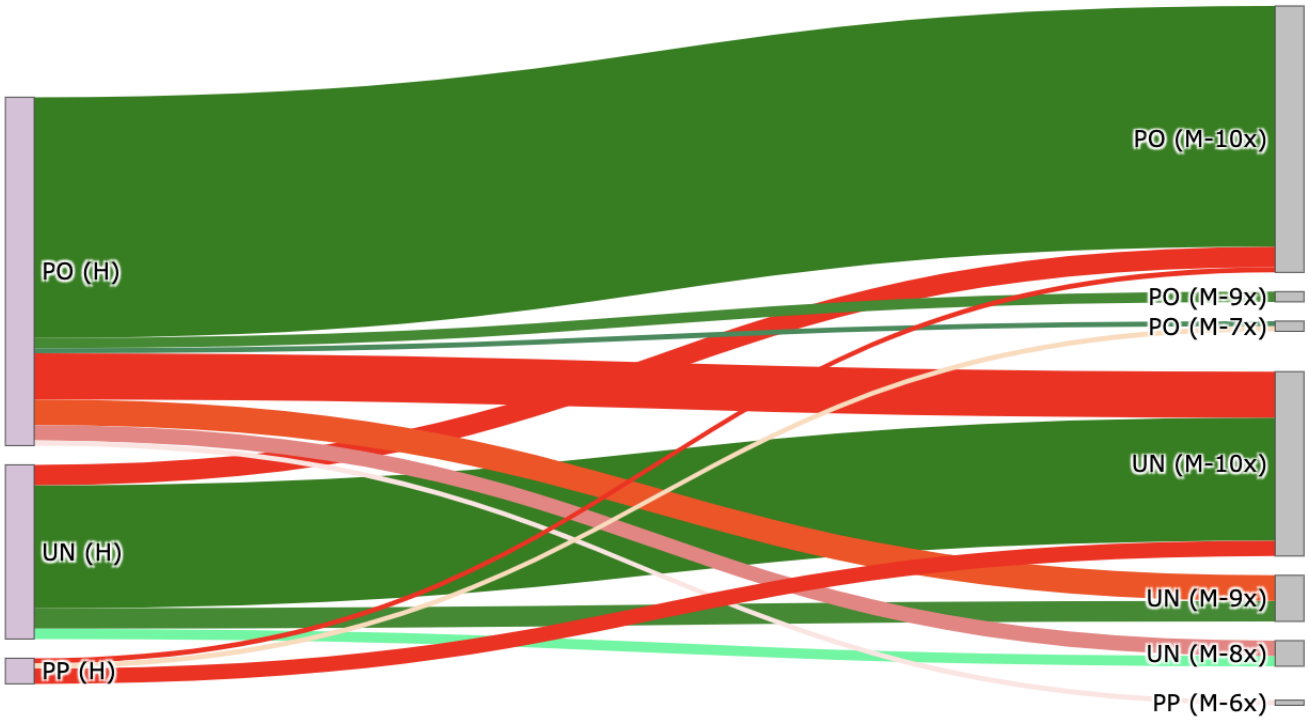

Fig. 2. Sankey plot showcasing the alignment of human (to the left) with model annotations (to the right). "-Kx" denotes that model predictions are certain in K (out of $n = 10$) runs. Green color indicates alignment while red color indicates misalignment.

to alternative approach of simple train/test splits with n repetitions.

While the two standard approaches (least confidence sampling and margin of confidence sampling) seem more elegant as they do not require the predefined thresholds, our approach became handy with multiple runs as we have not needed to calculate the average rank across multiple runs. If the number of uncertain tweets was higher (e.g. in the order of thousands or so) we would certainly resort to ranking approach in order to be able to select reasonable amount of tweets for annotation. On the other hand, using a threshold helps us identify if the number of uncertain observations decreases across rounds which would mean that the model is gradually getting confident about its predictions.

We are aware of several limitations of our work. First, the misalignment of our initial annotation guidelines has definitely taken a toll on the annotator agreement. In a hindsight, the annotation (c/s)ould have been repeated, but we were concerned about the annotator engagement and annotation quality in case they were to repeat similar task. Also in the latter case, we would need to assess annotators' consistency with previous annotation which is not a problem from the computational side but it would be very time consuming to redo the whole procedure in case of many inconsistencies. On the other hand, even worse annotator agreement in the second round leaves a doubt that initial misalignment had no effect and that the task is simply difficult (or that the tweets the model was unsure about are really difficult to be classified).

Second, the model performances could be easily improved due to two reasons. One is that we used only 70% of data for the model training instead of all, which certainly decreased classifier accuracy given that we already have very little annotated data, (especially in the round 1 - only 160). Another is that we opted for using only human annotated tweets (160+193) for training classifier in round 2, even though, according to the active learning paradigm we could have included also the observations for whom the round 1 model provided certain predictions. We made this choice deliberately as we rather preferred to obtain reliable than overly optimistic classification results.

XIV. Conclusion and Future Work

This paper addresses the understudied problem of detecting panic-related content in textual social media posts (tweets). More specifically, it proposes an active learning-based approach supported with psychometric, symptomatic and Transformer-based features as well as domain expertise, in order to provide a relevant annotated dataset. During annotation, we differentiate between personal panicking experiences, other panic-related content and completely panic-unrelated content. Our evaluation results showcase that with the proposed active learning framework, classifier performance improves remarkably after only two rounds. Furthermore, our model confidence improves which permits us to generate a final annotated dataset of 13,036 (including 353 human annotated) tweets. To the best of our knowledge, this is the

first dataset of this kind and scope, which we share with the research community in order to encourage further research. As per future work, based on obtained human annotations we are planning to explore other approaches in the directions of data augmentation. Additionally, we hope to experiment with other social media datasets in order to gain better insights into the actual share of panic-related (and especially panic attack-related) content in social media, even in the less critical periods (unlike COVID-19 pandemic).

REFERENCES

[1] T. L. G. Health, "Mental health matters," The Lancet. Global Health, vol. 8, no. 11, p. e1352, 2020.

[2] E. A. Ríssola, S. A. Bahrainian, and F. Crestani, "A dataset for research on depression in social media," in Proceedings of the 28th ACM conference on user modeling, adaptation and personalization, 2020, pp. 338–342.

[3] D. William and D. Suhartono, "Text-based depression detection on social media posts: A systematic literature review," Procedia Computer Science, vol. 179, pp. 582–589, 2021.

[4] E. Mohammadi, H. Amini, and L. Kosseim, "Quick and (maybe not so) easy detection of anorexia in social media posts." in CLEF (Working Notes), 2019.

[5] A. Murarka, B. Radhakrishnan, and S. Ravichandran, "Detection and classification of mental illnesses on social media using roberta," arXiv preprint arXiv:2011.11226, 2020.

[6] D. H. Barlow, J. Vermilyea, E. B. Blanchard, B. B. Vermilyea, P. A. Di Nardo, and J. A. Cerny, "The phenomenon of panic." Journal of Abnormal Psychology, vol. 94, no. 3, p. 320, 1985.

[7] S. Taylor and G. J. Asmundson, "Panic disorder," in The Handbook of Adult Clinical Psychology. Routledge, 2014, pp. 482–510.

[8] M. H. Pollack, "The pharmacotherapy of panic disorder," J Clin Psychiatry, vol. 66, no. suppl 4, pp. 23–27, 2005.

[9] F. Barale, B. Mauro, G. Vittorio, S. Mistura, and A. Zamperini, "Psiche. dizionario storico di psicologia, psichiatria, psicoanalisi e neuroscienze. vol. i ak," 2007.

[10] G. A. Clum, G. A. Clum, and R. Surls, "A meta-analysis of treatments for panic disorder." Journal of consulting and clinical psychology, vol. 61, no. 2, p. 317, 1993.

[11] N. Chawla, T. Anothaisintawee, K. Charoenrungrueangchai, P. Thaipisuttikul, G. J. McKay, J. Attia, and A. Thakkinstian, "Drug treatment for panic disorder with or without agoraphobia: systematic review and network meta-analysis of randomised controlled trials," bmj, vol. 376, 2022.

[12] D. American Psychiatric Association, A. P. Association et al., Diagnostic and statistical manual of mental disorders: DSM-5. American psychiatric association Washington, DC, 2013, vol. 5, no. 5.

[13] U. Galimberti, Nuovo dizionario di psicologia: psichiatria, psicoanalisi, neuroscienze. Feltrinelli, 2018.

[14] A. Sims and F. Oyebode, Introduzione alla psicopatologia descrittiva. Raffaello Cortina, 2010.

[15] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, "Active learning with statistical models," Journal of artificial intelligence research, vol. 4, pp. 129–145, 1996.

[16] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," Journal of machine learning research, vol. 2, no. Nov, pp. 45–66, 2001.

[17] N. Roy and A. McCallum, "Toward optimal active learning through sampling estimation of error reduction. int. conf. on machine learning," 2001.

[18] A. Culotta and A. McCallum, "Reducing labeling effort for structured prediction tasks," in AAAI, vol. 5, 2005, pp. 746–751.

[19] A. Esuli and F. Sebastiani, "Active learning strategies for multi-label text classification," in Advances in Information Retrieval: 31th European Conference on IR Research, ECIR 2009, Toulouse, France, April 6-9, 2009. Proceedings 31. Springer, 2009, pp. 102–113.

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[21] D. Q. Nguyen, T. Vu, and A. T. Nguyen, "BERTweet: A pre-trained language model for English Tweets," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 9–14.

[22] M. Di Giovanni and M. Brambilla, "Exploiting Twitter as source of large corpora of weakly similar pairs for semantic sentence embeddings," in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 9902–9910. [Online]. Available: https://aclanthology.org/2021.emnlp-main.780

[23] D. Borsboom and D. Molenaar, "Psychometrics," pp. 482–510, 2015.

[24] R. L. Boyd, A. Ashokkumar, S. Seraj, and J. W. Pennebaker, "The development and psychometric properties of liwc-22," Austin, TX: University of Texas at Austin, pp. 1–47, 2022.

[25] M. George, C. Bijitha, and B. R. Jose, "Crowd panic detection using autoencoder with non-uniform feature extraction," in 2018 8th International Symposium on Embedded Computing and System Design (ISED). IEEE, 2018, pp. 11–15.

[26] X. Zhang, X. Shu, and Z. He, "Crowd panic state detection using entropy of the distribution of enthalpy," Physica A: Statistical Mechanics and its Applications, vol. 525, pp. 935–945, 2019.

[27] K. Singh, S. Rajora, D. K. Vishwakarma, G. Tripathi, S. Kumar, and G. S. Walia, "Crowd anomaly detection using aggregation of ensembles of fine-tuned convnets," Neurocomputing, vol. 371, pp. 188–198, 2020.

[28] M. A. Hossain, D. Samanta, and G. Sanyal, "Extraction of panic expression depending on lip detection," in 2012 International Conference on Computing Sciences. IEEE, 2012, pp. 137–141.

[29] A. Hariharan, V. Dorner, C. Weinhardt, and G. W. Alpers, "Detecting panic potential in social media tweets," pp. 3181–3190, 2017.

[30] S. Mitrovic and V. Kanjirangat, "Enhancing bert performance with contextual valence shifters for panic detection in covid-19 tweets," in Proceedings of the 2022 6th International Conference on Natural Language Processing and Information Retrieval, 2022, pp. 89–92.

[31] D. Bonevski and A. Naumovska, "Panic attacks and panic disorder," in Psychopathology-An international and interdisciplinary perspective. IntechOpen, 2019.

[32] O. M. Hewitt, A. Tomlin, and P. Waite, "The experience of panic attacks in adolescents: an interpretative phenomenological analysis study," Emotional and Behavioural Difficulties, vol. 26, no. 3, pp. 240–253, 2021.

[33] J. Cohen, "A coefficient of agreement for nominal scales," Educational and psychological measurement, vol. 20, no. 1, pp. 37–46, 1960.

[34] J. L. Fleiss, "Measuring nominal scale agreement among many raters." Psychological bulletin, vol. 76, no. 5, p. 378, 1971.

[35] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 11 2019. [Online]. Available: https://arxiv.org/abs/1908.10084

[36] T. Scheffer, C. Decomain, and S. Wrobel, "Active hidden markov models for information extraction," in International symposium on intelligent data analysis. Springer, 2001, pp. 309–318.

[37] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," biometrics, pp. 159–174, 1977.

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

[39] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD '16. New York, NY, USA: ACM, 2016, pp. 785–794. [Online]. Available: http://doi.acm.org/10.1145/2939672.2939785